Advances in Machine Learning and Data Mining

Research in Computing Science

Series Editorial Board

Research in Computing Science es una publicación trimestral, de circulación internacional, editada por el Centro de Investigación en Computación del IPN, para dar a conocer los avances de investigación científica y desarrollo tecnológico de la comunidad científica internacional. **Volumen 139,** noviembre 2017. Tiraje: 500 ejemplares. *Certificado de Reserva de Derechos al Uso Exclusivo del Título* No.: 04-2005-121611550100-102, expedido por el Instituto Nacional de Derecho de Autor. *Certificado de Licitud de Título* No. 12897, *Certificado de licitud de Contenido* No. 10470, expedidos por la Comisión Calificadora de Publicaciones y Revistas Ilustradas. El contenido de los artículos es responsabilidad exclusiva de sus respectivos autores. Queda prohibida la reproducción total o parcial, por cualquier medio, sin el permiso expreso del editor, excepto para uso personal o de estudio haciendo cita explícita en la primera página de cada documento. Impreso en la Ciudad de México, en los Talleres Gráficos del IPN – Dirección de Publicaciones, Tres Guerras 27, Centro Histórico, México, D.F. Distribuida por el Centro de Investigación en Computación, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, México, D.F. Tel. 57 29 60 00, ext. 56571.

Editor responsable: *Grigori Sidorov, RFC SIGR651028L69*

Research in Computing Science is published by the Center for Computing Research of IPN. **Volume 139**, November 2017. Printing 500. Printed in Mexico City, in the IPN Graphic Workshop – Publication Office.

Advances in Machine Learning and Data Mining

Miguel González Mendoza (ed.)

Instituto Politécnico Nacional, Centro de Investigación en Computación
México 2017

ISSN: 1870-4069

Instituto Politécnico Nacional (IPN)
Centro de Investigación en Computación (CIC)
Av. Juan de Dios Bátiz s/n esq. M. Othón de Mendizábal
Unidad Profesional "Adolfo López Mateos", Zacatenco
07738, México D.F., México

http://www.rcs.cic.ipn.mx
http://www.ipn.mx
http://www.cic.ipn.mx

Indexed in LATINDEX, DBLP and Periodica

Printing: 500

Printed in Mexico

ISSN 1870-4069

Editorial

En este volumen de la revista "Research in Computing Science" se incluyen artículos relacionados con el aprendizaje máquina y la minería de datos. Todos los trabajos que aquí se presentan fueron cuidadosamente seleccionados por el comité editorial y revisados por lo menos por dos revisores externos considerando su originalidad científica y la calidad técnica.

En este volumen se incluyen quince artículos que abordan varios aspectos del aprendizaje máquina y la minería de datos. Por una parte, el diseño y uso de métodos de aprendizaje máquina permiten abordar problemas de inferencia, reconocimiento, diagnóstico, predicción y clasificación. Por ejemplo, se incluye el trabajo de Cerna-Vázquez, *et al.*, que propone el uso de una red neuronal para la predicción de datos de contaminación y prevención de ataques a personas con padecimientos de rinitis alérgica y asma. En otro campo de aplicación, los autores Pinzón, *et al*., plantean el uso de modelos auto-regresivos para predecir la generación de energía eléctrica a partir de datos adquiridos en la zona de la Ventosa, Oaxaca en México. En el trabajo "Algoritmo de desactivación de estaciones base para reducir el consumo de energía en redes celulares" se propone un algoritmo que detecta qué estaciones base de telefonía celular deben desactivarse para reducir el consumo de energía y ofreciendo los servicios demandados, mediante el uso de un algoritmo genético aplicado sobre un modelo de población estacionario. Por otra parte, Rodríguez, *et al.* presenta un estudio sobre la predicción a corto plazo de la velocidad del viento mediante series de tiempo incompletas ya que resulta importante en etapas posteriores de planeación, administración y producción de energías limpias. En otro dominio de aplicación, Meneses-Bautista y Alvarado presentan el análisis en series de tiempo para el pronóstico del tipo de cambio dólar-peso mexicano haciendo uso de redes neuronales de retro-propagación. En el artículo titulado "Una medida de distancia para construir árboles filogenéticos: un primer acercamiento" se propone una medida basada en los mejores aciertos bidireccionales para poder construir árboles filogenéticos para su uso posterior en genómica comparativa.

En el trabajo "Estudio de la indumentaria indígena mexicana" de Rodríguez-Mondragón, *et al.*, se propone el análisis de imágenes de íconos identificados en indumentaria indígena con la finalidad de poder encontrar patrones y generar nuevos diseños que mantengan la identidad visual del textil indígena. En el artículo de Flores-Bastida, *et al.* se propone un sistema de clasificación de hojas, lobuladas simples y no lobuladas simples, para la identificación de plantas. En temas de clasificación, el trabajo titulado "Generación de un vector característico para la detección de intrusos en redes computacionales" propone un sistema de detección de intrusos mediante la generación de un vector característico a partir de información real proveniente de la red con la finalidad de discernir entre un comportamiento autorizado o no permitido. En el trabajo de García-Domínguez y Galván-Tejada, se propone el estudio de actividades infantiles y su clasificación mediante el uso de modelos de agrupamiento jerárquico aplicados sobre información contenida en el espectro de sonidos emitidos al realizar las actividades. En el dominio médico, el artículo "Clasificación del cáncer

cervicouterino utilizando algoritmos ensamblados en imágenes microscópicas de Papanicolaou" propone la clasificación del cáncer cervicouterino mediante la combinación de los algoritmos Bagging+MultilayerPerceptron y AdaBoostM1+LMT.

Por otra parte, el volumen incluye trabajos relacionados a la minería de datos. En este sentido, el artículo de Reyes-Nava, *et al.*, "Minería de datos aplicada para la identificación de factores de riesgo en alumnos", presenta la automatización de un sistema integral de tutoría para detectar algunos factores de riesgo que presenten los alumnos de su institución aplicando técnicas de minería de datos. En el artículo "Modelo de deserción escolar en universidades mexicanas", los autores proponen un modelo para predecir la deserción escolar basado en la información auto-administrada por el estudiante y las puntuaciones en el examen de ingreso a la universidad, aplicando técnicas de modelado predictivo y minería de datos. Por otra parte, en el trabajo titulado "Método de fusión de datos de fuentes heterogéneas para mantener la consistencia de datos" propone el diseño conceptual de un método de fusión de datos provenientes de fuentes heterogéneas para mantener la información contextual al momento de extraer, pre-procesar, fusionar y cargar datos. En el trabajo de González-Marrón, *et al.*, se aplican técnicas de minería de datos para determinar si hay factores socioeconómicos que permitan predecir el factor de éxito en la realización de exámenes de ingreso al instituto educativo de los autores.

Finalmente, el proceso de revisión y selección de artículos se llevó a cabo usando el sistema libremente disponible EasyChair (www.easychair.org).

Miguel González-Mendoza
Editor Invitado
Tecnológico de Monterrey Campus Estado de México, México

Noviembre 2017

ISSN 1870-4069

Table of Contents

Page

ISSN 1870-4069

Automatic Classification of Lobed Simple and Unlobed Simple Leaves for Plant Identification

Juan Carlos Flores-Bastida, Asdrúbal López-Chau, Rafael Rojas-Hernández, Valentin Trujillo-Mora

Universidad Autónoma del Estado de México, Zumpango, Estado de México, México

juancarlosfloresbastida@gmail.com

Abstract. Commonly, classic plant identification methods use dichotomous or multi-access keys that compare characteristics of the leaves, asking if they are lobed, unlobed, simple or compound, among others leaf features. However, in the literature very little attention has been paid to make an automatic distinction of leaves using such features. In this paper, we contribute to fill this gap. We propose a novel method to differentiate between types of leaves. The proposal is invariant to rotation and also to scaling. In order to show the effectiveness of the proposal, we tested it with more than 1,900 images of leaves which are publicly available on the Internet, achieving correct identification rates greater than 86%.

Keywords: compound leaf, leaf feature, lobed simple leaf, unlobed simple leaf.

1 Introduction

Most of plant identification methods use leaves. There are some good reasons for this, for example, plants have leaves almost all year [6], the number of them is usually high [1]; depth can be ignored in images of leaves, unlike flowers or other organs of plants; and leaves are different between plants [10].

One of the first phases in automatic plant identification is to extract a set of features from leaves (after some pre-processing on images). In spite of many descriptors based on the shape, color, texture and veins of leaves [18,3,4,5,11,12,16] have been proposed in last decade, little attention has been paid to develop new methods to make an automatic distinction between the different types of leaves, such as unlobed, lobed, simple and compound. This categorization of leaves is important, and it is usually used in dichotomous keys for classic plant identification, see for example [14,2,9,17,15].

We propose a new method to differentiate between simple lobed and simple unlobed leaves. Our method counts the number of changes of color of lines that are traced over the image of leaves. The proposal is invariant to rotation and also to scaling of images.

The rest of the paper is organized as follows. Subsection 2.1 explains the types of leaves, and also explains the difference between lobed simple leaves and unlobed simple ones, 2.1 describes some basic types of plant leaves. Section 2.2 shows those materials used in this paper. We present our proposals in Section 3, then Section 4 shows experiments and results. Finally, last section of this paper presents conclusions and future works.

2 Preliminaries

2.1 Types of Plant Leaves

Categorizing leaves (and plants) is not a trivial task. Leaves can be classified according to their blade (simple or compound), edge (smooth, dentate, etc.), petiole (petiolated or sessile), shape of blade, etc. Among these categories, simple, compound, unlobed and lobed are very common in dichotomous keys. For the former category, the leaf blade is a single, continuous unit. For the second category, the blade is divided into two or more leaflets arising from the petiole. Simple leaves can be unlobed or lobed. For unlobed leaves, the blade is completely undivided. Lobed leaves have projections off the midrib with individual inside veins.

In some cases, such as the leaves shown in Figure 1, the definitions above can be directly applied to categorize a leaf. However, for other leaves, such as the ones shown in Figure 2, it could be a bit more complicated to categorize them.

Fig. 1. Example of a simple leave (left) and a compound leaf (right).

Fig. 2. Some types of lobed simple leaves.

Currently, there is not a single leaf descriptor that allows to identify all types of leaves perfectly. Instead, each leaf descriptor focuses on extracting one characteristic. The methods proposed in this paper identify to which of the following two groups belongs a plant leaf: a) unlobed simple leaves with smooth margins, and b) other types (compound, lobed, palmate, etc.). This information can be encoded as a binary leaf feature in plant identification.

2.2 Materials

Flavia data set is one of most widely used data sets for testing plant identification systems. It is publicly available at `http://flavia.sourceforge.net`. Flavia set contains 1,907 color images of 32 different species of plants. These images have a dimension of 1,600 × 1,200 pixels.

Figures 3 and 4 show the scientific name of plants, the class, and an example of a leaf for each one of the 32 species of plants in Flavia set.

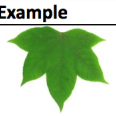
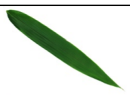
Name	Class	Example	Name	Class	Example	Name	Class	Example
Phyllostachys edulis (Carr.) Houz. **Attribute: FALSE**	C1		Kalopanax septemlobus (Thunb. ex A.Murr.) Koidz. **Attribute: TRUE**	C8		Viburnum awabuki K.Koch **Attribute: FALSE**	C16	
Aesculus chinensis **Attribute: FALSE**	C2		Cinnamomum japonicum Sieb. **FALSE**	C9		Osmanthus fragrans Lour. **Attribute:FALSE**	C17	
Berberis anhweiensis Ahrendt **Attribute:FALSE**	C3		Koelreuteria paniculata Laxm. **Attribute:FALSE**	C10		Cedrus deodara (Roxb.) G. Don **Attribute:FALSE**	C18	
Cercis chinensis **Attribute:FALSE**	C4		Ilex macrocarpa Oliv. **Attribute:FALSE**	C11		Ginkgo biloba L. **Attribute:FALSE**	C19	
Indigofera tinctoria L. **Attribute:FALSE**	C5		Pittosporum tobira (Thunb.) Ait. f. **Attribute:FALSE**	C12		Lagerstroemia indica (L.) Pers. **Attribute:FALSE**	C20	
Acer Palmatum **Attribute:TRUE**	C6		Chimonanthus praecox L. **Attribute:FALSE**	C14		Nerium oleander L. **Attribute:FALSE**	C21	
Phoebe nanmu (Oliv.) Gamble **Attribute:FALSE**	C7		Cinnamomum camphora (L.) J. Presl **Attribute:FALSE**	C15		Podocarpus macrophyllus (Thunb.) Sweet **Attribute:FALSE**	C22	

Fig. 3. Species of plants in Flavia set, first part.

Name	Class	Example	Name	Class	Example
Prunus serrulata Lindl. var. lannesiana auct. **Attribute:FALSE**	C23		Magnolia grandiflora L. **Attribute:FALSE**	C30	
Ligustrum lucidum Ait. f. **Attribute:FALSE**	C24		Populus ×canadensis Moench **Attribute:FALSE**	C31	
Tonna sinensis M. Roem. **Attribute:FALSE**	C25		Liriodendron chinense (Hemsl.) Sarg. **Attribute:TRUE**	C32	
Prunus persica (L.) Batsch **Attribute:FALSE**	C26		Citrus reticulata Blanco **Attribute:FALSE**	C33	
Manglietia fordiana Oliv. **Attribute:FALSE**	C27				
Acer buergerianum Miq. **Attribute:TRUE**	C28				
Mahonia bealei (Fortune) Carr. **Attribute:FALSE**	C29				

Fig. 4. Species of plants in Flavia set, second part.

3 Proposed Method to Distinguish Lobed Simple from Unlobed Simple Leaves

In this subsection, we introduce a novel method to detect whether a leaf is lobed from simple.

The first step is to align a binary image of a leaf. Instead of rotating L manually as in other works, we obtain automatically two new reference axes for L. The first reference axis is the line that joins the two most separate pixels in the leaf (diameter). The second reference axis is an orthogonal line to the first axis.

Algorithm 1 shows the procedure to compute the references axes for a leaf L. Figure 5 shows two examples of the references axes computed with Algorithm 1.

The second step in our method, is to drawn a set of equidistant lines over the body of L. A number of these lines are parallel to $\mathcal{L}_1$, and the same number of lines are parallel to $\mathcal{L}_2$.

Changes (from white to black or vice-versa) along each line are counted and stored in an array whose length is equal to the number of lines drawn. Then, a threshold is used to determine if the leaf is lobed. Algorithm 2 summarizes our method.

Figure 6 shows two examples of the lines obtained with Algorithm 2

Algorithm 1: Reference axes for a leaf.

Input: A binary image of leaf L

Output: New reference axes ($\mathcal{L}_1$ and $\mathcal{L}_2$) for L

1 Get de contour of the binary image

2 Compute $\underset{p_i,p_j}{\text{argmax}}\, d(p_i, p_j)$ such that p_i, p_j belong to contour.

3 Let be $\mathcal{L}_1$ a line which satisfies equation (1):

$$y = \frac{\Delta_y}{\Delta_x}x + \left(p_{iy} - \frac{\Delta_y}{\Delta_x}p_{ix}\right) \tag{1}$$

$$\Delta_x = p_{ix} - p_{jx}$$

$$\Delta_y = p_{iy} - p_{jy}$$

Compute $\underset{p_k,\mathcal{L}_1}{\text{argmax}}\, d(p_k, \mathcal{L}_1)$ such that p_k belong to contour.

4 Let be $\mathcal{L}_2$ a line which satisfies equation (2):

$$y = \frac{\Delta_x}{\Delta_y}(x - p_{kx}) + p_{ky} \tag{2}$$

5 **return** $\mathcal{L}_1$ and $\mathcal{L}_2$ as the new reference axes.

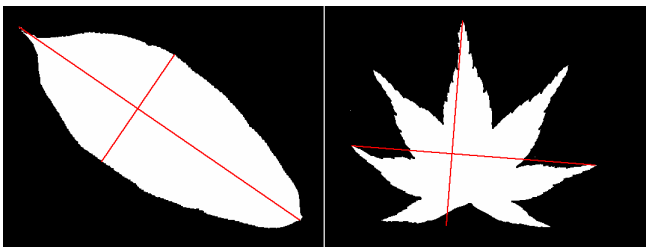


Fig. 5. Reference axes computed for a simple leaf (left), and a lobed leaf (right).

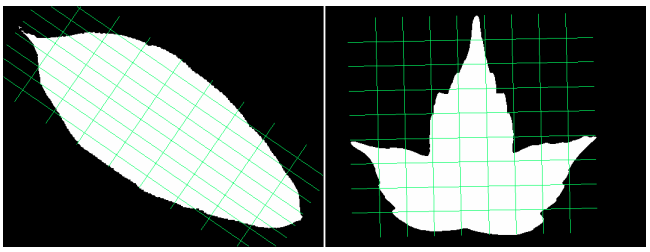


Fig. 6. Equidistant lines.

Algorithm 2: Method one to identify lobed from simple leaf.

Input: A binary image of leaf L, N: Number of lines, T: Threshold
Output: *0* if leaf is lobed, *1* if leaf is simple, *2* if the algorithm can not distinghish

```
Compute L1 and L2 using Algorithm 1
Draw N equidistant lines parallel to L1
Draw N equidistant lines parallel to L2
Count the number of changes along each line. Store it in an array A1 and
  A2 respectively;
Find the greatest number in array A1 and delete it;
Find the greatest number in array A2 and delete it;
Count the number of elements in A1 and A2 which are greater than T,
  call it W1 and W2
if W1 > T and W2 > T then
    return 0
else
    if W1 = 0 and W2 = 0 then
        return 2
    else
        return 1
```

Lines traced on the body of the leaf are based on reference axes. These axes are computed regardless the orientation of the image. It is important to say that our method does not vary when orientation changes. Likewise, changes of color along lines do not vary when scale changes. Our method is also invariant to scaling.

4 Experiments and Results

In this section, we present the results of experiments. We measured the capabilities of our proposals to identify lobed simple leaves. Our method was tested with the images in Flavia data set. Because in the literature there are not features specifically designed to identify lobed leaves, we do not compare the obtained results with others methods. Instead, we measure accuracy, specificity and sensitivity of the two introduced methods.

4.1 Detection of Lobed Simple and Unlobed Simple Leaves

Henceforth, our method will be referred as M_L. In order to measure the performance of M_L, we use Flavia data set [20]. In all our experiments we did not rotate or scale any image.

We manually identified the type of leaf and added a label (binary attribute) called Lobed to each leaf. The value of this attribute was set to true for the images of leaves of classes C6, C8, C28 and C32 (lobed simple leaves with smooth margins). For the rest of the leaves the value of the attribute was set to false (unlobed simple leaves). This identification is used to test the performance of our method.

The confusion matrix obtained for M_L is presented in Table 1. The positive cases correspond to lobed simple leaves, whereas the negative cases are the unlobed simple ones.

Table 1. Confusion matrix for M_L.

Prediction		
Unlobed	Lobed	**Type of leaf**
1,492	201	Unlobed
58	156	Lobed

Based on these last table, the following measures can be obtained:

- **Accuracy**: the proportion of the total number of predictions (positive and negative) that were correct.
- **Sensitivity or Recall**: the proportion of actual lobed leaves which are correctly identified.
- **Specificity**: the proportion of actual simple leaves which are correctly identified.

Table 2 shows the performances of our proposal.

Table 2. Performance of proposed method.

	Accuracy	**Sensivity or recall**	**Specificity**
M_L	86.42%	0.7290	0.8813

To measure the effect of our method in the performance of classification methods, we use 10-fold cross validation. Table 3 summarizes the classification accuracy achieved by each classification method. We observed that performances of classifiers are lower than those reported in the literature. However, in our experiments we only took into account six basic leaf features. This number is lesser and simpler than the used in many other works [7], [8,13,19]. Our goal is to compare basic leaf features with our proposal, as we consider it a basic leaf feature too.

The method which obtains the best performance is Multiclass classifier. This method decompose the multiclass problem into simpler ones, which are solved

Table 3. Effect of type of leaf (binary attribute) on seven classification methods.

Method	Classification accurary (%)	
	Six features	***With our method***
1 C4.5	58.78	**60.15**
KNN (K=1)	64.13	**64.24**
Random Forrest	65.44	**66.60**
Multi Class Classifier (Weka)	68.38	**71.95**
NN	**72.94**	72.93
Naive Bayes	55.06	**56.58**
Random Tree	56.63	**57.63**

with logistic regression. The method with second best performance is Random forest. This method uses 100 trees trained with a subset of attributes, and then uses a mechanism of votes to make predictions.

5 Conclusions

Many classic plant identification methods use dichotomous keys that take into account specific features of leaves, such as aspect ratio, leaf area, area convexity, diameter, among others. Motivated by this, we designed a new method to discriminate automatically between unlobed simple and lobed simple leaves. Our method detects changes between background and leaf (and vice versa) in binary images, previously obtained from color images. The unlobed simple and lobed simple leaves feature is an auxiliary characteristic to classification methods, since is independent of rest of leaf features, which when combined achieve better classification results.

The principal characteristic of our method is that it is invariant to rotation and scale of images, because we find a new axes in the leaf. With this axes all the operations can be defined as in linear algebra, line equation, parallel and orthogonal lines equations.

Currently, we are working on designing new compound-leaf features to detect the number of leaflets, also, we are improving basic leaf features to apply in real-world scenarios with challenging conditions.

Acknowledgements. Authors thank to Universidad Autónoma del Estado de México for all the facilities.

References

1. Aakif, A., Khan, M.F.: Automatic classification of plants based on their leaves. Biosystems Engineering 139, 66 – 75 (2015), http://www.sciencedirect.com/science/article/pii/S1537511015001373

2. Baker, R., Dengler, H.: Leaf Key to Common Trees in Maryland. Extension bulletin, Cooperative Extension Service, University of Maryland (1970), https://books.google.com.mx/books?id=DbosAQAAMAAJ
3. Bama, B.S., Valli, S.M., Raju, S., Kumar, V.A.: Content based leaf image retrieval (cblir) using shape, color and texture features. Indian Journal of Computer Science and Engineering 2(2), 202–211 (2011)
4. Berretti, S., Del Bimbo, A., Pala, P.: Retrieval by shape similarity with perceptual distance and effective indexing. Multimedia, IEEE Transactions on 2(4), 225–239 (2000)
5. Cerutti, G., Tougne, L., Coquin, D., Vacavant, A.: Leaf margins as sequences: A structural approach to leaf identification. Pattern Recognition Letters 49, 177 – 184 (2014), http://www.sciencedirect.com/science/article/pii/S0167865514002335
6. Cerutti, G., Tougne, L., Mille, J., Vacavant, A., Coquin, D.: Understanding leaves in natural images–a model-based approach for tree species identification. Computer Vision and Image Understanding 117(10), 1482–1501 (2013)
7. Chaki, J., Parekh, R., Bhattacharya, S.: Plant leaf recognition using texture and shape features with neural classifiers. Pattern Recognition Letters 58, 61 – 68 (2015), http://www.sciencedirect.com/science/article/pii/S0167865515000586
8. Di Ruberto, C., Putzu, L.: A fast leaf recognition algorithm based on svm classifier and high dimensional feature vector. In: Computer Vision Theory and Applications (VISAPP), 2014 International Conference on. vol. 1, pp. 601–609. IEEE (2014)
9. Dozier, H., Mills, R.: Leaf key to common trees in Louisiana. http://www.lsuagcenter.com/NR/rdonlyres/BA8FFA18-B7CD-4D98-88FF-AF234D5F9ACD/18437/pub1669LeafKey.pdf (Dec 2016), LSU AGCenter
10. Ehsanirad, A.: Plant classification based on leaf recognition. International Journal of Computer Science and Information Security 8(4), 78–81 (2010)
11. Gwo, C.Y., Wei, C.H., Li, Y.: Rotary matching of edge features for leaf recognition. Computers and Electronics in Agriculture 91, 124 – 134 (2013), http://www.sciencedirect.com/science/article/pii/S0168169912002906
12. Harish, B., Hedge, A., Venkatesh, O., Spoorthy, D., Sushma, D.: Classification of plant leaves using morphological features and zernike moments. In: Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on. pp. 1827–1831 (Aug 2013)
13. Kalyoncu, C., Toygar, O.: Geometric leaf classification. Computer Vision and Image Understanding 133, 102 – 109 (2015), http://www.sciencedirect.com/science/article/pii/S1077314214002124
14. Newcomb, L.: Newcomb's Wildflower Guide. Little, Brown (1989), https://books.google.com.mx/books?id=7DBvQgAACAAJ
15. Oregon State, U.: Dichotomous key. http://oregonstate.edu/trees/dichotomous_key.html (2016), consulted 20-04-2016
16. Qi, H.N., Yang, J.G.: Sawtooth feature extraction of leaf edge based on support vector machine. In: Machine Learning and Cybernetics, 2003 International Conference on. vol. 5, pp. 3039–3044 Vol.5 (Nov 2003)
17. Randall, D.J.: Trees of iowa: An interactive key. http://www.extension.iastate.edu/forestry/iowa_trees/key/key.html (December 2016), Iowa State University

18. Vijayalakshmi, B.: A new shape feature extraction method for leaf image retrieval. In: Proceedings of the Fourth International Conference on Signal and Image Processing 2012 (ICSIP 2012). pp. 235–245. Springer (2013)
19. Wang, B., Brown, D., Gao, Y., Salle, J.L.: March: Multiscale-arch-height description for mobile retrieval of leaf images. Information Sciences 302, 132 – 148 (2015), http://www.sciencedirect.com/science/article/pii/S0020025514007282
20. Wu, S.G., Bao, F.S., Xu, E.Y., Wang, Y., Chang, Y.F., Xiang, Q.L.: A leaf recognition algorithm for plant classification using probabilistic neural network. CoRR abs/0707.4289 (2007), http://dblp.uni-trier.de/db/journals/corr/corr0707.html#abs-0707-4289

ISSN 1870-4069

Generación de un vector característico para la detección de intrusos en redes computacionales

Ana Alcántara-Ramírez, Lourdes López-García, Juvenal Rueda Paz

Universidad Autónoma del Estado de México,
Centro Universitario UAEM Valle de Chalco, Estado de México,
México

aalcantarar@alumno.uaemex.mx,mllopezg@uaemex.mx, jruedap@uaemex.mx

Resumen. El control de acceso no autorizado en redes informáticas es un problema que inicia desde el surgimiento de los sistemas de información computarizados, donde la seguridad y la privacidad de la información son factores importantes. Una solución conveniente para resolver este problema es el uso de un Sistema de Detección de Intrusos (IDS, por sus siglás en inglés). La eficiencia de un IDS está determinada por la certeza en la detección, misma que depende de una correcta clasificación, que tendrá lugar si se cuenta con un vector que contenga las características adecuadas del objeto o entidad a clasificar. En este artículo, se propone la generación de un vector característico a partir de información real proveniente de la red que permita realizar una correcta interpretación sobre el comportamiento de los procesos habituales para los cuales la red fue creada, para así, discernir entre lo autorizado y no permitido en la red. Para comprobar la eficacia de la detección se utilizan 5 clasificadores incluido en ellos una red neuronal y árbol de decisión. Así, la certeza de una evaluación precisa de la red, permitirá protegerla de usuarios maliciosos que intenten invadirla sin ser detectados.

Palabras clave: ataques de intrusión, algoritmos de clasificación, vector característico.

Development of a Characteristic Vector for the Intrusion Detection in Computational Networks

Abstract. Unauthorized access detection in a computer networks is a problem that starts from the beginning of computerized information systems, where the security and privacy of the information are important factors. A good solution to solve this problem is the use of an Intrusion Detection System (IDS). The efficiency of an IDS is measured by the precision in the detection which depends on the an accurate classification, that can be possible, using a vector with the appropriate characteristics of the object or entity to be classified. In this paper, the generation of a

characteristic vector based on real information coming from the network, is proposed. The vector allows classifiers to do a correct interpretation of the behavior of the common processes for which the network was created, in order to discern between what is authorized or non authorized on the network. To verify the effectiveness of the detection, five classifiers are used, including a neural network and decision tree. Thus, the certainty of an accurate evaluation of the network, will protect it from malicious users who try to invade it undetected.

Keywords: intrusion attacks, classification algorithms, characteristic vector.

1. Introducción

La importancia de la comunicación radica en la necesidad de compartir información entre entidades. El canal de transmisión de los datos puede ser público o privado, sin embargo, en ambos casos debe proveerse eficiencia y seguridad ante información importante o secreta. Las redes computacionales son un medio de comunicación que permite compartir información a grandes distancias de manera rápida, fácil y en diferentes formatos.

La mayoría de los usuarios puede acceder a las redes de computadoras, que les permite tener una comunicación desde su ubicación hacia cualquier punto donde la red mantenga conexión. Utilizar este canal, sobre todo si es público, implica tener conocimiento sobre cómo usuarlo, pero no necesariamente sobre cómo funciona. Ésta diferencia, hace que las entidades maliciosas se aprovechen de los usuarios ingenuos para vulnerar el canal de comunicación usado y lograr ataques como una intrusión no deseada en la red.

Para proteger los datos de quienes usan estos canales, se implementan protocolos de seguridad, así como, la aplicación de métodos y herramientas especializadas para ciertas tareas. Los Sistemas de Detección de Intrusiones son útiles en la búsqueda de la seguridad, brindando un medio de detección ante una intrusión no autorizada en una red. En esta herramienta se incluyen métodos para el tratamiento y análisis de los datos que se transmiten, tales como la minería de datos, los algoritmos genéticos, la inteligencia artificial, entre otros. De tal manera, que sea posible clasificar el tipo de tráfico y decidir cuándo se presenta o no, un ataque de intrusión [1].

Contar con procedimientos eficientes de clasificación de patrones es esencial en muchas aplicaciones de gran interés. Una de ellas es el diseño de Sistemas de Detección de Intrusos en sistemas de cómputo. Un factor fundamental para alcanzar la eficiencia en procesos de clasificación, es la ejecución previa de técnicas de selección y la extracción de características sobre el conjunto de datos. Lo cual no sólo mejora la precisión de la clasificación, también mejora la capacidad de generalización en el caso de la clasificación supervisada, o contrarresta el riesgo de una mala clasificación que puede presentarse al usar clasificadores no supervisados.

Para desarrollar un sistema clasificador es necesario determinar el conjunto de características que describan la arquitectura de los datos con que se trabaja. La razón de identificar la elección de un subconjunto adecuado de características, es que permite la reducción de la dimensionalidad en el conjunto de datos, lo que contribuye a disminuir la complejidad computacional de la clasificación, mejorando el rendimiento del clasificador y evitando características redundantes o irrelevantes. Aunque la selección de características se puede definir formalmente como un problema de optimización con un solo objetivo, (es decir, la precisión de la clasificación obtenida usando el subconjunto de características seleccionadas), en los últimos años, se han propuesto algunos enfoques multi-objetivo para este problema.

En este trabajo se propone la generación de un vector característico como base fundamental para ser utilizado en un IDS. Dicho vector está conformado por datos reales obtenidos de una red. El escenario propuesto consta de la puesta en marcha de una red, diseñada para analizar su comportamiento en estado normal y bajo ataque por inundación de paquetes. De tal manera que sea posible generar un vector característico con información proveniente de una comunicación cotidiana en la red o en otro caso, bajo un ataque que tiene como objetivo introducir paquetes a la red para saturarla. El resultado es un vector con características distinguibles para cada escenario, lo que implica una clasificación correcta en la toma de decisión sobre si es un ataque o no. Así, la principal contribución de este trabajo es un enfoque empleado para la selección de características y su aplicación a un enfoque supervisado.

Para garantizar la efectividad de nuestra propuesta, los vectores caracteríticos generados son clasificados en una red neuronal y en los algoritmos J48, Random-Forest, Naive Bayes y Decision table, obteniendo una matríz de confusión con un porcentaje mínimo de falsos positivos.

El resto del artículo se organiza de la siguiente forma. En la sección 2 se presentan los conceptos básicos y definiciones necesarias respecto a los IDS y los algoritmos de clasificación. La sección 3 presenta el estado del arte en los trabajos relaciones a la propuesta en este artículo. Posteriormente, en la sección 4 se describe el escenario propuesto y se presenta un análisis de su comportamiento bajo los dos escenarios mencionados. En la sección 5 se detalla la forma en cómo es generado el vector característico y se muestra la efectividad de su clasificación, en la sección 6. En la sección 7 se realiza una discusión de los resultados obtenidos y una comparativa con los trabajos relacionados. Finalmente, en la sección 8, se presentan las conclusiones de este trabajo.

2. Preliminares

Esta sección comprende los conceptos básicos que intervienen y están relacionados con el desarrollo del objetivo principal, que es la creación de un vector característico. Iniciamos con una descripción de los IDS para comprender mejor la aplicación que se buaca alcanzar y la importancia que tiene lograrlo.

Posteriormente, se presenta una breve explicación sobre las características de los algoritmos para el análisis de los datos que componen el vector característico.

2.1. Sistemas de detección de intrusiones (IDS)

Los IDS son herramientas que escuchan el tráfico de la red y son capaces de detectar actividades inusuales, para así, reducir el riesgo de una intrusión no permitida. Los IDS pueden evaluar la información en tiempo real, o que esté contenida en una base de datos. Existen varias clasificaciones para los IDS, el denominado HIDS es un sistema de detección de intrusiones basado en host, que tiene como objetivo identificar ataques con base en la observación de los encabezados de los paquetes, para detectar a una entidad que intenta violar o modificar la seguridad del host. Por otra parte, el NIDS que es un sistema de identificación de intrusiones de red y su detección que se basa en el análisis de los paquetes de red y de los protocolos que se emplean para la transmisión de los datos, ambos pueden ser en tiempo real o no [2].

La Tabla 1 muestra los enfoques que puede tener un IDS, de acuerdo al modo de detección que maneja: comportamiento, firmas, anomalias o heurístico [3,4,5,6].

2.2. Algoritmos de clasificación

El proceso de clasificación es uno de los más útiles y comunes en el tratamiento de datos, ya que permite analizar el comportamiento de una o más variables dentro de un conjunto de información. Dicho conjunto es formado por datos agrupados y dependientes del atributo al que pertenecen, los datos son sometidos al sistema clasificador para así, determinar a qué clase corresponden. Los clasificadores requieren una fase de entrenamiento o construcción de la base de conocimientos [7,8]. En este trabajo, se utilizaron cinco clasificadores que se describen a continuación:

- **Red neuronal**: Se compone de varias neuronas (unidad mínima de procesamiento de la información, representa un dato de entrada) que están divididas en varias capas. Las neuronas de una capa se conectan con las neuronas de la capa siguiente y les pasan información. La arquitectura consiste en una capa de entrada que recibe la información del exterior; capas intermedias (ocultas) que realizan el trabajo de la red y una capa de resultados que muestra los resultados de la última capa intermedia [8,9].
- **Algoritmo J48**: Se deriva del algoritmo C4.5 y para la clasificación crea un árbol binario [10]. Se basa en la utilización del criterio *ratio de ganancia* (gain ratio) para evitar que las variables con mayor número de presencia salgan beneficiadas en la selección. Además, el algoritmo incorpora una poda del árbol una vez que éste ha sido inducido [11].
- **Random Forest**: Emplea una selección aleatoria de atributos y genera un conjunto de árboles predictores que serán evaluados posteriormente [12].

Tabla 1. Clasificación de los IDS.

Enfoque	Descripción
Comportamiento	*Funcionalidad*: busca variaciones de costumbres, como un tráfico elevado. *Ventajas*: método simple y efectivo para detectar ataques conocidos. Detalla el análisis contextual. *Desventajas*: inefectivo para ataques no conocidos, o variantes de los conocidos. Difícil mantener las firmas y patrones actualizados. Requiere mucho tiempo para aprender.
Firmas o MD-IDS	*Funcionalidad*: clasifica ataques con base en firmas y auditorias. *Ventajas*: efectivo para detectar vulnerabilidades nuevas, es menos dependiente del sistema operativo y puede detectar el abuso de privilegios. *Desventajas*: los perfiles cambian constantemente y no es efectivo en su reconstrucción.
Anomalías o AD-IDS	*Funcionalidad*: busca elementos fuera de lo común, se centra en patrones de tráfico. *Ventajas*: está basado en comportamiento de protocolos de red, detecta secuencias anormales de comandos. *Desventajas*: no distingue ataques que simulen un comportamiento usual en los protocolos y puede ser incompatible con algunos navegadores.
Heurístico	*Funcionalidad*: emplea algoritmos para analizar el tráfico que pasa por la red. *Ventajas*: puede predecir eventos y ser autodidacta, distingue secuencias de comando. *Desventajas*: consume muchos recursos y es de funcionamiento complejo.

Cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos. Es una modificación sustancial de bagging que construye una amplia colección de árboles no correlacionados y promediados posteriormente [13].

- **Naive Bayes**: Asume que la presencia o ausencia de una característica particular no está relacionada con la presencia o ausencia de cualquier otra característica, variable, tabulador, parámetro o atributo. Se evalúan de modo independiente sin establecer relaciones o coincidencias. Se puede entrenar en un ambiente de aprendizaje supervisado con pocos datos, obteniendo las medias y las varianzas de las variables necesarias para la clasificación. Debido a que las variables independientes se asumen, solo es necesario determinar las varianzas de las variables de cada clase y no toda la matriz de covarianza [14,15].
- **Decision Table**: Llamada DTM (Decision Table Majority), se compone de un conjunto de características que se incluye en la tabla atributos y por instancias etiquetadas (reglas). En su procesamiento cada dato de entrada

se asigna a la clase con la que ha tenido mayor número de correspondencias. De esta forma, a partir de un dato no etiquetado el clasificador busca correspondencias de este dato de entrada con el total de reglas para todos los atributos. Si no se encuentra alguna correspondencia, la tabla DTM asigna el dato a la clase mayoritaria [16,17].

3. Estado del arte

La información generada de procesos de red es cuantiosa, y tiende a crecer en cuanto la arquitectura de la red y los servicios que proporcionan se incrementan. Garantizar la seguridad de esta información obliga a buscar mejores herramientas. La base de un correcto funcionamiento de estas herramientas y su efectividad depende de lo certero que sea el juicio generado, es decir, la capacidad de distinguir correctamente el flujo que circula para así discernir entre lo permitido y lo no permitido.

Dentro de la literatura que contiene los avances en este tipo de trabajos podemos encontrar que siguen distintas vertientes, algunos apuestan por la variación de clases para evaluar correctamente, otros autores mencionan la necesidad de acotar las variables a evaluar y ser más mesurados en la cantidad de clases con las que se trabaja. Otro aspecto que se considera es el enfoque de aprendizaje supervisado, no supervisado, semi-supervisado, entre otros; que garanticen una mejor evaluación de las relaciones entre los datos.

Siguiendo con estos puntos, acciones tales como considerar un pre-procesado en los datos, proponer algoritmos que combinen algoritmos ya existentes, trabajar con bases de conocimientos previamente generadas o proponer el manejo de datos reales, se enfocan a conseguir la muestra apropiada y el evaluador o clasificador preciso que brinde confianza en su predicción.

En [18] analizan el conjunto de datos denominado KDD99 que cuenta con 41 atributos distintos, de los cuales, se seleccionaron 23 para su clasificación. El entrenamiento se realizó con el 10 % de los 51 millones de instancias contenidas en la base de datos, aplicándoles tres variantes de preprocesamiento, para después hacer una comparación basada en el uso de algoritmos representativos del aprendizaje automático. Entre éstos algoritmos se encuentran, una Red Neuronal Perceptron Multicapa (MLP), SMO que es una variante empleada en WEKA del algoritmo de Máquinas de Soporte Vectorial (SVM), el algoritmo J48, Naive Bayes y el algoritmo basado en instancias K con valores 3, 5 y 7.

Los resultados presentados arrojaron porcentaje del 98.14 % para Naive Bayes y un 99.02 % para J48, siendo éste el más preciso. Para la variante de preprocesado 2 se tiene a J48 con un 97.43 % ante lo obtenido con SMO con un 99.23 %. Finalmente, los resultados con la variante 3 presenta al algoritmo J48 con 95.85 % y MLP con 98.4 %.

El [19] se propone un sistema de identificador de intrusiones que use un clasificador basado en aprendizaje semi-supervisado. Los algoritmos empleados para el tratamiento de los datos son J48, Naive Bayes, NB tree, Random Forest, Random tree, Red Neuronal y SVM. Se implementan dos variantes que consisten

en el uso de la base de conocimiento KDDCUP99 con los 41 atributos que la integran y una variación de esta base compuesta por 21 atributos. Los resultados de certeza en la clasificación son SVM con 69.52 % que presenta los valores más bajos con un 42.29 % para la segunda variante de la base de conocimiento que cuenta con 21 atributos.

En [20] se encuentran tres variantes propuestas con distinto enfoque para evaluar los datos, previo al proceso del clasificador. Los datos son obtenidos de las bases KDD99 y Gure KDD. Tiene 6 posibles clases que representan 5 ataques a la red y una clase que describe un comportamiento normal. Dentro de los ataques que se incluyen en el evaluador, se encuentra el de Denegación de Servicio, además, de un algoritmo para clasificar las clases.

Los resultados porcentuales de certeza, se presentan en una tabla que los divide en los tres enfoques de clasificación trabajados. Para el método de clasificación 1 aplicada a 8 algoritmos se obtuvieron los siguentes resultados, 80.67 % para Random Forest y 99.21 % para una variante de su algoritmo propuesto. El segundo método de clasificación se aplicó a tres algoritmos donde su propuesta obtuvo 82.10 % frente al 96.5 % de Naive Bayes. Por último, la tercera clasificación, aplicada a 4 algoritmos, presentaron los valores 98.38 % para Decision Tree Based y 99.27 % para el algoritmo de su propuesta, con lo cual, garantizan una clasificación precisa.

El trabajo realizado en [21] recomienda el uso de Mobile Ad hoc redes (MANET) para la protección de redes inalámbricas. Emplea un modelo probabilístico que tiene la finalidad de reducir los tiempos activos del IDS, centrandose en la conexión realizada entre dispositivos al inicializar un juego. El juego es cooperativo y multijugador que analiza los efectos de un IDS con una actividad reducida en la red. El sistema funciona en redes estáticas y móviles. El algoritmo empleado es LDK para la detección de los vecinos cercanos o jugadores introducidos en el juego. El enfoque principal, entonces, no es diseñar un IDS, más bien es presentar un esquema para un uso eficiente que determina el ahorro de energía de los dispositivos mientras el IDS se ejecuta.

4. Escenario propuesto

La Fig. 1 muestra el escenario definido para el análisis de tráfico. Como puede observarse, consiste de dos redes conectadas a través de un ruteador. La red 2 se compone de un servidor web, el cual será atacado, y varios host, mientras que la red 1 contiene por lo menos una computadora que fungirá como adversario.

El ataque aplicado a la red definida es el de denegación de servicio (DoS), el cual busca la interrupción del flujo de datos y reduce la disponibilidad que otorga un servicio activo. El modo en que opera consiste en enviar paquetes con formato permitido en grandes cantidades para lograr la saturación del servidor, de tal manera que ya no le sea posible atender las solicitudes. Para lograr la saturación del servidor web se transmitieron paquetes del protocolo ICMP con carga elevada. Es importante mencionar que, el interés en este artículo es detallar una solución al ataque de DoS por inundación y no describir cómo se efectúa.

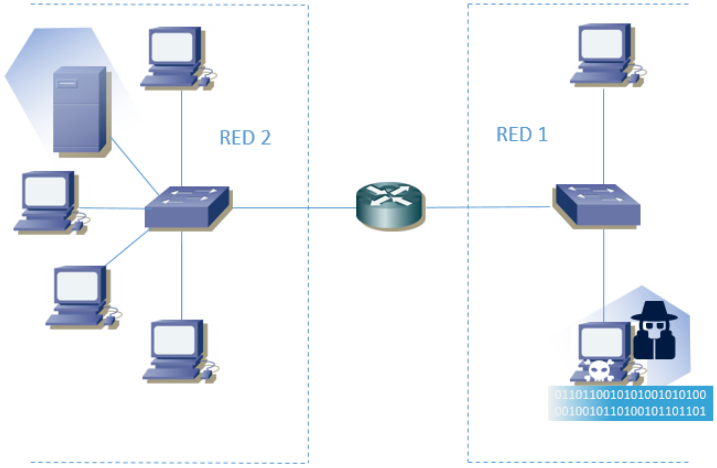


Fig. 1. Topología de la red.

La información resultante del ataque es obtenida a través de Whireshark, que es una aplicación para la escucha de la red, que nos permite guardar en un archivo de formato de texto, el cual será utilizado para obtener el vector característico.

5. Creación del vector característico

Los IDS protegen a un conjunto de computadoras de usuarios no autorizados, incluyendo, posiblemente, a entidades internas. Durante el periodo de entrenamiento, el IDS construye un modelo predictivo (un clasificador) capaz de distinguir entre las conexiones normales y las conexiones anormales, denominadas intrusiones o ataques. Como se mostró en la Fig. 1, el escenario establecido contempla dos redes que mantienen comunicación a través de un ruteador. El paso de información entre ambas redes, es generada para evaluar el tráfico en los estados *normal* y de *ataque*.

Los datos transmitidos en la conexión son generados aleatoriamente. Llamamos conexión a una secuencia de paquetes que fluyen desde una dirección IP en la red 1 hacia una dirección IP de la red 2, bajo algún protocolo bien definido como TCP, ICMP, entre otros. La captura del tráfico de red se tomo con un periodo de 5 segundos para cada lectura realizada. Se obtuvieron 45 lecturas de la red en estado normal y 45 en estado de ataque.

Ya que es un ambiente controlado, cada conexión es etiquetada ya sea como *normal* o como *ataque* y los registros resultantes tienen un tamaño entre 7.5KB y 2.5MB. Como la finalidad de este trabajo es distinguir cuándo es un ataque o no, las lecturas son divididas en dos clases: a0 y a1, respectivamente.

Como es de esperarse, las lecturas indican información proveniente de la red: IP origen, IP destino, tipo de puerto, descripción de la tarea como solicitud, acuse, transmisión del paquete, etc., lo que comúnmente se llama el volcado TCP sin procesar para una red de área local (LAN). Esta información no puede ser ingresada a los clasificadores, tal cual se encuentra en el registro, por lo que es necesario procesarla y obtener las características del archivo resultante de la lectura.

En la tabla 2 se muestran los atributos que conforman el vector característico, que están divididos en la información de la red como las IP de origen y la de destino; el tiempo de lectura (5 segs) y la información contenida en el archivo de registro como el número de patrones totales, número de patrones distintos, densidad léxica, etc. La tabla indica el tipo de información, si sus valores son constantes o variables.

El resultado es un vector característico de 17 elementos, que son normalizados y clasificados como $a0$ (estado normal) y $a1$ (estado de ataque).

Tabla 2. Lista de atributos.

Atributo	Descripción	Valor
a0	Ip origen	constante
a1	Ip origen	constante
a2	Ip origen	constante
a3	Ip origen	constante
a4	Ip destino	constante
a5	Ip destino	constante
a6	Ip destino	constante
a7	Ip destino	constante
a8	Tiempo de transmisión	constante
a9	Patrones totales	variable
a10	Número de patrones distintos	variable
a11	Densidad léxica	variable
a12	Total de sentencias	variable
a13	Longitud promedio comando	variable
a14	Longitud máxima de comando	constante
a15	Longitud mínima de comando	constante
a16	Legibilidad1	variable
a17	Legibilidad2	variable

6. Resultados de la clasificación

Para garantizar la efectividad del vector característico propuesto, se utilizaron los clasificadores descritos en la sección 2.2. De la información resultante de los clasificadores, tomamos al porcentaje de clasificaciones correctas, el tiempo

de ejecución y la matriz de confusión como ponderadores de los vectores caracteristicos puestos a prueba.

La Fig. 2 muestra la matriz de confusión para cada clasificador. Su interpretación es a través de la diagonal donde se muestra que los datos pertenecen a una clase (para este estudio clase a y b), separándolos en los que fueron clasificados correctamente y cuales incorrectamente. Como puede observarse, las diagonales de cada matriz muestran que los clasificadores realizaron una distribución correcta de la muestra, de acuerdo a la clase establecida por cada vector. Por ejemplo, para la Red Neuronal se tiene que de los 44 registros pertenecientes a la clase a todos fueron clasificados correctamente, ya que ninguno se catalogó como clase b, en tanto, para los 46 registros que pertenecian a la clase b, 45 fueron clasificados correctamente, es decir, sólo hubo un falso positivo para la clase a.

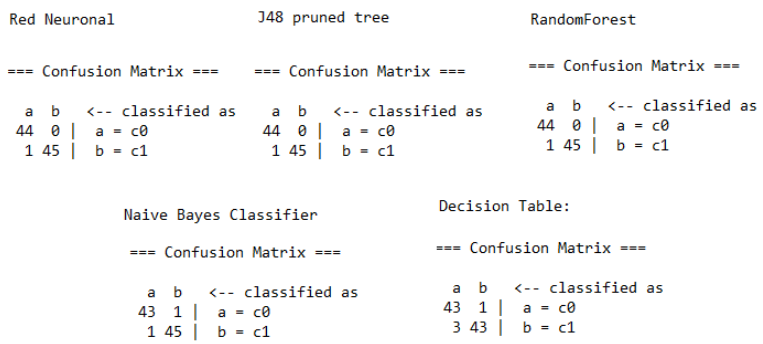


Fig. 2. Matriz de confusión de los resultados de cada clasificador.

La Tabla 3 presenta los valores restantes de la clasificación, referente a los porcentajes de la clasificación correcta y el tiempo de ejecución del clasificador en segundos. El algoritmo de *Decision table* fue el que reportó menor precisión, sin embargo, aún es un porcentaje elevado de eficacia. Por otro lado, el algoritmo J48 es el que reporta más eficiencia y eficacia, para este caso de estudio.

Tabla 3. Resultados obtenidos por los clasificadores que tienen como entrada el vector caraterístico propuesto.

Clasificador	Instancias Correctas	Tiempo de ejecución
Red Neuronal	98.8 %	0.21 segs.
J48	98.8 %	0.02 segs.
Random Forrest	98.8 %	0.07 segs.
Naive Bayes	97.7 %	0.001 segs.
Decision Table	95.5 %	0.04 segs.

7. Discusión

Los resultados que los clasificadores reportan al usar el vector característico propuesto, reportan porcentanjes elevados de precisión. Para hacer un análisis de su efectividad, es necesario realizar una comparación con respecto a los trabajos publicados en el estado del arte.

Es importante no perder de vista, que la información de comparación tiene variación para cada trabajo publicado, considerando el enfoque de aprendizaje empleado, los ataques de intrusión a los que está dirigido el detector, la base de datos de conocimiento usada, los atributos y las clases empleadas o si es supervisado o no. Por lo anterior, primero presentamos una lista de las coincidencias y otra de las diferencias, para que con ello, se tomen los elementos más importantes para la comparación.

- Diferencias:
  1. El origen de los datos empleados en la generación del vector característico propuesto proviene de información real de la red, mientras que los trabajos reportados en el estado del arte utilizan, principalmente, la base de datos KDD99. El número de variables
  2. Los trabajos relacionados usan multivariables con 23 atributos, en contraste con nuestra propuesta que tiene sólo 1 con 17 atributos.
  3. Los ataques de intrusión son diferentes, en este trabajo nos enfocamos al de denegación de servicio.
- Similitudes de comparación
  1. La precisión con la que el clasificador reporta resultados.
  2. Los algoritmos de clasificación J48, Redes neuronales y Naive Bayes y Randiom Forest.
  3. La variación más precisa reportada en cada trabajo relacionado.

La Tabla 5 presenta la relación entre los algoritmos empleados y el porcentaje de precisión que presentan en la clasificación. Como se puede apreciar, nuestra propuesta reporta un elevado porcentaje de precisión en todos los clasificadores, muy cercano a lo propuesto en [8] y en contraste con [19], que reporta un 81.05 % con el algoritmo J48, sin embargo, con el algoritmo Naive Bayes cae hasta el 76.56 %. Finalmente, en la Tabla 4 se presenta el tipo de información utilizada, en donde, claramente se identifica que el vector característico propuesto usa información obtenida de un escenario de ataque real a la red y en estado normal, mientras que los restantes, se apoyan de información generalizada, almacenada en la base de datos DKK99 y sus variaciones.

8. Conclusiones

En este artículo se presenta un método de clasificación para la detección de intrusos en una red, particularmente, del ataque de denegación de servicio, provocado por inundación de paquetes. La propuesta consta de la generación de un vector caraterístico a partir de información obtenida de una red. Este método

Tabla 4. Tabla de comparación en el porcentaje de precisión.

	Red Neuronal	Alg. J48	Naive Bayes	Random Forest
Rivero [18]	98.52 %	99.02 %	98.14 %	NA
Ashfaq [19]	77.41 %	81.05 %	76.56 %	80.67 %
Zhu [20]	NA	NA	76.56 %	80.67 %
Vector Propuesto	98.8 %	98.8 %	97.7 %	98.8 %

Tabla 5. Tabla de comparación usando tipos de muestra y enfoque.

	Muestra	Enfoque
Rivero [18]	KDD99	Supervisado
Ashfaq [19]	KDDCUP99	Semi-supervisado
Zhu [20]	KDD99 y Gure KDD	Supervisado
Vector Propuesto	Lectura directa de la red	Supervisado

permite diferenciar entre los registros provenientes de un ataque y los registros provenientes de un flujo normal.

El vector característico está compuesto por 17 atributos que corresponden a información propia de la red como las IP de origen y destino, el tiempo de captura y toda la información que participa en la transmisión, de tal manera, que permita una clasificación correcta.

Para comprobar que la información contenida en el vector característico permite distinguir claramente un ataque o no, se probó en 5 clasificadores, tales como, una red neuronal, el algoritmo J48 y el Naive Bayes, entre otros. Los resultados reportados por los clasificadores indican que el vector característico permite una categorización precisa y eficiente.

En comparación con los trabajos relacionados, nuestra propuesta consigue obtener un porcentaje elevado de precisión con diferentes tipos de clasificadores, usando información real proveniente de la red, en un enfoque supervisado, al estar constituido con únicamente 17 atributos para su caracterización.

Referencias

1. Debar, H., Becker, M., Siboni, D.: A neural network component for an intrusion detection system. In: Proceedings on Research in Security and Privacy. IEEE Computer Society Symposium, pp. 240–250 (1992)
2. Horng, S., Su, M., Kao, T., Chen, R., Lai, J., Perkasa, C.: A novel intrusion detection system based on hierarchical clustering and support vector machines. Expert Systems with Applications, 38(1), pp. 306–313 (2011)
3. Liao, H., Lin, C., Lin, Y., Tung, K.: Intrusion detection system: A comprehensive review. Journal of Network and Computer Applications, 36(1) pp. 16–2 (2012)
4. Capraru, C.: Detección de anomalías HTTP trazando la sesión web de un usuario. Tesis de Maestría en Seguridad de las Tecnologías de la Información y las

Comunicaciones (MISTIC), Universidad Oberta de Catalunya, España, pp. 1–30 (2016)
5. Rivero, J.: Técnicas de aprendizaje automático para la detección de intrusos en redes de computadoras. Revista Cubana de Ciencias Informáticas, 8(4), pp. 52–73 (2014)
6. Diaz, G., Flores, R., Silva, V.: Sistema Monitor Detector de Intrusos usando TRIPLE-DES96. Tesis de Maestría en Tecnología de Cómputo, Centro de Innovación y Desarrollo Tecnológico en Cómputo, Instituto Politécnico Nacional, México (2014)
7. Flores, J., Lara, Pedro., Gutierrez, M., De los Cobos Silva, S. , Rincón, E.: Un sistema clasificador utilizando coloración de gráficas suaves. Revista de Matemática: Teoría y Aplicaciones, 24(1), pp. 129–156 (2017)
8. Silva, E., Chura, E.: Clasificación de dígitos manuscritos de imágenes digitales. Revista Ciencia & Desarrollo, 1(19), pp. 61–67 (2017)
9. Rubio, J., Hérnandez-Aguilar, J., Stein-Carrillo, J., Ávila-Camacho, F., Meléndez-Ramírez, A.: Sistema sensor para el monitoreo ambiental basado en redes Neuronales. Ingeniería, Investigación y Tecnología, 17(2), pp. 211–222 (2016)
10. Patil, T., Sherekar, S.: Performance analysis of Naive Bayes and J48 classifiation algorithms for data classification. International Journal of Computer Science and Applications, 6(2), pp. 256–261 (2013)
11. Salazar, C.: Generación de Modelos Predictivos de Satisfacción Transaccional para un Centro de Atención a Clientes. Tesis de Maestría en Ciencias Computacionales con Especialidad en Redes y Seguridad Informática, Tecnológico de Monterrey Campus Estado de México, México (2016)
12. Bai, S.: Growing random forest on deep convolutional neural networks for scene categorization. Expert Systems with Applications, 71(1), pp. 279–28 (2017)
13. Tang, F., Ishwaran, H.: Random forest missing data algorithms. Journal arXiv preprint, eprint: 1701.05305 (2017)
14. García, A., Camacho, O., Yáñez, C.: Clasificador de Heaviside. Nova scientia, 7(14), pp. 365–397 (2015)
15. Krishnan, D., Balasubramanian, K.: A Fusion of Multiagent Functionalities for Effective Intrusion Detection System. Security and Communication Networks, 2017(1), pp. 1–15 (2017)
16. Berdun, F., Armentano, M., Amandi, A.: Inferencia de roles de equipo a partir de conductas colaborativas detectadas en 5 interacciones textuales. En: Simposio Argentino de Inteligencia Artificial (ASAI 2016), Buenos Aires, Argentina, Febrero 3, pp. 78–85 (2016)
17. Univaso, P., Ale, J.,Gurlekian, J.: Data Mining applied to Forensic Speaker Identification. IEEE Latin America Transactions, 13(4), pp. 1098–1111 (2015)
18. Rivero Pérez, J. L., Ribeiro, B., Ortiz, K. H.: Comparación de algoritmos para detección de intrusos en entornos estacionarios y de flujo de datos. Universidad y Sociedad, 8(4). pp. 32–42 (2016)
19. Ashfaq, R. A. R., Wang, X. Z., Huang, J. Z., Abbas, H., He, Y. L.: Fuzziness based semi-supervised learning approach for intrusion detection system. Information Sciences, 378(1), pp. 484–497 (2017)
20. Zhu, Y., Liang, J., Chen, J., Ming, Z.: An improved NSGA-III algorithm for feature selection used in intrusion detection. Knowledge-Based Systems, 116(1), pp. 74–85 (2017)
21. Marchang, N., Datta, R., Das, S.: A Novel Approach for Efficient Usage of Intrusion Detection System in Mobile Ad Hoc Networks. IEEE Transactions on Vehicular Technology, 66(2), pp. 1684–1695 (2017)

ISSN 1870-4069

Método de fusión de datos de fuentes heterogéneas para mantener la consistencia de datos

Julio Muñoz[1], Guillermo Molero–Castillo[1,2], Edgard Benítez–Guerrero[1]

[1] Universidad Veracruzana, Xalapa, Veracruz, México

[2] CONACYT–Universidad Veracruzana, Xalapa, Veracruz, México

{juliomunoz, edbenitez}@uv.mx, ggmoleroca@conacyt.mx

Resumen. Los sistemas consientes del contexto utilizan datos obtenidos de diversas fuentes para adaptarse y proveer servicios de interés al usuario de acuerdo a sus necesidades, localización o interacción con el ambiente. Sin embargo, el uso de fuentes heterogéneas crea un amplío volumen de datos que pueden diferir en formato, velocidad de transmisión y pueden ser afectados por el ruido ambiental. Esto genera cierta inconsistencia en los datos, la cual debe ser detectada a tiempo para evitar análisis erróneos. Para esto se hace uso de la fusión de datos, que es la acción de integrar fuentes diversas para ser analizadas de acuerdo a un contexto determinado. En este trabajo se presenta el diseño conceptual de un método de fusión de datos de fuentes heterogéneas, obtenidas de información contextual, con el propósito de mantener la consistencia de los datos en el proceso de fusión (extracción, pre–procesamiento, fusión y carga de datos).

Palabras clave: fusión de datos, fuentes heterogéneas, inconsistencia de datos, método de fusión, sistemas consientes del contexto.

Data Fusion Method of Heterogeneous Sources to Maintain Data Consistency

Abstract. Context–aware systems use data obtained from various sources to adapt and provide services of interest to the user according to their needs, location or interaction with the environment. However, the use of heterogeneous sources creates a large volume of data that may differ in format, transmission speed and may be affected by environmental noise. This generates some inconsistency in the data, which must be detected in time to avoid erroneous analysis. This is done using data fusion, which is the action of integrating diverse sources to be analyzed according to a determined context. This paper presents the conceptual design of a data fusion method of heterogeneous sources, obtained from contextual information, with the aim of maintaining the consistency of the data during the fusion process (extraction, pre–processing, fusion, and loading data).

Keywords: context-aware systems, data inconsistency, data fusion, fusion method, heterogeneous sources.

1. Introducción

En los últimos años, la tecnología ha evolucionado y se ha adaptado a las necesidades de los usuarios, así como al desarrollo de nuevas técnicas de comunicación y a la aparición de nuevos dispositivos con alto poder de cómputo y de tamaño cada vez más pequeños [1]. Este desarrollo tecnológico en la actualidad apertura nuevos temas de interés relacionados con los sistemas de cómputo y la forma en que perciben, capturan y analizan los datos del entorno que los rodea. A esto se le conoce como contexto, el cual es el entorno físico, emocional y social en el que se encuentra inmerso el usuario y que le dan significado, sentido y valor a las acciones o actividades que se realizan a su alrededor [1, 2].

Mediante el contexto puede caracterizarse la situación de una entidad, como persona, lugar u objeto, considerada relevante para la interacción entre un usuario y un sistema [3]. Para analizar esta interacción es importante tener información del contexto, que responda algunos cuestionamientos, conocidos como las cinco Ws [4]: a) quién (who), que se enfoca en la identidad del usuario; b) qué (what), que hace referencia a lo que el usuario está haciendo; c) dónde (where), que es la localización del usuario; d) cuándo (when), asociado al tiempo; y e) por qué (why), que incluye elementos del estado emocional de la persona.

La acción de contextualizar la información implica poner en contexto una situación que es percibida, de manera aislada o conjunta, de todos aquellos elementos que la rodean y que influyen sobre esta acción. A esto se le conoce como cómputo consiente del contexto o sistemas consientes del contexto, que detectan la actividad del usuario y reaccionan a los cambios del mismo para proveerle servicios que le sean de utilidad en la realización de sus actividades cotidianas [1, 5].

Para hacer esta detección y reacción se necesita analizar diferentes fuentes de datos, ya sean provenientes de sensores físicos o fuentes lógicas [1]. Los sensores físicos como los acelerómetros miden el desplazamiento y posición, los fotodiodos miden la proximidad e intensidad de la luz, las termo–resistencias miden la temperatura, entre otros. Mientras que las fuentes lógicas proveen información no física, como: agenda del usuario, preferencias, configuraciones, entre otros. En general, estas fuentes de datos son heterogéneas debido a la variedad de los formatos y velocidades de captura. Estas características hacen que los datos obtenidos sean completamente diferentes [6].

Una actividad compleja, previo al análisis de la detección y reacción a los cambios en un determinado contexto, es la fusión de datos de fuentes heterogéneas. Esta fusión de datos es el proceso de detección, asociación, correlación, estimación y combinación de datos en varios niveles [7], que provienen de diferentes fuentes, como [6, 8]: sensores, bases de datos, bitácoras, observaciones, señales e incluso decisiones.

A la fecha, la fusión de datos ha sido empleada en diversas áreas, como [6, 7]: procesamiento de señales, teoría de la información, estimación, estadística, inferencia e inteligencia artificial; teniendo mayores avances en aplicaciones militares mediante el reconocimiento automático de objetivos, navegación de vehículos autónomos,

sensado remoto e identificación de amenazas. Otras aplicaciones no militares son el monitoreo de procesos industriales, robótica, aplicaciones médicas, entre otros.

Indudablemente, el interés por la fusión de datos es cada vez mayor debido a la creciente incorporación de sensores en los dispositivos y sistemas de cómputo, el objetivo es tener información útil como apoyo en el proceso de la toma de decisiones sobre un determinado suceso, objeto o acción [5, 6]. Recientemente se está empleando también la fusión de datos para integrar fuentes variadas para hacer detecciones y clasificaciones de actividades en hogares inteligentes [9], ambientes virtuales inmersivos [10], interfaces tangibles [11] y escritorios inteligentes [12], apoyo en la realización de actividades del usuario [4], fusión de datos de observaciones humanas [13], entre otros.

Sin embargo, debido a la interacción que hace el usuario con el sistema consiente del contexto se produce una amplia variedad de datos heterogéneos producto del uso de sensores multimodales, objetos y aplicaciones que tienen como propósito facilitar la eficiencia del trabajo del usuario, adaptándose a los cambios de contexto y caracterización de la situación [14]. Hacer la fusión de datos de estas fuentes heterogéneas constituye un importante reto, esto debido al amplio conjunto de sensores y otras fuentes de datos que se utilizan para llevar a cabo una interacción dinámica del usuario, los objetos y el entorno [1, 14].

En este artículo se presenta el diseño conceptual de un método de fusión de datos de fuentes heterogéneas, obtenidas de información contextual, con el propósito de mantener la consistencia de los datos en el proceso de fusión. Las fases que contempla el método son: extracción, pre–procesamiento, fusión y carga de datos. El documento está estructurado en tres partes principales: la primera son los antecedentes del trabajo de investigación, la segunda es el método del trabajo científico, y la tercera es la propuesta de solución del método de fusión como diseño conceptual.

2. Antecedentes

Los primeros trabajos relacionados con los métodos de fusión de datos se remontan a 1786 con el método Condorcet, quien utilizó este enfoque para las votaciones en los modelos de la democracia [15]. Posteriormente, los métodos de fusión de datos se siguieron aplicando en diversas disciplinas, como: fiabilidad [16], reconocimiento de patrones [17], redes neuronales artificiales [18], proceso de toma de decisiones [8, 19], estimación estadística [20, 21], y predicción climática [22]. Además, actualmente una de las áreas que mayor impulso ha dado a la fusión de datos es la ciencia de datos, donde se emplean métodos para extraer, transformar y cargar fuentes de datos como parte del proceso de ingeniería de datos, previo a la analítica de éstos [3, 23].

En los sistemas consientes del contexto, los métodos de fusión de datos han demostrado ser especialmente útiles [7], debido a que proporcionan la posibilidad de utilizar los datos de múltiples fuentes. En este sentido, la fusión de los datos es un aspecto clave y crítico en sistemas con diversas fuentes (sensores, usuarios, entre otras). El objetivo es la acción de fusionar o combinar, de una forma eficiente, los datos de múltiples fuentes para superar las limitaciones de utilizar una sola fuente.

En el caso concreto del cómputo consiente del contexto, donde se obtienen datos de diversas fuentes que intentan describir las acciones que se generan alrededor del sistema o del usuario, existe una amplia red de sensores distribuidos de forma lógica, espacial o geográfica en un entorno y conectados por una red de transmisión. Los sensores pueden ser visuales (cámaras), auditivos (micrófonos), infrarrojos, sensores de humedad, temperatura, entre otros. Existen ventajas derivadas de utilizar sistemas de múltiples fuentes frente a sistemas tradicionales [24, 25]. Un ejemplo de fusión de datos en un sistema consiente del contexto se presenta en la Figura 1, donde los datos se obtienen de diversos sensores (S1, S2,…, Sn), así como de fuentes variadas, como: bases de datos, almacenes de datos, bitácoras, agendas, entre otras.

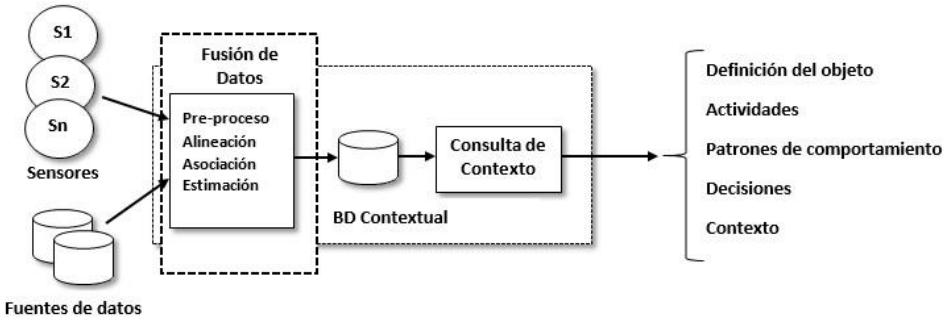


Fig. 1. Fusión de datos en un sistema consiente del contexto. Adaptada de [2].

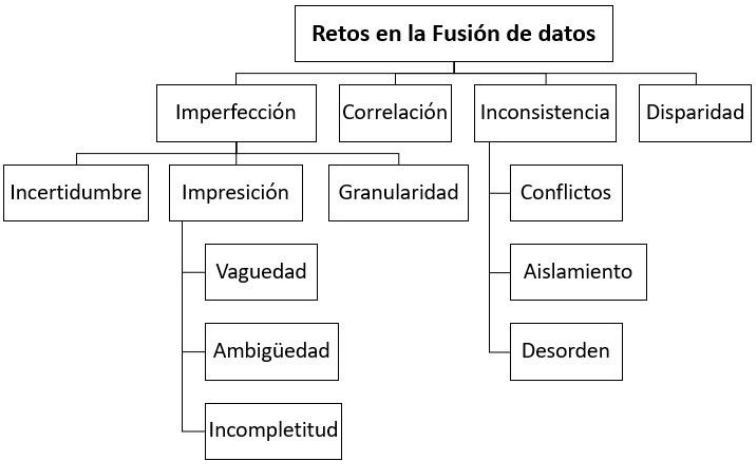


Fig. 2. Clasificación de los retos actuales en la fusión de datos. Adaptada de [31].

Los datos adquiridos se fusionan mediante un proceso de pre–procesamiento, alineación, asociación y estimación, y a través de consultas son comparados con una base de datos contextual, esto con el propósito de obtener una determinada inferencia del contexto. Esta inferencia proporciona información sobre la identidad del objeto, usuario, patrones de comportamiento o incluso permite al sistema tomar decisiones

adaptándose a la situación en un determinado momento, es decir, se integra la información de entrada (datos) para obtener datos refinados o características que describen las acciones del usuario y su comportamiento.

Así, para la implementación del proceso de fusión de datos existen diversos modelos, como Joint Directors Laboratories (JDL) [7], Thomopoulos [26], Integración de multi–sensores [27], Basado en conocimiento del comportamiento [28], Cascada [29] y Omnibus [30], y arquitecturas, como centralizada, descentralizada y distribuida [7], los cuales definen los niveles de complejidad, procesos y el momento cuando los datos deben ser fusionados. Sin embargo, debido a la naturaleza de los datos, a la amplia variedad de sensores y fuentes, y a la inexistencia de algoritmos ideales de fusión de datos, en [31] se propone una clasificación de retos actuales en la fusión de datos: a) imperfección, b) correlación, c) inconsistencia, y d) disparidad. Cada uno de estos retos son descritos en la sección siguiente.

2.1. Retos en la fusión de datos

Los retos en la fusión de datos surgen debido al variado formato, tipo y velocidad de muestreo de los sensores, la diversidad de las fuentes y la imperfección de los datos. En la Figura 2 se presenta los retos identificados, los cuales fueron clasificados de acuerdo a su naturaleza [31]:

- Imperfección. Los datos suministrados por los sensores pueden ser afectados por cierto nivel de imprecisión, así como por la incertidumbre en las mediciones. Esta incertidumbre se presenta no solo por la imprecisión o el ruido de las mediciones de los sensores, sino también por las ambigüedades y la incapacidad del sistema de fusión de distinguirlas.

- Correlación. Este tipo de problema es común en configuraciones distribuidas, donde algunos de los sensores obtienen los mismos datos, de rutas diferentes o debido a rutas cíclicas del flujo de los datos.

- Inconsistencia. Constituye uno de los problemas más importantes en la fusión de datos debido a la incertidumbre inherente de las mediciones de los sensores, obteniéndose datos incorrectos. La causa de obtener estos datos incorrectos puede ser debido a fallas permanentes o fallas de corta duración en las mediciones.

- Disparidad. Los datos deben ser transformados localmente por cada sensor en un formato común, antes del proceso de fusión. Este problema de disparidad se presenta a menudo en los sensores debido a errores de calibración.

En la actualidad, cubrir todos los retos y limitaciones en la fusión de datos no es trivial, puesto que se produce una amplia variedad de datos heterogéneos debido al uso de sensores multimodales, objetos y aplicaciones que tienen como propósito facilitar la eficiencia del trabajo del usuario, adaptándose a los cambios de contexto y caracterización de una situación. Por tanto, la fusión de datos de múltiples fuentes es un aspecto clave y crítico en sistemas con variados sensores, usuarios y otros actores.

En este sentido, debido a la naturaleza de los datos existen trabajos que proponen soluciones parciales, pero no se tienen métodos que cubran los diferentes retos. Por lo que, para esta investigación se considera útil incluir para la fusión de datos métodos de fusión de fuentes heterogéneas, con el propósito de minimizar la inconsistencia en la fusión de datos contextuales. Así, se pretende obtener resultados con mejor calidad, a partir de múltiples sensores y fuentes variadas, realizando combinaciones de éstas de forma eficiente.

2.2. Inconsistencia de datos

Los datos que provienen de sensores son afectados por cierto grado de imprecisión e incertidumbre en las mediciones. Esto genera inconsistencia en los datos. Además, las fuentes pueden tener diferentes velocidades de comunicación, esto provoca que los datos sean enviados de diversas maneras, sin mantener un orden al momento de ser recibidos, incluso pueden obtener valores atípicos debido a situaciones inesperadas, fallas en la medición y mediciones erróneas.

Algunos métodos utilizados para minimizar la inconsistencia de los datos son [4, 13, 31]: teoría de razonamiento evidencial Dempster Shafer, lógica difusa, estimación Bayesiana, filtro de Kalman, entre otros. Para la fusión de datos estos métodos toman en cuenta los factores ambientales y la imprecisión de los sensores que afectan las mediciones, así como la ambigüedad de los datos y la dificultad para distinguirlos [32]. Estas imprecisiones y ambigüedades generan inconsistencias que deben ser identificadas y tratadas para evitar errores durante el proceso de fusión, tales como [31]:

- Conflicto en los datos. Son datos atípicos recibidos por el sistema de fusión, los cuales deben ser descartados para evitar resultados erróneos. Estos pueden ser datos corruptos por factores ajenos al sistema de fusión, o provenir de fuentes con diversos formatos y dimensiones que necesitan ser integrados. Ante esto, se debe contar con mecanismos de estimación para prevenir esta clase de conflictos.

- Datos en desorden. Los datos para la fusión por lo general están organizados por la marca de tiempo que se les asigna cuando éstos fueron creados. Factores como la variación en los tiempos de propagación y fuentes heterogéneas hacen que los datos lleguen en una secuencia diferente a la esperada por el sistema de fusión. El principal problema es utilizar datos obsoletos para actualizar estados presentes, es decir, utilizar datos obtenidos con cierto retraso para la actualización de estados en tiempo real.

- Aislamiento de los datos. Esta clase de datos son mediciones erróneas por algún fallo en los sensores o el ruido de la señal provocado por el ambiente. Estos datos deben ser aislados para evitar estimaciones erróneas al ser fusionados con datos correctos. El objetivo es identificar o predecir estos datos para eliminarlos antes del proceso de fusión.

Por tanto, un aspecto fundamental en la fusión de datos es la capacidad de los métodos de fusión para mantener la consistencia de los datos, evitando conflictos,

aislamiento y desorden, provenientes de fuentes diversas, integrándolos de una manera adecuada, robusta y consistente [33] para obtener una visión unificada del evento, fenómeno o suceso monitoreado [34]. Sin embargo, en la fusión de datos de fuentes heterogéneas existen variados problemas que surgen durante el proceso de integración, tales como la asociación de los datos, la incertidumbre de los sensores provocada por el ruido inducido, esto debido a factores externos del ambiente, o incluso por imperfecciones en los mismos sensores, así como la administración de los datos, entre otros [33, 35].

Precisamente, ante los variados problemas en el proceso de fusión de datos, en la actualidad existen algunos estudios que se enfocan en la validación de sensores, antes al proceso de fusión; así como en el uso de conocimiento previo sobre los datos. Sin embargo, por lo general no siempre se tiene esta información debido a las condiciones reales en que opera el sistema de fusión [7, 31, 32], esto es, no es posible modelar en su totalidad todas las fallas inherentes a los sensores y fuentes variadas.

3. Método

Dado el propósito de mantener la consistencia de datos en el proceso de fusión de fuentes heterogéneas, para esta investigación se definieron cuatro etapas de trabajo; las cuales son de tipo exploratoria y aplicada que en su conjunto forman el método definido para esta investigación.

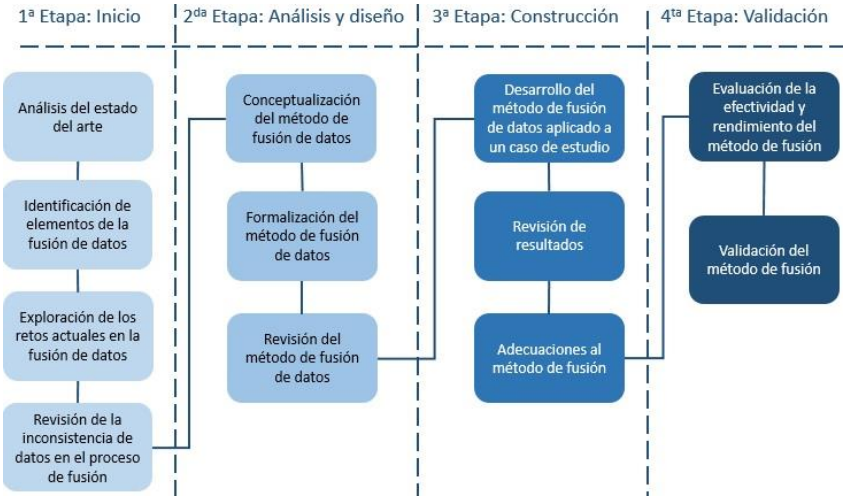


Fig. 3. Método de trabajo.

La primera es el inicio o preparación, la cual es una etapa fundamental que sustenta el análisis teórico de la fusión de datos y sus retos actuales. La segunda es el análisis y diseño conceptual de la propuesta de solución para la caracterización del método de fusión de datos de fuentes heterogéneas. La tercera etapa es el desarrollo del método de fusión con base en la elección de técnicas adecuadas aplicadas a un caso de estudio. La

cuarta es la evaluación del método de fusión de datos desde el punto de vista del desempeño y el cumplimiento del objetivo de mantener la consistencia de los datos en el proceso de fusión. En la Figura 3 se hace una proyección del método científico que se utiliza en esta investigación y las actividades contempladas en éste.

En virtud de lo anterior, en este artículo se presenta un avance de la investigación que cubre las primeras dos etapas del método de trabajo: i) inicio y ii) análisis y diseño. Ambas etapas fueron fundamentales como proceso analítico para diseñar esta propuesta conceptual del método de fusión de datos de fuentes heterogéneas, cuyo objetivo es mantener la consistencia de los datos en el proceso de fusión.

4. Propuesta

Para los sistemas consientes del contexto, el objetivo es incrementar la sensación de interacción del usuario con el sistema, ya sea mediante la manipulación física de objetos que afecten el comportamiento del sistema [36] o mediante la interacción y relación del usuario con el ambiente para mejorar el trabajo colaborativo con otros usuarios y dar soporte durante el proceso de la toma de decisiones [4]. Estas interacciones son detectadas mediante el uso de diversos sensores de video y proximidad, cámaras, acelerómetros, transceptores (receptores y transmisores), entre otros. El sistema reacciona a estas interacciones que representan información digital y adecúa su comportamiento como sea necesario, incluso creando interacciones físicas que respondan a las iniciadas por el usuario.

En este sentido, el uso de sensores diversos es importante para que estas interacciones puedan realizarse de manera adecuada, por lo que se debe tener un esquema de fusión de datos que permita explotar estos beneficios. La propia diversidad de los sensores y la naturaleza de las interacciones del usuario con el sistema pueden crear inconsistencia en los datos. Siendo estos datos necesarios para la caracterización del contexto.

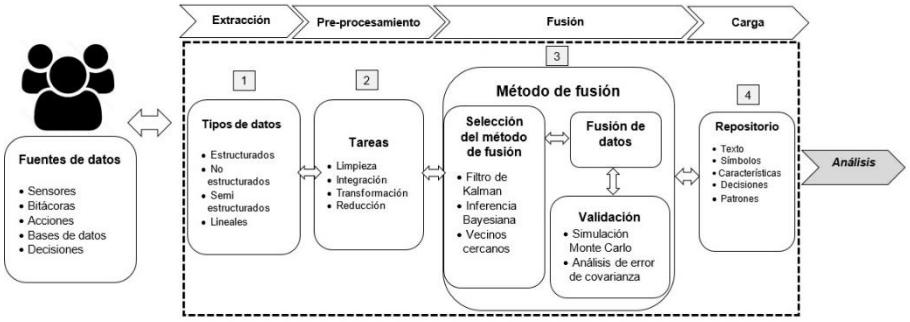


Fig. 4. Diseño conceptual del método de fusión de datos de fuentes heterogéneas.

La inconsistencia de datos es un factor crítico que debe ser detectado y evitado en el proceso de fusión. Esta fusión de datos inconsistentes con datos correctos puede producir una estimación errónea, que eventualmente puede conducir a un análisis

incorrecto [5]. Ante esto, un aspecto fundamental en la fusión de datos es la capacidad de los métodos de fusión para mantener la consistencia de los datos [31], evitando conflictos, aislamiento y desorden.

En este trabajo se propone el diseño conceptual de un método de fusión de datos de fuentes variadas (Figura 4), donde las interacciones del usuario con otros usuarios, con el ambiente o con el sistema generan datos heterogéneos obtenidos de sensores, bases de datos, agendas, configuraciones, preferencias, por mencionar algunas. Estas interacciones pueden ser acciones que realiza el usuario en un determinado espacio de trabajo, ambiente o entorno en el que se encuentra, e incluso pueden ser mediciones fisiológicas, como frecuencia respiratoria, cardíaca, temperatura, entre otros.

El diseño de este método de fusión está estructurado en cuatro etapas o módulos de procesamiento (extracción, pre–procesamiento, fusión y carga) que tiene como propósito principal mantener la consistencia de los datos a lo largo del proceso de fusión.

En la etapa de *extracción* los datos son obtenidos de diversas fuentes de acuerdo a su naturaleza, como sensores, bases de datos, bitácoras, acciones, mediciones fisiológicas, entre otras. Todas éstas con formatos diversos o incluso sin un formato establecido. Entre los tipos de datos considerados para procesar en este módulo están:

- Estructurados. Son datos que pueden estar almacenados en bases de datos, tablas, matrices, arreglos, bitácoras, entre otros.
- No estructurados. Son datos que no tienen un formato específico, como documentos de texto, datos crudos de sensores, redes sociales, entre otros.
- Semi–estructurados. Son datos que no están limitados a ciertos campos o tablas, pero tienen una forma de separar los elementos que los componen, como XML (eXtensible Markup Language), JSON (Java Script Object Notation), entre otros.
- Datos lineales. Estos datos están compuestos de texto plano, como archivo de texto delimitado (txt), valores separados por coma (csv), entre otros.

El *pre–procesamiento* constituye una etapa clave, previo a la fase de fusión de datos, puesto que los datos son preparados para ser fusionados y parte de la inconsistencia puede ser resuelta o minimizada en esta etapa. Para esto fueron definidas cuatro tareas generales:

- Limpieza. Consiste en hacer un proceso de limpieza o depuración para evitar datos incompletos, duplicados o conflictivos. Esto con el fin de obtener datos de mayor calidad, minimizando la inconsistencia de éstos.
- Integración. Consiste en agrupar los datos que vienen de diversas fuentes, aunque éstos tengan formatos diferentes. El objetivo es obtener fuentes de datos coherentes y consistentes, evitando problemas de representación y codificación.
- Transformación. Consiste en transformar valores numéricos a nominales o viceversa, ya sea mediante discretizaciones o derivaciones. Esta última permite crear nuevos atributos, incluso cambiando el formato de los datos. Además,

pueden hacerse normalizaciones para eliminar variaciones en los datos ajustándolos a una escala específica.

- Reducción. Consiste en reducir la alta dimensionalidad de los datos. En este caso el tamaño de éstos puede ser a consecuencia de datos redundantes. Para esto se obtienen muestras o representaciones de los datos, manteniendo su integridad, o usando mecanismo de reducción como análisis de componentes principales y análisis correlacional de datos.

Posterior al pre–procesamiento, en la fase de *fusión* se integran los datos con base en la selección de uno o más algoritmos. La selección del algoritmo depende de la naturaleza, el tipo y el nivel de abstracción de los datos, como: datos, atributos, símbolos, identidad, patrones o decisiones. Estos algoritmos deben ser capaces de distinguir la inconsistencia y combinar diferentes tipos de datos, a diferentes niveles, para obtener una descripción consistente [13, 32]. Por lo que, para este trabajo los algoritmos definidos como parte del proceso de fusión son:

- Inferencia Bayesiana. Este algoritmo asume una representación de un estado de interés dada una observación, es decir, la relación entre el estado y la observación pueden ser codificados como una probabilidad conjunta o una distribución de probabilidad conjunta [32]. De esta forma la información es actualiza cada vez que se fusionan nuevos datos derivados de una nueva observación [31]. Esto provee una forma de combinar datos observados con información previamente adquirida [37].

- Vecinos cercanos. Consiste en agrupar o seleccionar los datos cercanos unos con otros. La cercanía de un valor depende de la métrica de la distancia utilizada, por ejemplo basada en la distancia absoluta, euclidiana o una función estadística de distancia [15, 31].

- Filtro de Kalman. Este algoritmo utiliza una serie de estimaciones de estados previos para hacer una predicción de estados siguientes, adelantando las decisiones del sistema [7, 31, 37], esto es, hace una estimación del estado que evoluciona con el tiempo basándose en observaciones periódicas. Para esto el filtro de Kalman emplea un modelo estadístico sobre la evolución de un parámetro contra el tiempo y cómo las observaciones se relacionan con este parámetro.

Para la *validación* de la fusión de datos se incluye como métodos de evaluación la simulación de Monte Carlo y las técnicas de análisis de error de covarianza.

- Simulación de Monte Carlo. Se encarga de realizar un análisis de la fusión de datos basado en la creación de modelos de posibles resultados, esto es, mediante la sustitución de un rango de valores para un determinado valor con incertidumbre [7]. Una característica de este tipo de simulación es el uso de mediciones estadísticas para medir el desempeño del proceso de fusión [37].

- Análisis de error de covarianza. Es un modelo lineal general con una o más variables cuantitativas y uno o más factores. Se calcula mediante una regresión

lineal múltiple, que es un procedimiento estadístico que elimina la heterogeneidad causada en la variable de interés (dependiente) por la influencia de una o más variables cuantitativas (covariables) [15]. La inclusión de covariables puede aumentar la potencia estadística debido a que por lo general reduce la variabilidad.

Finalmente, en la etapa de *carga* se gestiona el almacenamiento de los datos ya sea en un repositorio, almacén, archivo, u otros medios de acuerdo a las necesidades de uso (texto, símbolo, características, decisiones o patrones), como la analítica de datos para detectar patrones de comportamiento del usuario, roles, perfiles, emociones, configuraciones, trabajo colaborativo, y otros temas de interés asociados a los sistemas consientes del contexto; dominio que ha demostrado ser sumamente útil para percibir, capturar y analizar datos de diversos ámbitos, como sociales, comerciales, biomédicos, ambientales, militares, entre otros.

5. Consideraciones finales

La fusión de datos en la actualidad ha tomado una creciente importancia debido a la incorporación de sensores en los sistemas de cómputo y a la necesidad de hacer que estos sistemas interactúen con el usuario y el ambiente, haciendo que éstos se adapten a situaciones que ocurren en el entorno y den el servicio adecuado en el momento en que se requiera.

La principal ventaja de fusionar datos, en comparación de tener una sola fuente, es tener mayor información para caracterizar una determinada situación o actividad que es observada. Sin embargo, para tener diferentes fuentes es necesario usar diversos sensores que pueden variar en naturaleza, formato, velocidad e incluso interactuar con fuentes lógicas, como bases de datos, bitácoras o preferencias del usuario.

La heterogeneidad de datos trae consigo retos que deben ser solucionados, donde la imperfección en los datos, el dinamismo del entorno y la propia naturaleza de éstos crean inconsistencias que deben resolverse mediante métodos adecuados.

Resolver el reto de la inconsistencia de datos en el proceso de fusión no es trivial debido a la amplia variedad de datos heterogéneos que se producen a través de sensores, objetos y aplicaciones. Los cuales son útiles para el análisis del contexto y la caracterización de una determinada situación.

Derivado de la primera etapa del trabajo de investigación sobre la fusión de datos de fuentes heterogéneas, se propuso el diseño conceptual de un método de fusión de datos estructurado en cuatro etapas: a) extracción, b) pre–procesamiento, c) fusión y d) carga. El enfoque de esta propuesta es detectar la inconsistencia de los datos, previo al proceso de fusión, para mantener su consistencia durante la integración, evitando el conflicto, el desorden y el error en los datos.

Cómo trabajo posterior se implementará el método de fusión de datos y se evaluará el desempeño de los algoritmos desde el punto de vista de cumplimiento del objetivo, que es de mantener la consistencia de los datos en el proceso de fusión. Para esto se utilizarán datos heterogéneos obtenidos de un sistema consiente del contexto y la actividad del usuario.

Finalmente, definir y diseñar nuevos métodos de fusión de datos representa en la actualidad un reto importante para mejorar la precisión de los sistemas de integración de datos, que pueden ser generados por las interacciones del usuario con el sistema y el ambiente. Esto representa un campo de oportunidad para la investigación de sistemas consientes del contexto y la interacción humano–computadora con el propósito de mejorar el trabajo colaborativo entre usuarios, optimizar las tareas realizadas, dar soporte al proceso de toma de decisiones, realizar inferencias, entre otros.

Agradecimientos. Este trabajo forma parte del proyecto "Infraestructura para agilizar el desarrollo de sistemas centrados en el usuario" financiado por el Consejo Nacional de Ciencia y Tecnología (CONACYT) en el marco del programa de Cátedras CONACYT (Ref. 3053). Además, se agradece a CONACYT por la beca de doctorado número 449416, así como a la Universidad Veracruzana por el apoyo en el desarrollo de esta investigación.

Referencias

1. Gellersen, H., Schimidt, A., Beigl, M.: Multi–sensor context–awareness in mobile devices and smart artifacts. Mobile Networks and Applications, 7(5), pp. 341–351 (2002)
2. Benítez–Guerrero, E.: Context–Aware Mobile Information Systems: Data Management Issues and Opportunities. In: Proceedings of the International Conference on Information & Knowledge Engineering, Las Vegas (2010)
3. Wu-Zhu, X., Wu, G., Ding, W.: Data mining with big data. IEEE transactions on knowledge and data engineering, 26(1), pp. 97–107 (2014)
4. Abowd, G. D., Mynatt, E. D.: Charting past, present, and future research in ubiquitous computing. ACM Transactions on Computer Human Interaction, 7(1), pp. 29–58 (2000)
5. Bernardos, A. M., Tarrio, P., Casar, J. R.: A data fusion framework for context–aware mobile services. In: Proceedings Multisensor Fusion and Integration for Intelligent Systems, pp. 606–613 (2008)
6. White, F.: Data fusion lexicon. Reporte técnico, Joint Directors of Laboratories (1991)
7. Hall, D., Llinas, J.: An introduction to multisensory data fusion. Proceedings of the IEEE, 85(1), pp. 6–23 (1997)
8. Dasarathy, B.: Sensor fusion potential exploitation–innovative architectures and illustrative applications. Proceedings of the IEEE, 85(1), pp. 24–38 (1997)
9. Rodriguez, S., De Paz, J., Villarrubia, G., Zato, C., Bajo, J., Corchado, J.: Multi-agent information fusion system to manage data from a WSN in residential home. Information Fusion, 23, pp. 43–57 (2015)
10. Surie, D., Pederson, T., Lagriffoul, F., Janlert, L., Sjölie, D.: Activity recognition using an egocentric perspective of everyday objects. In: International conference on ubiquitous intelligence and computing, Hong Kong, pp. 246–257 (2007)
11. Branton, C., Ullmer, N., Wiggins, A., Rogge, L., Setty, N., Beck, S. D., Reeser, A.: Toward rapid and iterative development of tangible, collaborative, distributed user interfaces. In: Proceedings of the 5th ACM SIGCHI symposium on Engineering interactive computing systems, London, pp. 239–248 (2013)
12. Yang, M., Al Kutubi, M., Pham, D.: Continuous acoustic source tracking for tangible acoustic interfaces. Measurement, 46(3), pp. 1272–1278 (2013)

13. Jenkins, M. P., Gross, G. A., Bisantz, A. M., Nagi, R.: Towards context aware data fusion: Modeling and integration of situationally qualified human observations to manage uncertainty in a hard+soft fusion process. Information Fusion, 21, pp.130–144 (2015)
14. Qin, W., Suo, Y., Shi, Y.: CAMPS: A middleware for providing context–aware services for smart space. In: International conference on grid and pervasive computing, Taichung, pp. 644–653 (2006)
15. Castanedo, F.: A review of data fusion techniques. The Scientific World Journal (2013)
16. Von-Neumann, J.: Probabilistic logics and the synthesis of reliable organisms from unreliable components. Automata studies, 34, pp. 43–98 (1956)
17. Chow, C.: Statistical independence and threshold functions. IEEE Transactions on Electronic Computers, 1, pp. 66–68 (1965)
18. Hashem, S.: Algorithms for optimal linear combinations of neural networks. Neural networks, 1, pp. 242–247 (1997)
19. Varshney, P. K.: Distributed Bayesian detection: Parallel fusion network. In: Distributed detection and data fusion, Springer, New York, pp. 36–118 (1997)
20. Breiman, L.: Stacked regressions. Machine learning, 24(1), pp. 49–64 (1996)
21. Juditsky, A., Nemirovski, A.: Functional aggregation for nonparametric estimation. In: IRISA, Rennes (1996)
22. Granger, C. W.: Invited review-combining forecasts twenty years later. Journal of Forecasting, 8(3), pp. 167–173 (1989)
23. Jagadish, H., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., Shahabi, C.: Big data and its technical challenges. Communications of the ACM, 57(7), pp. 86–94 (2014)
24. Luo, R., Kay, M.: Multisensor integration and fusion: Issues and approaches. In: Proceedings SPIE, Florida, pp. 42–49 (1988)
25. Luo, R., Yih, C., Su, K.: Multisensor fusion and integration: approaches, applications, and future research directions. IEEE Sensors Journal, 2(2), pp. 107–119 (2002)
26. Thomopoulos, S. C.: Sensor integration and data fusion. In: Symposium on visual communications, image processing, and intelligent robotics systems, pp. 178–191 (1990)
27. Luo, R. C., Kay, M. G.: Multisensor integration and fusion: issues and approaches. In: Orlando technical symposium, pp. 42–49 (1988)
28. Pau, L. F.: Sensor data fusion. Journal of Intelligent and Robotic Systems, 1(2), pp. 103–116 (1988)
29. Harris, C., Bailey, A., Dodd, T.: Multi-sensor data fusion in defence and aerospace. Aeronautical Journal, 102(1015), pp. 229–244 (1998)
30. Bedworth, M., O'Brien, J.: The omnibus model: a new model of data fusion? IEEE Aerospace and Electronic Systems Magazine, 15(4), pp. 30–36 (2000)
31. Khaleghi, B., Khamis, A., Karray, F., Razavi, S.: Multisensor data fusion: A review of the state-of-the-art. Information Fusion, 14(1), pp. 28–44 (2013)
32. Kumar, M., Garg, D. P., Zachery, R. A.: A method for judicious fusion of inconsistent multiple sensor data. IEEE Sensors Journal, 7(5), pp. 723–733 (2007)
33. Kale, D. R., Aparadh, S. Y.: A Study of a Detection and Elimination of Data Inconsistency in Data Integration. International Journal of Scientific Research in Science, Engineering and Technology, 2(1), pp. 532–535 (2016)
34. Saha, B., Srivastava, D.: Data quality: The other face of big data. In: IEEE 30th International Conference on Data Engineering, pp. 1294–1297 (2014)
35. Wang, X., Huang, L., Xu, X., Zhang, Y., Chen, J. Q.: A Solution for Data Inconsistency in Data Integration. Journal of Information Science and Engineering 27(2), pp. 681–695 (2011)

36. Patten, J., Ishii, H., Hines, J., Pangaro, G.: Sensetable: a wireless object tracking platform for tangible user interfaces. In: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 253–260 (2001)
37. Siciliano, B., Khatib, O.: Handbook of Robotics. Springer (2016)

ISSN 1870-4069

Red neuronal Backpropagation para la predicción de datos de contaminación y prevención de ataques a personas con padecimientos de rinitis alérgica y asma

Daniel Cerna-Vázquez[1], Carlos Lino-Ramírez[1], Arnoldo Díaz-Ramírez[2], Juan Francisco Mosiño[1], Miguel Ángel Casillas-Araiza[1], Rosario Baltazar-Flores[1], Guillermo Mendez-Zamora[1]

[1] Tecnológico Nacional de México, Instituto Tecnológico de León, León, Guanajuato, México

[2] Tecnológico Nacional de México, Instituto Tecnológico de Mexicali, Mexicali, Baja California, México

cerna.daniel.dc@gmail.com, carloslino@itleon.edu.mx, jfmosino@gmail.com, miguel.casillas@gmail.com, charobalmx@yahoo.com.mx, guillermomendez06@gmail.com, adiaz@itmexicali.edu.mx

Resumen. La prevención de enfermedades respiratorias causadas por los altos índices de contaminación ambiental, es un tema de importante discusión en las grandes ciudades, donde la industrialización y la sobrepoblación causan un incremento de partículas alergénicas que agravan el padecimiento de rinitis alérgica y asma, sobre todo en la comunidad infantil. El problema radica en la desinformación de la población acerca de la calidad del aire y las medidas preventivas a tomar en cuenta para evitar un deterioro en la salud. En este artículo, se presentan datos monitoreados por una red de sensores que registran cinco de los principales alérgenos para diferentes zonas de la ciudad de León, Guanajuato. Una red neuronal artificial (RNA) con un entrenamiento supervisado de tipo Backpropagation, es empleada para la predicción de datos futuros hasta alcanzar un error mínimo. Posteriormente, se emiten recomendaciones previamente avaladas, con base a los resultados arrojados por la red neuronal. La metodología propuesta genera resultados eficientes, medidos en el error de las soluciones y en tiempo de ejecución.

Palabras clave: redes de sensores, red neuronal artificial, entrenamiento backpropagation, predicción de datos climáticos.

Backpropagation Neural Network for the Prediction of Contamination Data and Prevention of Attacks to People with Allergic Rhinitis and Asthma

Abstract. The prevention of respiratory diseases caused by high air pollution rates is an important issue in big cities, where industrialization and overpopulation cause an increase in allergenic particles that aggravate the disease

of allergic rhinitis and asthma, especially in childhood. The problem lies in the disinformation of the population about air quality and the preventive measures to be taken in order to avoid deterioration in health. In this paper, data are monitored by a sensor network that registers five of the main allergens for different areas of the city of León, Guanajuato. An artificial neural network (ANN) with a supervised Backpropagation training is used to predict future data until a minimum error is reached. Then, we emit approved recommendations, based on the results of the neural network. The proposed methodology generates efficient results, measured in the error of the solutions and in execution time.

Keywords: sensor network, artificial neural network, backpropagation training, climate data prediction.

1. Introducción

Una de las principales causas de enfermedades respiratorias y que genera gran preocupación para las grandes ciudades, es la emisión aérea de contaminantes, que es ocasionada por diversas actividades humanas, entre ellas, la industria [1], [2]. El empleo de recursos no renovables en la producción de energía, como el petróleo o el carbón, genera importantes emisiones de dióxido de azufre (SO_2), monóxido de carbono (CO), entre otras. Por otro lado, los medios de transporte empleados en la vida cotidiana son otra fuente alarmante de contaminación. Una gran parte de estos contaminantes emitidos al ambiente es generada por los automóviles [3]. Según las Naciones Unidas, actualmente hay alrededor de 7 billones de personas en el mundo [4], [5], lo que representa una enorme fuente de contaminación, agravando el problema a medida que las personas tienden a migrar a grandes ciudades, lo que conlleva a una mayor emisión de contaminantes que deterioran la calidad del aire. Según datos del Consejo Nacional de Población (CONAPO), el 72.3% de la población en México, vive en zonas metropolitanas. Además, según la Organización de las Naciones Unidas (ONU), en los próximos 10 años las poblaciones rurales comenzarán a disminuir significativamente [6]. Esto genera que la salud de las personas, en las grandes ciudades, se deteriore cada vez más, ya que los individuos que viven en lugares con altos índices de contaminación son más propensos a adquirir enfermedades del tipo respiratorias, como alergias o asma [7].

Los alérgenos son aquellas partículas que pueden causar alergias y/o agravar el asma. Entre los principales alérgenos encontrados en el aire, se encuentran el ozono (O_3), el dióxido de azufre (SO_2), el monóxido de carbono (CO), el plomo (Pb), la materia particulada ($PM_{2.5}$ y PM_{10}), entre otros [8], [9]. El asma es una de las enfermedades que más afecta principalmente a la población infantil. A nivel escolar existe una prevalencia del 9.4% en los infantes a nivel mundial, y un 12.6% durante la adolescencia. De estos, el 80% presentan alergias con respecto al entorno en el que viven [10]. Según datos de la encuesta ISAAC, en México la prevalencia del asma va del 2.7% al 21.8% en diferentes ciudades. Para los infantes, la prevalencia fue del 5.7% y de 5.9% para los adolescentes [11]. La rinitis alérgica es una de las enfermedades crónicas más frecuentes a nivel mundial [12]. Este padecimiento es más frecuente en regiones con alto índice de tráfico vial. Esto se debe a que el aumento de la contaminación genera también un aumento en los síntomas de las personas que sufren

esta enfermedad [7]. En México, según el estudio ISAAC, existe una prevalencia del 4.6% de rinitis alérgica en infantes. La rinitis alérgica se puede agravar debido a los alérgenos que genera la contaminación y con esto llegar a generar asma [13].

El monitoreo constante de la calidad del aire, empleando redes de sensores [14], permite a las personas que sufren de rinitis alérgica y asma, mantenerse informadas acerca de las condiciones ambientales de su entorno, para así tomar acciones pertinentes y evitar un deterioro en la salud, por causa de los altos niveles de contaminación que les rodea. En la ciudad de León, Guanajuato, México, se cuenta con la monitorización de la calidad del aire por parte del Instituto de Ecología del Estado (IEE). El IEE cuenta con un Sistema de Monitoreo de Calidad del Aire del Estado de Guanajuato (SIMEG). El sistema está conformado por tres estaciones de monitoreo fijas, distribuidas en la ciudad de León. En este trabajo se hace uso de los datos generados por una de las tres estaciones de monitoreo del SIMEG [15], la cual es llamada CICEG y recibe este nombre dado que se encuentra ubicada en las instalaciones de la Cámara de la Industria del Calzado del Estado de Guanajuato y genera mediciones de los alérgenos contaminantes PM_{10}, O_3, NO_2, SO_2 y CO. Éstos, junto con el Pb, entran en la categoría de principales contaminantes del aire [16].

En el presente trabajo, con el propósito de hacer de conocimiento a las personas con padecimientos causadas por alérgenos información relacionada con el nivel actualizado de la calidad del aire de su entorno, se hace uso de una Red Neuronal Backpropagation para predicción de los datos futuros de los alérgenos contaminantes obtenidos por la red de sensores instalada en la estación CICEG. Las redes neuronales son una herramienta que ha demostrado su eficacia a la hora de predicción de datos futuros. Como en [17], donde se hace uso de una Red Neuronal Backpropagation para la predicción a corto plazo de energía eólica, o en [18], donde se logran buenas predicciones en mercado de acciones. El poder predecir los niveles futuros de la calidad del aire, nos ayudará a enviar alertas que aconsejen a las personas con los padecimientos antes mencionados, para que tomen medidas preventivas en caso de un futuro deterioro de la calidad del aire.

El resto de este artículo está organizado de la siguiente manera: en la Sección 2 se presentan conceptos teóricos utilizados en la elaboración de este trabajo. En la Sección 3 se muestra la metodología. En la Sección 4, se discuten los resultados que se obtuvieron. Finalmente, en la Sección 5 se presentan las conclusiones.

2. Marco teórico

2.1. Air Quality Index (AQI)

El AQI es un indicador de la calidad del aire diaria, el cual muestra qué tan limpio está el aire que nos rodea y qué efectos se pueden llegar a generar por dicha calidad del aire. Las mediciones del AQI van de 0 a 500 ppb. Estas mediciones están divididas en cinco categorías (Bueno, Moderado, Insalubre para grupos sensitivos, Insalubre, Muy Insalubre y Peligroso) que se determinan dependiendo del nivel de contaminante que se encuentra en mayor proporción [19]. El AQI se concentra en la medición de cinco de los principales contaminantes del aire: PM_x, O_3, NO_2, SO_2 y CO. El nivel de la calidad del aire, determinado por el AQI [20], se obtiene utilizando la ecuación 1:

$$I_p = \frac{I_{HI} - I_{LO}}{BP_{HI} - BP_{LO}}(C_P - BP_{LO}) + I_{LO}\ , \quad (1)$$

donde I_p es el índice de la calidad del aire, C_P es la concentración de contaminante observada, I_{HI} es el punto de quiebre AQI mayor al C_P observado, I_{lo} es el punto de quiebre AQI menor al C_P observado, BP_{HI} es el punto de quiebre contaminante mayor al C_P observado y BP_{LO} es el punto de quiebre contaminante menor al C_P observado. Los puntos de quiebre AQI se muestran en la Tabla 1, los cuales se sustituyen en la ecuación 1 para conocer el índice de la calidad del aire:

Tabla 1. Clasificación de calidad del aire AQI y sus puntos de quiebre (menor y mayor) correspondientes.

Categoría	Punto de quiebre menor	Punto de quiebre mayor	Color Característico de categoría
Bueno	0	50	Verde
Moderado	51	100	Amarillo
Insalubre para grupos sensibles	101	150	Naranja
Insalubre	151	200	Rojo
Muy Insalubre	201	300	Morado
Peligroso	301	500	Marrón

Tabla 2. Clasificación de los cinco principales contaminantes, sensados por la estación CICEG, con sus repsectivas unidades de medidas.

Contaminante	**PM10**	**O3**	**SO2**	**NO2**	**CO**
Unidad de medida	**$\mu g/m^3$**	**Ppb.**	**Ppb.**	**Ppb.**	**Ppm.**
Bueno	0-54	0-64	0-99	0-198	0-9
Satisfactoria	55-74	65-69	100-109	190-209	9-10
No Satisfactoria	75-174	70-130	110-174	210-315	11-15
Mala	175-274	131-184	175-239	316-420	16-22
Muy Mala	>275	>185	>240	>420	>22

Para los puntos de quiebre de los contaminantes, se utilizó el semáforo de calidad del aire que se publica en el Informe de Estado y Tendencia de la Calidad del Aire Guanajuato 2014 [15]. Los puntos de quiebre para los contaminantes obtenidos por la estación CICEG, se muestran en la Tabla 2.

2.2. Red neuronal artificial Backpropagation (RNA-BP)

Rumelhart, Hinton, y Williams, en 1986, introdujeron la red Backpropagation, que es un tipo de descenso de gradiente [21], ya que utiliza el cálculo de los gradientes de una red neuronal para ajustar los pesos [22]. Debido a las ventajas que ofrece esta modalidad de red neuronal artificial, es de las más utilizadas [17].

La función de activación utilizada para la red neuronal fue la sigmoide, también llamada función logística. Esta función de activación tiene un buen rendimiento cuando

los datos para el entrenamiento son positivos, en un rango de valores entre 0 y 1 [22]. La ecuación 2 muestra la función sigmoide:

$$f(x) = \frac{1}{1 + e^{-x}}, \tag{2}$$

para cumplir con el requerimiento de la red, se realizó una normalización de los valores, ubicándolos en un rango entre 0 y 1, a través de la ecuación 3 [23]:

$$y_i = \frac{d - x_{min}}{x_{max} - x_{min}}, \tag{3}$$

donde d, es el dato a normalizar, x_{max} es el valor máximo en la serie de datos y x_{min} el mínimo. Se determinó un número total de épocas como criterio de paro, además de un umbral de error definido por un valor épsilon [24]. Este error fue medido de acuerdo a la ecuación 4:

$$\text{Error} = (y_i - f(x_i))^2, \tag{4}$$

donde y_i es el dato conocido, y $f(x_i)$ el dato calculado por la red neuronal.

La predicción utilizada en este trabajo mediante la RNA-BP emplea el mecanismo de predicción de series temporales a un paso en el tiempo, ilustrado en la Fig. 1 [25]. Para el conjunto n de datos de entrada de la RNA-BP, este tipo de predicción permite, una vez entrenada la red, conocer el valor $n + 1$, que es desconocido.

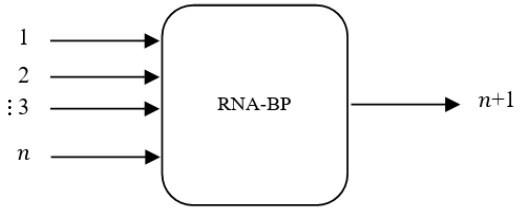


Fig. 1. Predicción de series temporales a un paso en el tiempo.

3. Metodología

Para el desarrollo de este trabajo, se contó con la colaboración del IEE, que facilitó la obtención de la base de datos generada por la estación CICEG, ubicada en la ciudad de León, Guanajuato. La base de datos cuenta con 96,408 registros, que comprenden un histórico del 1 de enero de 2005 al 31 de diciembre de 2015. Además de la medición de cinco contaminantes (PM_{10}, O_3, NO_2, SO_2 y CO), la base de datos cuenta con la información de "Año, Mes, Día, Hora y Temperatura". En este trabajo se tomaron los datos correspondientes a los contaminantes O_3, SO_2, CO y PM_{10}, debido a que no se tiene disposición de los datos de NO_2 para el intervalo de días seleccionados para las pruebas. La Tabla 3 muestra cómo está constituida la base de datos.

Primeramente, se seleccionaron 120 datos por cada contaminante, que corresponden a cinco días de mediciones, del 21 al 25 de diciembre del 2015 (lunes a viernes). Estos

datos se seleccionaron, ya que en la ciudad de León son los días de más movilización urbana. También, con base a experimentación, la cantidad de datos correspondientes a cinco días generaron buena eficacia en la predicción de la red neuronal; es decir, con un mínimo error obtenido, además de que el tiempo de procesamiento computado es corto. La importancia de estos resultados radica en que se requiere predecir el valor de calidad de aire a una hora en el futuro, para así poder notificar al usuario de las medidas preventivas a tomar. La Fig. 2 representa la metodología empleada para la predicción del valor desconocido por cada contaminante.

Tabla 3. Base de datos de niveles para los cinco diferentes contamientes, generados por la estación CICEG.

Fecha	Hora	O3	SO2	NO2	CO	PM 10	Temperatura
1/1/2005	0	7.02	10.40	151	3	362	14
1/1/2005	1	8.04	17.2	95	2.8	498	16
1/1/2005	2	10	18	101	3.3	490	17
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
31/12/2015	21	9.24	16.1	NA	2.14	82	22
31/12/2015	22	9.01	15.68	NA	2.32	101.05	21.35
31/12/2015	23	8.77	15.06	NA	1.86	90.34	20.98

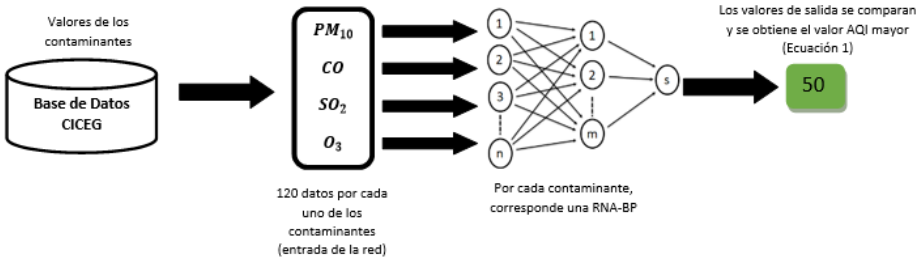


Fig. 2. Metodología utilizada para la predicción de datos de contaminantes de la base de datos generada por la estación CICEG.

Para cada uno de los cuatro contaminantes utilizados de la base de datos de la estación CICEG, se generó una red neuronal artificial Backpropagation (RNA-BP). Cada uno de los valores ingresados como entrada de la red fue normalizado en el rango de 0 a 1 para cumplir con los requerimientos de la RNA-BP, mediante la ecuación 3. La configuración que se utilizó en cada una de las RNA-BP, se muestra en la tabla 4.

Tabla 4. Configuración de la red neuronal.

Variable	Épocas	Neuronas Capa Entrada/Oculta/Salida	Criterio de paro	Taza de aprendizaje	Pesos inicializados en rango:
PM10, O3, NO2, SO y CO	50000	85/95/1	0.001	0.03	-2 a 2

Para el proceso de entrenamiento y calibración de los pesos, se tomaron 120 datos (correspondientes a cinco días de monitoreo). De éstos, 85 datos fueron tomados para entrenamiento, correspondiendo cada valor a una neurona de entrada. Para prueba, se utilizaron los 35 datos restantes. La salida de la RNA-BP fue comparada con el valor ya conocido. Este proceso se realizó hasta alcanzar el criterio de paro (0.001), o un número determinado de épocas (50000). Una vez entrenada la red, se ingresó un vector de 85 datos para conocer el dato desconocido a una hora en el futuro generado por la RNA-BP; este dato se comparó con el dato real para conocer la eficacia de la predicción. Se realizaron 35 experimentos para cada una de las RNA-BP (Fig. 2), obteniendo la mediana de los resultados, como parámetro para conocer la eficiencia de la predicción.

La configuración mostrada en la tabla 4 fue seleccionada de acuerdo a los mejores resultados dados por experimentación previa, empleando diferentes parámetros. Una vez obtenido el valor del contaminante introducido en la red neuronal, se hace el proceso inverso a la normalización, haciendo uso de la ecuación 5, la cual se obtiene de despejar la ecuación 3:

$$y = d * (x_{max} - x_{min}) + x_{min}\,, \tag{5}$$

donde d, es el dato normalizado, x_{max} es el valor máximo en la serie de datos normalizados y x_{min} el mínimo. Una vez obtenido el dato a su forma original, éste se introduce en el sistema para, mediante la ecuación 1, conocer cuál es el nivel de calidad del aire. Estos niveles de calidad del aire se comparan entre sí para determinar cuál es el que genera un AQI mayor. El mayor determina el mensaje de alerta que es enviado al usuario.

4. Resultados

Una vez entrenada la red, para el proceso de validación, se agregó un vector cuya salida no era conocida por la red. La salida no conocida correspondía a la hora siguiente (1:00 horas) del día siguiente (sábado 26 de diciembre de 2015) a los seleccionados para entrenamiento y prueba. El dato de validación es conocido para verificar la eficacia de la red neuronal. Dado que la salida de la red es un dato que se encuentra normalizado, se realizó el proceso inverso a la ecuación 3, para conocer el dato real y compararlo, resultando la ecuación 5. En la Tabla **5** se aprecian las medianas de los resultados que arrojaron los 35 experimentos realizados en la predicción de los datos seleccionados de la estación CICEG con la red neuronal.

Tabla 5. Resultado de los 35 experimentos realizados con la red neuronal.

	PM	**CO**	**SO2**	**O3**
Real	54.78	1.49	10.41	8.58
Salida de la red	52.45	1.56	10.64	10.77

Los valores de los contaminantes fueron substituidos en la Ecuación 1 AQI, para conocer cuál era el que determinaba el nivel de calidad del aire. En la Tabla 6 se aprecian los resultados.

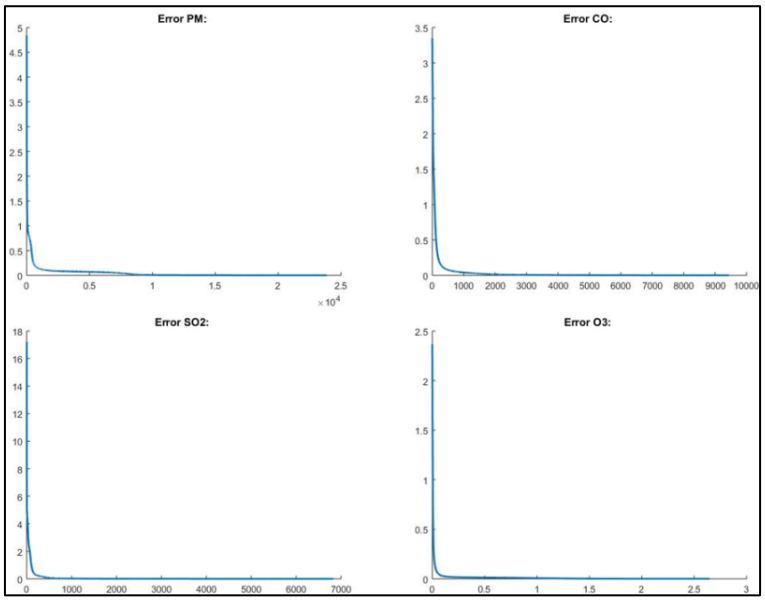


Fig. 3. Comportamiento del error cuadrático medio en el proceso de entrenamiento de la red para los cuatro contaminantes.

Tabla 6. Valor AQI, para los resultados de los experimentos de la red neuronal mostrados en la tabla 5.

	PM10	**CO**	**SO2**	**O3**
AQI	48.56	8.7	5.37	40.9
Categoría	Bueno	Bueno	Bueno	Bueno

El comportamiento del error observado en el proceso de entrenamiento de la red, para cada uno de los contaminantes anteriores (Tabla 6), se observa en la Fig. 3.

En cada uno de los experimentos realizados se obtuvo el criterio de alcanzar un error igual a 0.001, en la etapa de entrenamiento de la RNA-BP. La mediana de los errores finales (tras 35 experimentos), para cada contaminante, se aprecia en la Tabla 7.

Tabla 7. Resultados de las medianas del valor de los errores de los 35 experimentos realizados.

	PM10	**CO**	**SO2**	**O3**
AQI	0.000999912	0.000999902	0.000999751	0.000999991

Como se observa en la Tabla 6, el índice AQI que mayor valor mostró fue el contaminante PM10. Debido a que el índice AQI se toma en base al contaminante que se encuentra en mayor proporción en el ambiente, el AQI para la hora predicha por la red sería 48.56, que equivale a la categoría de calidad del aire "Bueno". La Figura 4 muestra una recomendación para esta categoría.

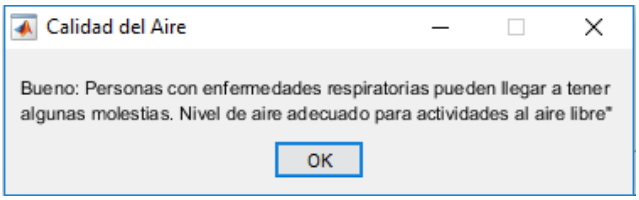


Fig. 4. Ejemplo de mensaje enviado al usuario, al conocer el nivel de calidad del aire predicho por la red.

5. Conclusiones y trabajo futuro

La técnica utilizada para este trabajo, su arquitectura y configuración, y los datos obtenidos con los que se realizaron los experimentos, demostraron una vez más, el buen funcionamiento de las redes neuronales Backpropagation al momento de realizar predicciones. Como se puede observar en la tabla 5, los valores obtenidos en la salida de la red neuronal son muy cercanos a los valores reales (no conocidos por la red neuronal) de los contaminantes medidos por la estación CICEG. Debido a que los resultados permiten tener confianza en las predicciones de la red neuronal y así, poder mantener al tanto a las personas del nivel de calidad del aire que pueden experimentar en el futuro cercano, manteniéndolas informadas para tomar acciones preventivas y no tener que sufrir reacciones adversas en su salud, se concluye que para los datos con los que se trabajó, la arquitectura de la red neuronal genera buenos resultados.

Es importante mencionar que los valores de las mediciones de contaminantes obtenidos por la estación CICEG varían mucho de una hora a otra, esto debido a factores como el nivel de tráfico, los horarios de trabajo en fábricas cercanas a la estación, condición climática, entre otros. Como este puede variar (para bien o para mal) de una hora a otra, el sistema es capaz de enviar alertas con base a las observaciones de la estación. Así, cuando la red neuronal falle en alguna predicción, se mantiene al mínimo el nivel de riesgos en la salud de las personas con las enfermedades ya mencionadas. Se espera con esto el incentivar a las personas que sufren de padecimientos del tipo respiratorio, al estar más al tanto de la calidad del aire del entorno en el que se desarrollan. Manteniéndose informados, pueden decidir en qué hora del día es mejor para su salud realizar actividades al aire libre, por citar un ejemplo. Basándonos en la opinión de expertos, el problema de la contaminación es algo que continuará existiendo y que seguirá siendo un foco de atención para la sociedad, y al mantenerse informado se puede crear conciencia para cuidar más el medio ambiente.

Como trabajo futuro se propone realizar el procesamiento de las bases de datos generadas por las demás estaciones ubicadas en la ciudad de León y así poder tener un panorama general del nivel de calidad del aire en la ciudad. Además, se propone buscar más información con expertos en el área de las alergias, para que las recomendaciones emitidas sean más precisas y se vayan adaptando conforme se vayan dando nuevos descubrimientos en el problema. También, se propone utilizar otras herramientas de predicción como el ajuste de datos por mínimos cuadrados, y realizar experimentos para ver cual ofrece mejor desempeño en la predicción y/o coste computacional u otras características que ayuden a agilizar al sistema de predicción y así emitir mejores alertas y con mayor rapidez. Además, emplear sensores que nos permitan conocer la calidad

del aire en interiores, ya que dependiendo de los contaminantes que existen en interiores se puede inclusive, llegar a generar asma estacional.

Agradecimientos. Se agradece el apoyo económico brindado al Consejo Nacional de Ciencia y Tecnología (CONACYT), para la realización de este trabajo. También, agradecimientos especiales al Dr. M. Ornelas-Rodríguez y a los ingenieros D. Duarte-Carrera y A. Godínez-Bautista, por sus apreciados comentarios.

Referencias

1. Yang, X., Du, J., Liu, S., Li, R., Liu, H.: Air pollution source estimation profiling via mobile sensor networks. Int. Conf. Comput. Inf. Telecommun. Syst. (2016)
2. Baralis, E., Cerquitelli, T., Chiusano, S., Garza, P., Kavoosifar, M. R.: Analyzing air pollution on the urban environment. pp. 1464–1469 (2016)
3. Bose, B.: Global Warming: Energy, Environmental Pollution, and the Impact of Power Electronics. IEEE Ind. Electron. Mag., Vol. 4, No. 1, pp. 6–17 (2010)
4. Guo, D., Zhang, Y., He, L., Zhai, K., Tan, H.: Chebyshev-polynomial neuronet, WASD algorithm and world population prediction from past 10000-year rough data. In: Proc. 27th Chinese Control Decis. Conf. CCDC, pp. 1702–1707 (2015)
5. Melorose, J., Perroy, R., Careas, S.: World population prospects. United Nations, Vol. 1, No. 6042, pp. 587–92 (2015)
6. C. de N. ONU.: Proyectan aumento en población urbana para 2050. United Nations, Disponible en: http://www.un.org/spanish/News/story.asp?NewsID=11754#. WNGX_Ts1-Um (2008)
7. Garrido-Lestache, J. S., Rodríguez-García, V.: Las enfermedades alergicas (2012)
8. Onatra, W., Vargas, S., Paez, E., Rojas, D., Lopez, A.: Correlación entre la enfermedad respiratoria aguda (era) en mujeres embarazadas y la calidad del aire. Rev. U.D.C.A Act. Div. Cient., Vol. 12, No. 2, pp. 27–37 (2009)
9. Saad, S. M., Saad, A. R. M., Kamarudin, A. M. Y., Zakaria, A., Shakaff, A. Y. M.: Indoor air quality monitoring system using wireless sensor network (WSN) with web interface. In: Int. Conf. Electr. Electron. Syst. Eng., pp. 60–64 (2013)
10. Alcalá-padilla, G.: Prevalencia de sensibilización a alergenos en niños escolares con asma que viven en la zona metropolitana de Guadalajara. Rev. Alerg. Mex., Vol. 63, No. 2, pp. 135–142 (2016)
11. de la Luz Cid, M.: Estudio aerobiológico de pólenes anemófilos en la ciudad de Toluca, México. Rev. Alerg. Mex. Vol. 62m No. 1, pp. 8–14 (2015)
12. Baena-Cagnani, C. E.: Actualización de rinitis alérgica y su impacto en el asma (ARIA 2008). La perspectiva latinoamericana. Rev. Alerg. Mex., Vol. 56, No. 2, pp. 56–63 (2009)
13. Sánchez-González, A.: ARIA México 2014 Adaptación de la Guía de Práctica Clínica ARIA 2010 para México. Metodología ADAPTE ARIA Mexico 2014 Adaptation of the Clinical Practice Guide ARIA 2010 for Mexico. Vol. 61, pp. 3–116 (2014)
14. Pieri, T., Michaelides, M. P.: Air Pollution Monitoring in Lemesos using a Wireless Sensor Network. pp. 18–20 (2016)
15. Gobierno del Estado de Guanajuato e Instituto de Ecología del Estado.: Informe

de Estado y Tendencia de la Calidad del Aire Guanajuato 2014. Guanajuato (2014)
16. Bagieński, Z.: Traffic air quality index. Sci. Total Environ., Vol. 505, pp. 606–614 (2015)
17. Zhang, G.: Prediction of Short-Term Wind Power in Wind Power Plant based on BP-ANN. pp. 75–79 (2016)
18. Guanqun, D., Fataliyev, K., Wang, L.: One-step and multi-step ahead stock prediction using backpropagation neural networks. In: ICICS 2013 - Conf. Guid. 9th Int. Conf. Information, Commun. Signal Process., No. weights 2, pp. 2–6 (2013)
19. Kyrkilis, G., Chaloulakou, A., Kassomenos, P. A.: Development of an aggregate Air Quality Index for an urban Mediterranean agglomeration : Relation to potential health effects. Sci. Direct, Vol. 33, No. 2, pp. 670–676 (2007)
20. García, B.: Los Índices de Calidad del Aire: Alcances y Limitaciones. En: Concienc. Tecnológica, pp. 74–76 (2011)
21. Heaton, J.: Artificial Intelligence for Humans. Deep Learning and Neural Networks, Vol. 3 (2015)
22. Heaton, J.: Introduction to the Math of Neural Networks (2012)
23. Heaton, J.: Artificial Intelligence for Humans. Volume 1: Fundamental Algorithms. Heaton Research. Incorporated (2013)
24. Kandel, A.: Introduction to pattern recognition: statistical, structural, neural, and fuzzy logic approaches, Illustrate. Vol. 32, World Scientific (1999)
25. Isasi-Viñueala, P., Galván-Leon, I. M.: Redes de neuronas artificiales: un enfoque práctico. Pearson Education (2004)

ISSN 1870-4069

Aplicación de modelos auto regresivos para la predicción de generación de energía eléctrica a partir de datos eólicos

Sara Edith Pinzón Pineda[1], José Alberto Hernández Aguilar[2], Gustavo Arroyo-Figueroa[1]

[1] Instituto Nacional de Electricidad y Energías Limpias, Ciudad de México, México

[2] Universidad Autónoma del Estado de Morelos, Cuernavaca, Morelos, México

sepinzon@iie.org.mx, jose_hernandez@uaem.mx, garroyo@iie.org.mx

Resumen. Se discuten los modelos estadísticos auto regresivos clásicos *Box and Jenkins* utilizados en problemas de predicción, y se analiza su aplicación en el sector energético. Posteriormente, se propone una metodología basada en la aplicación de estos modelos para la predicción de generación de energía eléctrica a partir de datos obtenidos en un parque eólico de la Ventosa, Oaxaca Finalmente, se presenta el modelo y los resultados preliminares obtenidos destacando, que se alcanza un $R^2 = 0.974$ y un RMSE = 18.862

Palabras clave: modelos auto regresivos, *Box and Jenkins*, predicción, energía eléctrica, energía eólica, datos eólicos.

Application of Autoregressive Models for the Prediction of Electric Power Generation from Wind

Abstract. We discuss the classic auto regressive statistical models Box and Jenkins used in prediction problems, and their application is analyzed in the energy sector. Subsequently, we propose a methodology based on the application of these models for the prediction of electric power generation, from data obtained at a wind farm in Ventosa, Oaxaca. Finally, the model and the preliminary results obtained shows an R2 = 0.974 and an RMSE = 18.862.

Keywords: autoregressive models, Box and Jenkins, prediction or forecasting, electric power, wind power, wind data, wind power prediction.

1. Introducción

Actualmente se está alcanzando el límite de capacidad de carga dentro de los ecosistemas para regenerarse de la contaminación producida por la actividad humana. De acuerdo a [1]: "Un tercio del total de la contaminación generada a nivel mundial

proviene del proceso de producción de electricidad". Por lo que, es necesario el desarrollo de fuentes de energías renovables.

De acuerdo a [2]: "La energía eólica tiene su origen en la energía solar. El calentamiento de masas de aire como consecuencia de la radiación solar, contribuye a la aparición de corrientes ascendentes, siendo el espacio que liberan ocupado por otras masas adyacentes de aire más frío. La desviación del viento es proporcional a la velocidad del mismo: A mayor velocidad mayor desviación con respecto a la dirección inicial".

Al viento se le considera un recurso energético y también como suministro. Su característica particular lo describe como una fuente con importantes variaciones temporales, tanto en su superficie como en su altura, al que se agrega una componente aleatoria que afecta en buena medida su variación total [2].

Por este motivo, se considera importante predecir la velocidad del viento para la planificación de la cantidad de energía que se produce en el tiempo [1]. De acuerdo a lo anterior, se debe evaluar y planificar la producción eólica debido a la incorporación de parques eólicos en el mercado eléctrico, por lo que es necesario realizar ofertas de energía para plazos que oscilan entre 24 y 36 horas, lo que requiere un conocimiento preciso de las condiciones de funcionamiento de los parques durante esos periodos.

La dificultad de estas predicciones radica en el grado de detalle y la cantidad de variables a analizar, y de la necesidad de contar con horizontes de tiempo [2].

Debido a lo anterior se han realizado diversas investigaciones de predicción de generación del viento, basados en modelos matemáticos, estadísticos y de inteligencia artificial [3], [4].

Problema de investigación

Aplicar la Metodología Box-Jenkins para la predicción de la generación de energía eléctrica a partir de datos obtenidos de un parque eólico localizado en la Ventosa, Oaxaca, México.

Para este propósito, esta investigación se ha dividido en cuatro secciones: en la primera sección, se describe la Introducción y se señala el problema de investigación, en la segunda sección, se presenta la metodología empleada en la cual se utilizan como base los modelos auto regresivos clásicos basados en la metodología Box-Jenkins utilizando el software SPSS v.22. En la tercera sección, se presentan y discuten los resultados obtenidos. Finalmente, se presentan las conclusiones y trabajos futuros.

Los datos, fueron obtenidos de 3 periodos históricos proporcionado por el Centro Regional de Tecnología Eólica en el estado de Oaxaca, que es considerado como el quinto más grande del país, está ubicado en la parte sur de México, con una superficie de aproximadamente 95.364 km2. Cuenta con una orografía de zonas montañosas, mesetas planas, valles y costas, con una gran variedad de climas, tropical en toda la costa y templado en el interior. La lluvia aparece generalmente a finales de abril y continúa hasta finales de octubre. La temperatura media oscila entre 26 y 28 1C en toda la costa [5]. Los datos con los que se realiza el análisis se tomaron de una muestra de 3 periodos históricos de una torre anemométrica a 80m. Los periodos específicos son del 2007 (enero a diciembre), 2012 (febrero a diciembre) 2013 (enero) y 2015 (enero a diciembre) del CERTE en el estado de Oaxaca. Esta zona se caracteriza por la complejidad de su terreno que es: montañosa en el Norte a 5 Km y plano en el Este, Sur y Oeste.

2. Metodología

Para llevar a cabo el desarrollo de esta metodología, se utilizaron datos del Centro Regional de Tecnología Eólica – CERTE, localizado en la localidad de la Ventosa del Istmo de Tehuantepec del estado de Oaxaca, que es considerado como el quinto más grande del país, está ubicado en la parte sur de México, con una superficie de aproximadamente 95.364 km2. Cuenta con una orografía de zonas montañosas, mesetas planas, valles y costas, con una gran variedad de climas, tropical en toda la costa y templado en el interior. La lluvia aparece generalmente a finales de abril y continúa hasta finales de octubre. La temperatura media oscila entre 26 y 28 1C en toda la costa [5].

El parque eólico CERTE de acuerdo a [6]: fue construido con el apoyo económico del Fondo Mundial para el Medio Ambiente (GEF por sus siglas en inglés) a través del Programa de las Naciones Unidas para el Desarrollo (PNUD), como parte de las metas del proyecto "Plan de Acción para Eliminar Barreras para el Desarrollo de la Generación Eolo eléctrica en México". El 1 de julio de 2010 al haber cumplido con los requisitos y pruebas establecidas por la Comisión Federal de Electricidad (CFE), entró en operación normal el CERTE perteneciente de la Gerencia de Energías No Convencionales del anterior Instituto de Investigaciones Eléctricas - IIE que actualmente fue renombrado Instituto Nacional de Electricidad y Energías Limpias – INEEL.

Es un centro de investigación y desarrollo tecnológico que tiene como objetivos además de buscar dar solución a los problemas de adecuación, instalación y mantenimiento, el desarrollo de sus propias tecnologías para fortalecer y potencializar el sector eólico nacional.

A continuación se describen los modelos auto regresivos clásicos, posteriormente se presentan los pasos empleados para el procesamiento de los datos obtenidos del parque eólico para la predicción de la generación de la energía eléctrica.

2.1. Modelos para predicción o pronóstico

Los métodos para predecir la generación de energía eólica pueden ser categorizados en métodos físicos, métodos estadísticos, métodos basados en redes neuronales, y métodos híbridos [7].

2.2. Modelos auto regresivos clásicos para predicción o pronóstico

En 1970, Box and Jenkins desarrollaron un cuerpo metodológico destinado a identificar, estimar y diagnosticar modelos dinámicos de series temporales en los que la variable tiempo juega un papel fundamental [1].

De acuerdo a [8]:

"Este modelo implica identificar un proceso ARIMA adecuado realizando el ajuste de los datos y luego realizando la predicción. El procedimiento original del modelo Box-Jenkins implicó un proceso iterativo de tres etapas:

1. Identificación o selección del modelo,
2. Estimación de los parámetros,

3. Verificación del modelo.

Explicaciones recientes del proceso (por ejemplo, Makridakis, Wheelwright y Hyndman, 1998) [9] añaden una etapa de pre procesamiento de datos y una representación final de la aplicación del modelo o pronóstico.

El modelado de Box-Jenkins implica identificar un proceso ARIMA adecuado, ajustándolo a los datos, y luego utilizando el modelo de previsión. Una de las características atractivas del enfoque de Box-Jenkins para la predicción es que los procesos ARIMA son una clase muy rica de posibles modelos y suele ser posible encontrar un proceso que proporcione una descripción adecuada de los datos".

Pasos para el desarrollo del modelo según [8] [9] se deberá efectuar una serie de etapas que se describen a continuación:

Preparación de los datos (Pre procesamiento de los datos)

En esta etapa, se realiza la limpieza de los datos y se puede hacer un análisis exploratorio de los mismos, también de manera gráfica se puede observar si la media y la varianza son constantes en el tiempo para determinar a grandes rasgos si la serie de tiempo es estacionaria.

En caso de no ser estacionaria, se pueden realizar transformaciones o diferenciaciones de los datos. Las transformaciones de los datos pueden ser a través de raíces cuadradas o logarítmicas y se utilizan para estabilizar la varianza en las series de tiempo.

La diferenciación de los datos se realiza tomando observaciones consecutivas o de años anteriores, generalmente para casos donde la varianza no permanece constante en el tiempo.

Box-Jenkins de acuerdo a [9] **y** [10] propone tres pasos para el desarrollo del modelo**:**

1. **Identificación y selección del modelo.** Este modelo utiliza varios gráficos basados en transformaciones y diferenciaciones para tratar de identificar posibles procesos ARIMA los cuales podrían proporcionar un buen ajuste a los datos de los desarrollos posteriores. Existen algunas herramientas de selección de modelos como el Criterio de Información de Akaike o los criterios ACF y PACF.

2. **Estimación de parámetros.** Se deben encontrar los valores para los coeficientes del modelo que proporcionen un mejor ajuste de los datos.

3. **Comprobación del modelo mediante ensayo.** Consiste en probar las suposiciones del modelo para identificar cualquier área donde el modelo es inadecuado. Si el modelo se considera inadecuado, es necesario volver al Paso 2 y tratar de identificar un modelo mejor.

Realización del Pronóstico.

Entre las técnicas univariantes existen algunas sencillas, como por ejemplo los modelos autoregresivos de primer orden, o los modelos de tendencia lineal o exponencial, mientras que otras técnicas resultan más complejas. [11]

Existen diversas técnicas para hacer pronóstico de series de tiempo de las cuales algunos autores afirman que la metodología Box and Jenkins es muy efectiva con el pronóstico a corto plazo. [11]

En la figura 1 se detalla de manera general el comportamiento del modelo:

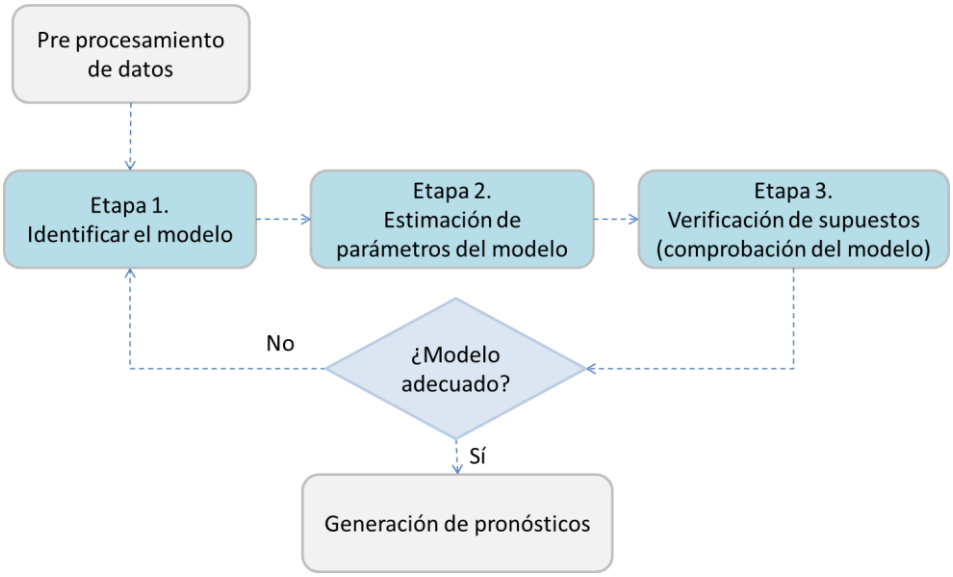


Fig. 1. Etapas de la Metodología Box - Jenkins.

En la Figura 2 se muestran los modelos clásicos Box – Jenkins para la predicción del viento:

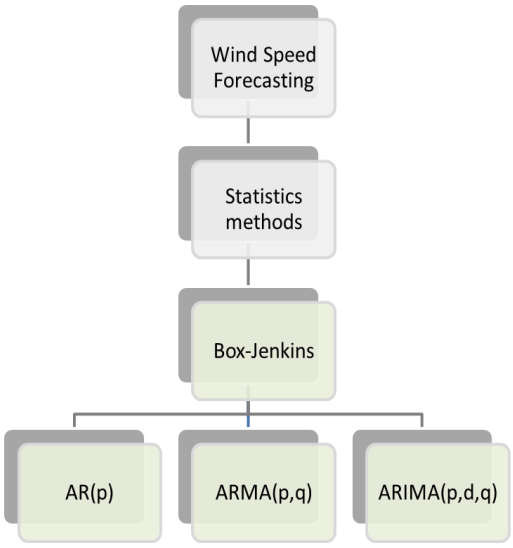


Fig. 2. Modelos clásicos Box-Jenkins para predicción o pronóstico del viento.

A continuación, de acuerdo a [12] y [8] se describen las ecuaciones los modelos Estadísticos Box and Jenkins:

Modelo Auto regresivo de primer orden AR(p). Sus valores expresan la dependencia de $\mathbf{X_t}$ en sus valores pasados. Definido mediante la siguiente ecuación:

$$X_t = \Phi X_{t-1} + \mathcal{E}_t, \quad (1)$$

dónde:

X_t= Valores de correlación de la serie de tiempo,
X_{t-1} = Valor pasado (Variable determinística fija que al no ser fija se convierte en modelo dinámico),
Φ = Constante definida por un tiempo,
$\mathcal{E}_t$ = Residuo desconocido.

Modelo Auto regresivo de Medias Móviles ARMA(p,q). Es un modelo donde se involucran un término de orden auto regresivo y un término de media móvil. Es una combinación entre AR y MA (Medias móviles). Su ecuación se define por:

$$X_t = \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + \theta_1 \mathcal{E}_{t-1} + \mathcal{E}_t, \quad (2)$$

dónde:

X_t = Suma de la parte independiente no correlacionadas entre sí y sus pasados inmediatos anteriores,
$\Phi_1, \Phi_2 =$ Constantes definidas en el tiempo y su valor pasado a esa constante,
X_{t-2} = Valor precedente al X_{t-1},
$\theta_1 \mathcal{E}_{t-1}$ = Residuo de orden a_{t-1} (Residuo que se puede conocer),
$\mathcal{E}_t$= Residuo desconocido.

Modelo Auto regresivo Integrado de Medias Móviles ARIMA(p,d,q) - No estacional. Si combinamos la diferenciación con auto regresión y un modelo de media móvil, obtenemos un modelo ARIMA no estacional. ARIMA es un acrónimo para el modelo *AutoRegressive Integrated Moving Average* ("integración" en este contexto es el inverso de la diferenciación). El modelo completo puede escribirse como [9]:

$$Y'_t = C + \Phi_1 Y'_{t-1} + \cdots + \Phi_p Y'_{t-p} + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} + e_t, \quad (3)$$

dónde Y'_t es la serie diferenciada (Puede haber sido diferenciada más de una vez). Los "Predictores" en el lado derecho incluyen tanto los valores de retardo de Y_t y los errores de retardo. Según [9] y [13] el modelo es llamado ARIMA(p,d,q) cuando:

p = El orden de la parte auto regresiva (El número de términos auto regresivos).
d = Grado de la diferenciación involucrada (Número de diferencias no estacionales necesarias para generar la estacionariedad).
q = El orden de la parte de media móvil (El número de errores de predicción de retardo en la ecuación de predicción).

3. Diseño experimental

A continuación se describen los pasos utilizados de acuerdo a la metodología para analizar y procesar los registros obtenidos de tres periodos completos (del 2007 al 2010) del parque eólico.

3.1. Pre procesamiento de los datos

El método propuesto está basado en los trabajos de [14] [15], Del primer utiliza el enfoque para el análisis de los datos y pronóstico de fenómenos físicos mediante modelos auto regresivos y del segundo se toma la metodología KDD (Knowledge Discovery from Data – Descubrimiento de Conocimiento de los Datos) para la obtención, pre procesamiento, limpieza y almacenamiento de los datos, y visualización e identificación de patrones. Cabe señalar que para esta investigación se utilizó el software SPSS Versión 22 de IBM y R en su versión 3.2.

En la figura 1. Se muestra la curva de potencia del aerogenerador del modelo T300-28 del fabricante Turbowinds (Bélgica), de potencia máxima de 300 kw [16], modelo que se encuentra ubicado dentro del parque eólico y bajo la cual se realizó la interpolación de datos para simular la potencia generada.

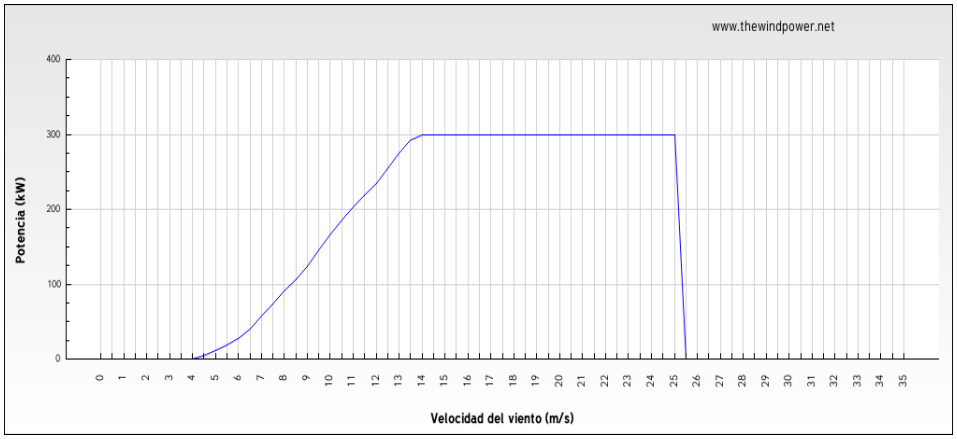


Fig. 3. Curva de potencia del modelo T300-28.

3.2. Identificación del modelo

1. Selección de la muestra

La muestra consistió de 157824 registros correspondientes a las lecturas diarias tomadas con una frecuencia de 10 minutos durante el periodo.

2. Visualización de los datos

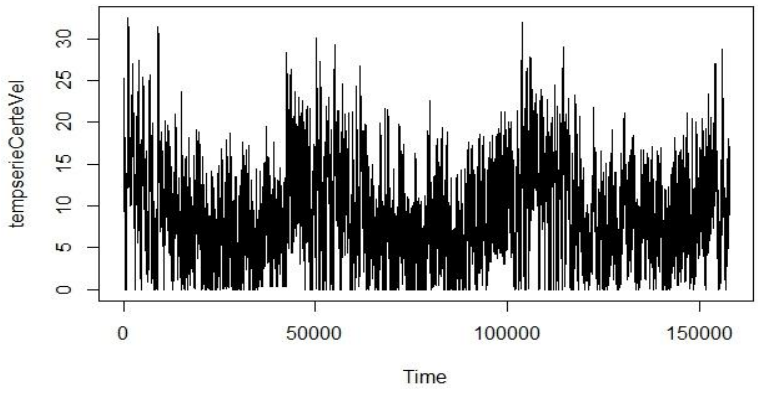
En esta etapa se analiza el comportamiento y tendencias de los datos, se evalúa si existen comportamientos estacionarios dentro de la serie de tiempo.

Para graficar la Información se utilizó el módulo de predicciones en su opción gráficos de secuencia de SPSS versión 22 y R en su versión 3.2 [17].

Los resultados de este análisis preliminar se muestran en las Figuras 4, 5 y 6. En la figura 5 se muestra la dirección del viento medida a partir de los datos eólicos. En la figura 6 se muestra la rosa del viento en donde se aprecia que el viento proviene preponderantemente del norte y noroeste. De acuerdo al análisis exploratorio de los datos, se concluye que la serie de tiempo es estacionaria ya su tendencia se mantiene estable (Priestley, 1988).

3.3. Identificación de los parámetros del modelo

Para la identificación de los parámetros del modelo se utilizó el módulo de Predicción de SPSS versión 22 a partir del cual se creó el modelo utilizando la opción: Crear Modelos de la figura 7.



.

Fig. 4. Grafica preliminar de los datos a partir de la velocidad del viento a 80 m (fuente propia) de los datos procesados en R versión 3.2.

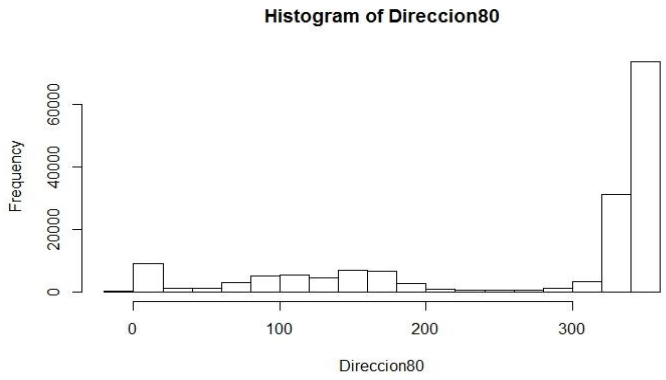


Fig. 5. Histograma de los datos de la dirección del viento a 80 m (fuente propia) procesados en R Versión 3.2.

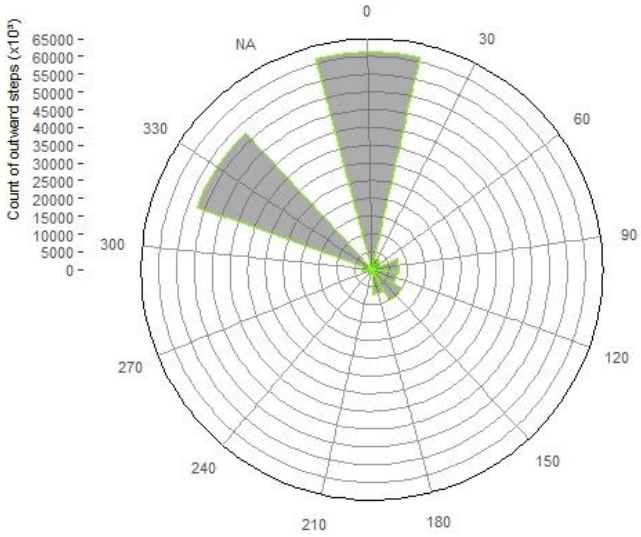


Fig. 6. Rosa de viento a partir de los datos a una altura de 80 m (fuente propia) de la dirección en R Versión 3.2.

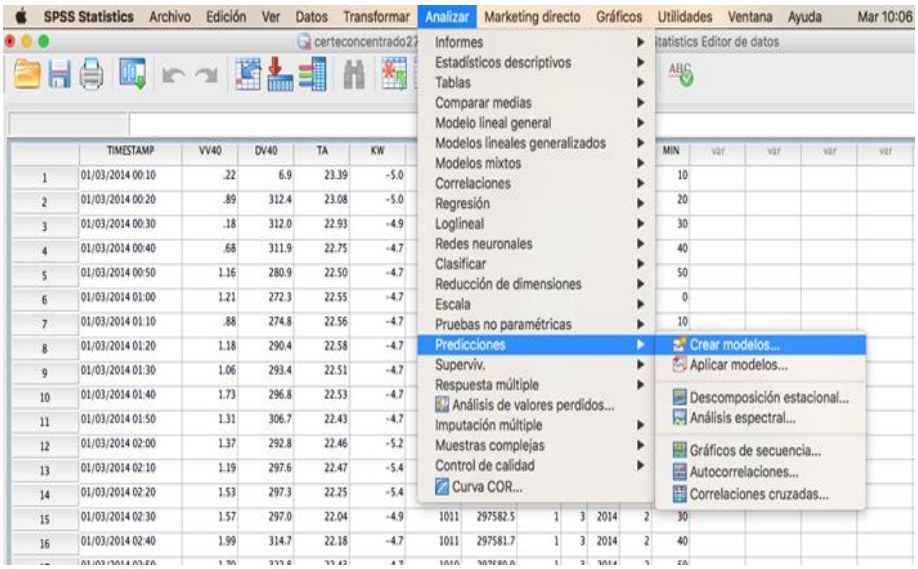


Fig. 7. Módulo para la creación de modelos predictivos.

El generador de modelos que se utilizó fue el experto, el cual permite seleccionar la variable dependiente y las variables independientes o predictoras. Se probó con las variables existentes Velocidad del Viento (VV80), dirección del viento (DV-80), y se encontró que la velocidad del viento es la principal variable predictora.

4. Resultados

4.1. Verificación o comprobación del modelo

Para la prueba del modelo se seleccionaron tres muestras, la primera de 50000 al 32%, la segunda 126580 que corresponde al 81% y la tercera de 154824 del 100% de los datos. Para ello se utilizó el módulo "Aplicar modelos", disponible en el menú de Analizar -> Predicciones -> Aplicar modelos.

Los resultados se presentan en la Tabla 2, en él se resumen la R^2 estacionaria y el RMSE.

Tabla 1. Comparativo para los diferentes modelos de predicción generados.

Experimento	n=50000	n=126580	n=154824
Variable predictora	Velocidad del Viento y Dirección del Viento	Velocidad del viento	Velocidad del viento
R cuadrado	0.966	0.974	0.972
RMSE	21.654	18.862	19.792
Modelo generado	ARIMA(5,1,5)	ARIMA(3,1,3)	ARIMA(3,1,5)

Con el comparativo anterior las variables más adecuadas para predecir la generación de energía eléctrica es la velocidad del viento de acuerdo a la mayor R cuadrada y la menor RMSE.

Tabla 2. Descripción del mejor.

		Tipo de modelo
ID modelo	W Modelo_1	ARIMA(3,1,3)

El modelo anterior nos indica que para predecir la generación de energía eléctrica a partir de la velocidad del viento para nuestros datos es un modelo ARIMA(p,d,q*)* con:
p = 3,
d = 1,
q = 3.

5. Conclusiones y trabajo a futuro

Los resultados demuestran la necesidad de tener una base de datos confiable para poder predecir la generación de energía eléctrica a partir de fuentes eólicas. Es indispensable que esta base de datos se valide por expertos en el área, en el sentido de que los datos representen la realidad y tengan sentido.

A partir de múltiples experimentos se pudo obtener un primer modelo predictivo ARIMA (3,1,3), con 1 sola variable predictora que corresponde a la velocidad del

viento. Cabe mencionar que aún se requiere la prueba del modelo y la validación del mismo, lo cual es un área en la que aún se está trabajando.

Los resultados aquí presentados no son de ninguna manera los finales, representan tan solo los primeros pasos para la generación de un modelo robusto que nos permita la Predicción de la Generación de energía eléctrica. Este ejercicio está en línea con lo publicado en la literatura, en los que se sabe que la predicción del viento se encuentra correlacionada con la velocidad del mismo.

La relevancia de este trabajo radica en una primera aproximación para la comprensión de la predicción de la generación de energía eléctrica a partir de modelos auto regresivos.

Parte del trabajo futuro consiste en implementar los modelos en lenguaje R para disminuir los tiempos de ejecución que se llevaron a cabo con SPSS.

Referencias

1. De Arce, R., Mahía, R.: Modelos Arima. Programa CITUS: Técnicas de Variables Financieras (2003)
2. Diez, R. P., Fernández, J. P., Priore, P., Gómez, A. G.: Pronóstico de la Velocidad y Dirección del Viento mediante Redes Neuronales Artificiales. In: VIII Congreso de Ingeniería de Organización, pp. 905–914 (2004)
3. Lawan, S. M., Wawzawy, C., Baharun, A., Masri, T.: Different models of wind speed prediction: A compre hensive review. International Journal of Scientific and Engineering Research, pp. 1760–1768 (2014)
4. Wang, X., Guo, P., Huang, X.: A review of wind power forecasting models. In: Energy Procedia, The Proceedings of International Conference on Smart Grid and Clean Energy Technologies (ICSGCE 2011), 12, pp. 770–778 (2011)
5. Cadenas, E., Rivera, W.: Wind speed forecasting in the south coast of Oaxaca, Mexico. Renewable energy, 32(12), pp. 2116–2128 (2007)
6. López, R. A. A.: Transferencia Tecnológica y Creación de Capacidades Tecnológicas, en el Enfrentamiento al Cambio Climático: El caso de Proyectos MDL en Energía Eólica en México.
7. Liu, Z., Gao, W., Wan, Y. H., Muljadi, E.: Wind power plant prediction by using neural networks. In: Energy Conversion Congress and Exposition (ECCE), IEEE, pp. 3154–3160 (2012)
8. Box, G. E., Jenkins, G. M., Reinsel, G. C., Ljung, G. M.: Time series analysis: forecasting and control. John Wiley & Sons (2015)
9. Hyndman, R. J., Athanasopoulos, G.: Forecasting: principles and practice. OTexts (2014)
10. Pankratz, A.: Forecasting with univariate Box-Jenkins models: Concepts and cases. John Wiley & Sons, Vol. 224 (2009)
11. Guerrero, J. F. J., Fernández, R. S., Abad, J. C. G.: La capacidad predictiva en los métodos Box-Jenkins y Holt-Winters: una aplicación al sector turístico. Revista Europea de Dirección y Economía de la Empresa, 15(3), pp. 185–198 (2006)
12. Statistical forecasting: Notes on regression and time series analysis, 26 02 2016. Disponible en: http://people.duke.edu/~rnau/411home.htm
13. Santamaría-Bonfil, G., Reyes-Ballesteros, A., Gershenson, C.: Wind speed forecasting for wind farms: A method based on support vector regression. Renewable Energy, 85, pp. 790–809 (2016)
14. Han, J., Pei, J., Kamber, M.: Data mining: concepts and techniques. Elsevier (2011)
15. Wind Power: The Wind Power. Disponible en: http://www.thewindpower.net/turbine_es_276_turbowinds_300.php (2017)

16. IBM: IBM SPSS Software. Disponible en: http://www.ibm.com/analytics/us/en/technology/spss (2017)
17. Priestley, M. B.: Non-linear and non-stationary time series analysis (1988)
18. Jiménez, L. A. F.: Modelos avanzados para la predicción a corto plazo de la producción eléctrica en parques eólicos. Doctoral dissertation, Universidad de La Rioja (2007)
19. Torgo, L., Torgo, L.: Data mining with R: learning with case studies. Boca Raton, FL: Chapman & Hall/CRC (2011)
20. Mattig, I. E.: Predicción de la potencia para la operación de parques eólicos. Memoria para optar al título de Ingeniero civil electricista. Universidad de Chile (2011)
21. CERTE: Centro Regional de Tecnología Eólica (CERTE 2014). Disponible en: https://www.ineel.mx//detalle-sede-oaxaca.html

ISSN 1870-4069

Reconocimiento de actividades infantiles utilizando sonido ambiental: Un enfoque preliminar

Antonio García-Domínguez, Carlos E. Galván-Tejada

Universidad Autónoma de Zacatecas, Unidad Académica de Ingeniería Eléctrica, Zacatecas, Zac., México

antonio.garcia@uaz.edu.mx, ericgalvan@uaz.edu.mx

Resumen. El reconocimiento de actividades humanas es un nicho de interés por diferentes entidades, como lo son médicas, empresariales para-automatización y aquellas enfocadas a las casas habitación de los usuarios promedio, sin embargo, el estudio de actividades realizadas por infantes de 12 a 36 meses suele quedar fuera del estudio por ser un grupo complejo de focalizar. Debido a lo anterior, en este trabajo se propone el estudio de actividades infantiles y clasificarlas utilizando un modelo de agrupamiento jerárquico utilizando la información contenida dentro del espectro de los sonidos emitidos al realizar las actividades. Este estudio preliminar demuestra que el espectro tiene suficiente información para clasificar correctamente las actividades infantiles y estimar dicha actividad.

Palabras clave: reconocimiento de actividades infantiles, sonido ambiental, espectro de Fourier, agrupamiento jerárquico aglomerativo.

Child Activity Recognition Using Environmental Sound: A Preliminary Approach

Abstract. Human activity recognition is a topic of interest for different entities, such as medical, bussines for automation and those focused on the homes of average users, however, the study of activities performed by infants from 12 to 36 months usually be out of the study because it is a complex focus group. Due to the above, in this work we propose the study of child activities and classify them trhough a hierarchical grouping model using the information contained within the spectrum of the sounds emitted during the performance of the activity. This preliminary study shows that the spectrum has enough information to correctly classify children activities and to estimate such activity.

Keywords: child activity recognition, environmental sound, Fourier spectrum, agglomerative hierarchical grouping.

1. Introducción

En los últimos años, la inteligencia ambiental ha sido objeto de numerosos estudios e importantes investigaciones, con la finalidad de ofrecer soluciones enfocadas a facilitar la vida diaria de los seres humanos mediante una interacción automatizada y natural con el medio ambiente en el cual se desenvuelven. Uno de los temas analizados en esta área es el reconocimiento y clasificación de actividades humanas, el cual sirve como estudio base para numerosas aplicaciones planteadas y soluciones desarrolladas.

Un aspecto crucial a tomar en cuenta para el reconocimiento y clasificación de actividades humanas es la fuente de datos a utilizar, ya que de ello depende la elección de las actividades con las cuales se trabaja, además de las técnicas utilizadas para el estudio. La tendencia en los años recientes, para la adquisición de datos, ha sido la utilización de sensores incrustados en objetos de uso cotidiano [6,8]. Otros trabajos han explorado el uso de sensores presentes en los dispositivos móviles, principalmente el acelerómetro, como los presentados por Brayat et al. [3] y Casale et al. [5]. Buscando una mayor precisión y efectividad en el reconocimiento de actividades, se han realizado también trabajos que utilizan el sonido como fuente de datos [7,9,12], principalmente haciendo uso del micrófono con el que cuentan los dispositivos móviles.

Algunos de los trabajos en el área del reconocimiento y clasificación de actividades humanas se han enfocado en el monitoreo de grupos de personas con necesidades específicas, como lo son los adultos mayores y los niños, como los presentados por Shaikh et al. [11] y Boughorbel et al. [4], principalmente como apoyo para salvaguardar su integridad física. Cuando el estudio se centra en el reconocimiento de actividades infantiles, esta tarea se vuelve un enorme reto por la gran diversidad de actividades que realizan, dependiendo del rango de edad en la que se encuentran, debido a su desarrollo fisico e intelectual natural. Dichas actividades van desde caminar o correr hasta actividades más complejas como la manipulación de juguetes o artículos potencialmente peligrosos.

Actualmente, la mayoría de los trabajos sobre reconocimiento de actividades infantiles hacen uso de acelerómetros, como el presentado por Nam et al. [10], para la identificación de actividades simples como camimar, correr o saltar. Dichos sensores utilizados deben ser colocados en las prendas que utilizan los niños cotidianamente, o en prendas especiales diseñadas específicamente para la experimentación.

Este trabajo se centra en el reconocimiento de actividades infantiles en niños de entre 12 y 36 meses, debido al comportamiento en común que comparten y que hace posible la identificación de actividades específicas para su reconocimiento y clasificación. La fuente de datos utilizada es el sonido ambiental, de manera que la adquisición de datos no interfiera con el comportamiento natural de los niños. Las actividades a tomar en cuenta para su análisis en el presente trabajo son: caminar, correr, llorar y manipular juguetes (bloques de plástico).

A continuación se describe cómo está organizado el artículo. Después de esta breve introducción, se describe la propuesta de metodología y materiales utilizados. En la tercera sección describimos los procesos de experimentación aplicados.

Seguido de la publicación de los resultados obtenidos durante el proceso. Por último, presentamos las conclusiones, y el trabajo a realizar en el futuro.

2. Materiales y métodos

En la presente sección se describen las características de los audios que modelan las actividades infantiles, así como los métodos utilizados para llevar acabo el presente estudio.

2.1. Descripción del conjunto de datos

Para este trabajo sobre el reconocimiento de actividades infantiles utilizando sonido, el conjunto de datos (Data-Set) se encuentra formado por grabaciones de audio de cuatro actividades que realizan niños en el rango de edad especificado (12 a 36 meses), en un ambiente controlado (dentro de una habitación): caminar, correr, llorar y manipular juguetes (bloques de plástico). La tabla 1 muestra la descripción de cada una de las actividades a tomar en cuenta para el análisis.

Tabla 1. Descripción general de las actividades.

Actividad	Descripción
Caminar	Trasladarse de un lugar a otro a velocidad media
Correr	Trasladarse rápidamente de un lugar a otro
Llorar	Emitir sonido de llanto como reacción a algún suceso
Manipular juguetes	Jugar con bloques de plástico de modo que éstos emitan ruido al chocar

Dispositivos de grabación. Para el proceso de adquisición de los datos, los audios de las actividades fueron grabados utilizando los micrófonos presentes en dos dispositivos móviles. Las características principales de los dispositivos utilizados se muestran en la tabla 2.

Tabla 2. Dispositivos de grabación.

Dispositivo	Procesador	Sistema Operativo
Lanix Ilium s620	MediaTek MT6582 quad-core	Android 4.2.2
Motorola Moto G4	Snapdragon 617	Android 6.0.1

Metadatos. Los audios grabados se encuentran en formato WAV (WaveForm Audio File), con la finalidad de no experimentar la pérdida de calidad que

Tabla 3. Metadatos.

Actividad	Tasa de Muestreo	Canales	Resolución
Caminar	44,100 Hz	Stereo	16 bit
Correr	44,100 Hz	Stereo	16 bit
Llorar	44,100 Hz	Stereo	16 bit
Jugar	44,100 Hz	Stereo	16 bit

otros formatos presentan debido al proceso de compresión. Las características generales como tasa de muestreo, canales y resolución, se muestran en la tabla 3.

Las características presentadas anteriormente aseguran una calidad aceptable para los archivos de audio grabados, además de definir los parámetros necesarios para grabaciones futuras con la finalidad de expander el Data-Set.

2.2. Espectro de Fourier

En el área de procesamiento de señales, la Transformada de Fourier es ampliamente utilizada debido a que, al expresar las señales en términos de las frecuencias que la componen, facilita su análisis y procesamiento para distintas aplicaciones.

En el presente trabajo se utiliza el espectro de Fourier de las muestras de audio para realizar el respectivo procesamiento dentro del modelo de clasificación utilizado.

2.3. Modelo de clasificación

Un aspecto importante a tomar en cuenta en el proceso de reconocimiento de actividades humanas es el modelo de clasificación utilizado, debido a que esto se reflejará en el comportamiento de los experimentos realizados y la precisión de los resultados obtenidos. Cuando se analizan señales de audio, como es el caso del presente trabajo, la utilización de un modelo adecuado para la clasificación de las muestras ayuda a obtener un procesamiento adecuado y confiable.

El modelo de clasificación elegido para este trabajo es un modelo de agrupamiento jerárquico aglomerativo, la elección se hizo tomando en cuenta la forma en que este método plantea la formación de grupos con individuos (muestras) de características similares. El modelo toma inicialmente tantas muestra se tengan, tratándolas como grupos individuales. El procesamiento consiste en ir formado grupos de forma ascendente, basándose en la similitud de sus individuos, hasta llegar a la cantidad de grupos definida para el estudio.

3. Experimentación

Mediante el proceso de captura de audios para las actividades a analizar en el presente trabajo, se obtuvieron un total de 23 grabaciones, en la tabla 4 se

muestran las cantidades de archivos de audio que se tienen para cada actividad. El procedimiento realizado a continuación fue llevado acabo en el lenguaje de programación R, debido a que éste está enfocado al análisis de datos y modelado estadístico, con poderosas características que lo hacen uno de los lenguajes más utilizados en investigación.

Tabla 4. Archivos de audio por actividad.

Actividad	Grabaciones
Caminar	5
Correr	4
Lorar	7
Jugar	7

Para el análisis de los archivos, se trabajó con el espectro de frecuencias de las señales de audio, obtenido mediante el análisis de Fourier correspondiente. Ya que las actividades pueden variar en cuanto a su longitud temporal, debido a su naturaleza y a los infantes, los archivos grabados para el presente trabajo tienen una duración de entre 10 y 160 segundos. Para poder realizar el análisis de frecuencias mencionado, el primer paso fue dividir todos los archivos de audio en muestras de 10 segundos, con la finalidad de obtener un espectro de Fourier utilizando la transformada rápida de fourier (FFT por sus siglas en inglés) [2] de igual longitud para todas las muestras y proceder al análisis respectivo. La tabla 5 muestra la cantidad total de muestras para cada actividad.

Tabla 5. Muestras por actividad.

Actividad	Muestras de 10 segundos
Caminar	24
Correr	6
Lorar	10
Jugar	42

Los datos de entrada al modelo de clasificación son cada uno de los espectros de frecuencias de las muestras de 10 segundos obtenidas a partir de los archivos de audio grabados de las actividades. El primer paso a realizar en el proceso de clasificación es formar la matriz de distancias con dichas muestras, esta matriz relaciona a todos los sujetos involucrados mediante una medida de su similitud. La obtención de la matriz de distancias se realizó con la función dist() contenida en el paquete stats de R Project [1].

La formación de grupos con individuos similares, de manera ascendente, se realiza en base a los datos contenidos en la matriz de distancias. Para este

proceso se sigue la estrategia de la distancia mínima o similitud máxima, en la cual las muestras con mayor similitud se juntan para formar grupos cada vez más grandes. El proceso finaliza cuando se ha llegado a obtener un solo grupo que contiene a todos los individuos del experimento o una vez que se ha alcanzado el número de grupos definido para el estudio.

La forma de procesar los datos de la matriz de distancias, formando grupos de manera jerárquica, es representada mediante un dendrograma, en el cual se puede visualizar el proceso completo y los grupos formados en cada paso. Este proceso se realizó utilizando la función hclust() contenida también en el paquete stats [1].

4. Resultados

En la figura 1 se observa el espectro de Fourier para una de las muestras de 10 segundos obtenidas de los archivos de audio. Este proceso fue realizado para cada una de las muestras totales. El espectro de Fourier permite visualizar la intensidad de cada frecuencia presente en la señal y, en base a esta característica, poder detectar el tipo de actividad al que pertenece la muestra de audio. Es importante mencionar que trabajar con el espectro de las señales de audio es un buen primer enfoque con el objetivo de reconocer actividades infantiles, sin embargo, sin un análisis más profundo, los experimentos realizados podrían parecer tender al caos, debido a la naturaleza de las grabaciones y varios factores más que podrían afectar, como el ruido de fondo.

Los espectros de Fourier de cada una de las muestras fueron tomados como datos de entrada para la elaboración de la matriz de distancia. Mediante esta matriz es posible conocer la medida de similitud entre las muestras de audio analizadas, basándose en la distancia euclidiana. En la tabla 6 se muestra una parte de la matriz de distancia resultante. Los nombres de las muestras están formados por la actividad a la que pertenece el sonido, el número de archivo y el número de muestra.

Tabla 6. Matriz de distancia para 5 muestras.

	Cry_1.wav_1	Cry_2.wav_1	Cry_2.wav_2	Cry_2.wav_3	Cry_3.wav_1	Cry_3.wav_2
Cry_2.wav_1	701.8916					
Cry_2.wav_2	683.3781	582.4608				
Cry_2.wav_3	681.9768	634.6091	581.2837			
Cry_3.wav_1	693.5581	748.2898	684.6646	637.5035		
Cry_3.wav_2	719.3180	746.5135	712.3953	670.5225	5733450	

Tomando la matriz de distancia obtenida en el paso anterior, se procesó utilizando el modelo de clasificación jerárquico utilizando el método de la distancia mínima entre muestras para el agrupamiento. La figura 2 muestra el dendrograma resultante.

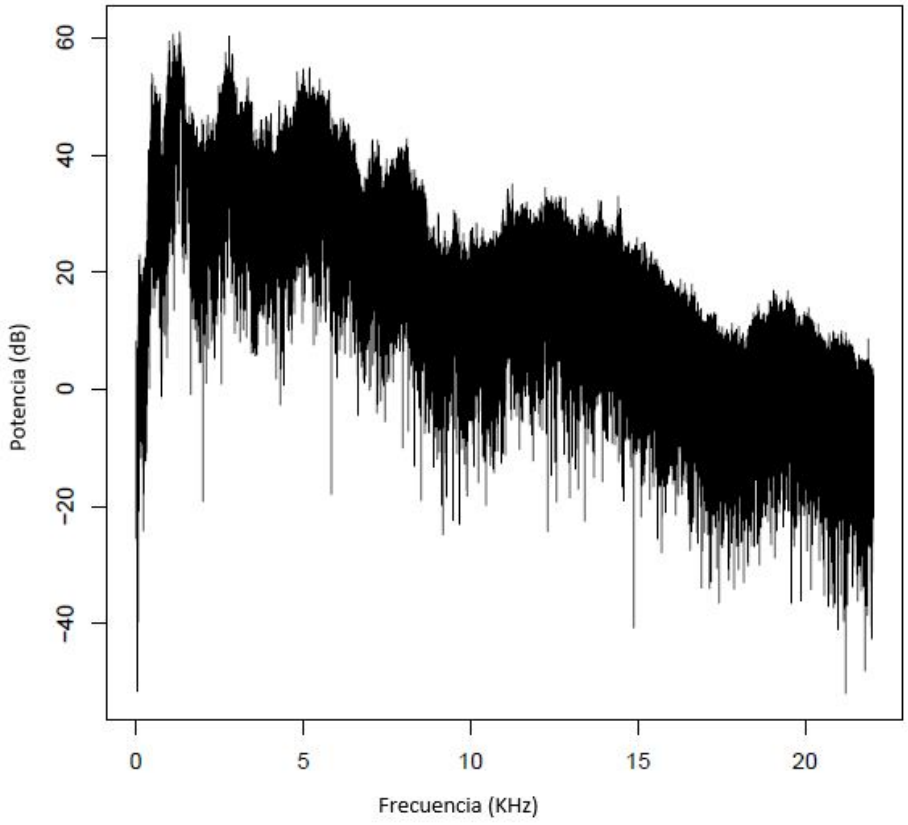


Fig. 1. Espectro de Fourier para una muestra de 10 segundos de duración.

5. Discusión y conclusiones

El enfoque principal de este trabajo de investigación es realizar un estudio preeliminar para poder estimar las actividades realizadas por infantes de 12 a 36 meses (considerados biológicamente como bebés) utilizando la evolución espectral de archivos de audio con duración de 10 segundos. Una técnica de agrupamiento jerárquico es utilizada para dicho reconocimiento; de esta experimentación se puede discutir y concluir lo siguiente:

- El reconocimiento de actividades infantiles pueden ser clasificadas utilizando la información del espectro de la señal de audio: La información contenida dentro del espectro permite obtener un agrupamiento de las actividades, sin embargo, se requieren 10 ramas para obtener un agrupamiento correcto de las 4 actividades.
- Actividades similares, como lo son correr y caminar, pueden ser correctamente clasificadas: A pesar de que en su evolución temporal es similar

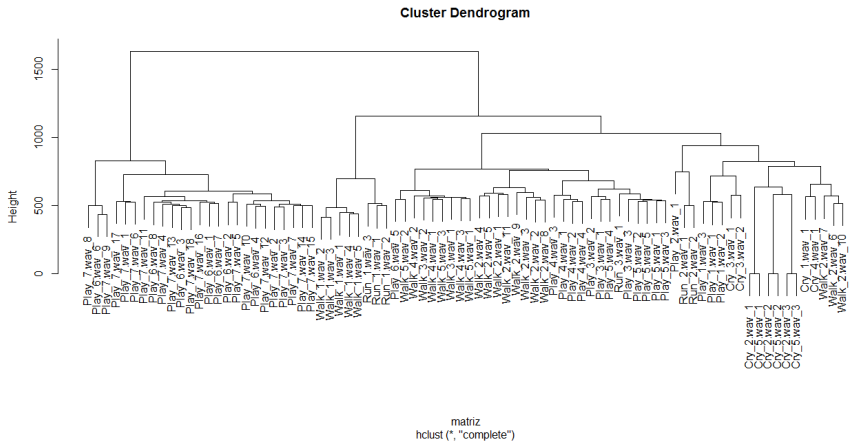


Fig. 2. Dendrograma resultante.

tanto visualmente como de manera auditiva, el espectro contiene suficiente información para diferenciar actividades que generan un sonido ambiental similar.

Por otro lado, se encontró indicios de un comportamiento que demuestra la necesidad de un data set más grande, sin embargo, una función de agrupamiento jerarquico tiene la capacidad de clasificar este tipo de señales correctamente, permitiendo así obtener un modelo de estimación de actividades infantiles utilizando solamente la información en sonidos ambientales.

6. Trabajo futuro

Como parte del trabajo futuro, se propone añadir más actividades que son comunmente desarrolladas por infantes de 12 a 36 meses, así como un estudio mas profundo de características de la evolución espectral y temporal de la señal para clasificar dichas actividades. de manera adicional se propone los siguientes puntos específicos:

- Realizar un estudio de curvas de aprendizaje para estimar la necesidad de extender el conjunto de datos por cada actividad.
- Agregar más actividades realizadas por los infantes.
- Utilizar el enfoque de extracción de características para reducir la cantidad de datos necesarios así como optimizar la clasificación.
- Implementar técnicas complejas de aprendizaje por computadora.

Referencias

1. R: Distance matrix computation, `https://stat.ethz.ch/R-manual/R-devel/library/stats/html/dist.html`
2. R: Fast discrete fourier transform (fft), `https://stat.ethz.ch/R-manual/R-devel/library/stats/html/fft.html`
3. Bayat, A., Pomplun, M., Tran, D.A.: A study on human activity recognition using accelerometer data from smartphones. Procedia Computer Science 34, 450–457 (2014)
4. Boughorbel, S., Breebaart, J., Bruekers, F., Flinsenberg, I., Kate, W.T.: Child-activity recognition from multi-sensor data. Proceedings of the 7th International Conference on Methods and Techniques in Behavioral Research - MB '10 (2010)
5. Casale, P., Pujol, O., Radeva, P.: Human activity recognition from accelerometer data using a wearable device. Pattern Recognition and Image Analysis Lecture Notes in Computer Science p. 289–296 (2011)
6. Cornacchia, M., Ozcan, K., Zheng, Y., Velipasalar, S.: A survey on activity detection and classification using wearable sensors. IEEE Sensors Journal 17(2), 386–403 (2017)
7. Galván-Tejada, C.E., Galván-Tejada, J.I., Celaya-Padilla, J.M., Delgado-Contreras, J.R., Magallanes-Quintanar, R., Martinez-Fierro, M.L., Garza-Veloz, I., López-Hernández, Y., Gamboa-Rosales, H.: An analysis of audio features to develop a human activity recognition model using genetic algorithms, random forests, and neural networks. Mobile Information Systems 2016, 1–10 (2016)
8. Lara, O.D., Labrador, M.A.: A survey on human activity recognition using wearable sensors. IEEE Communications Surveys & Tutorials 15(3), 1192–1209 (2013)
9. Lee, M., Yang, H., Han, D., Yu, C.: Crowdsourced radiomap for room-level place recognition in urban environment. 2010 8th IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops) (2010)
10. Nam, Y., Park, J.W.: Child activity recognition based on cooperative fusion model of a triaxial accelerometer and a barometric pressure sensor. IEEE Journal of Biomedical and Health Informatics 17(2), 420–426 (2013)
11. Shaikh, M.A.M., Hirose, K., Ishizuk, M.: Recognition of real-world activities from environmental sound cues to create life-log. The Systemic Dimension of Globalization (Jan 2011)
12. Sim, J.M., Lee, Y., Kwon, O.: Acoustic sensor based recognition of human activity in everyday life for smart home services. International Journal of Distributed Sensor Networks 11(9), 679123 (2015)

ISSN 1870-4069

Pronóstico a corto plazo de velocidad del viento a partir de datos incompletos

Hector Rodriguez Rangel[1], Noel A. Garcia Carrillo[1], Juan J. Flores[2], Luis A. Morales Rosales[3], Giovanni Manjarrez Montelongo[1]

[1] Instituto Tecnológico de Culiacán, División de Estudios de Posgrado e Investigación, Culiacan, México

[2] División de Estudios de Postgrado, Facultad de Ingeniería Eléctrica, Universidad Michoacana de San Nicolás de Hidalgo, Morelia, Michoacán, México

[3] Facultad de Ingeniería Civil, CONACYT-Universidad Michoacana de San Nicolás de Hidalgo, Morelia, Michoacán, México

hrodriguez@itculiacan.edu.mx, noelandresg@gmail.com, juanf@umich.mx, amorales@conacyt.mx, gmanjarrez@itculiacan.edu.mx

Resumen. En este artículo se presenta un estudio realizado a la predicción a corto plazo de la velocidad del viento en series de tiempo incompletas. Se ha propuesto realizar este estudio dado el incremento en el interés hacia la transición global de la producción de energías limpias. Siendo el pronóstico de dicha variable de suma importancia para las etapas de planeación, administración y producción. El proceso de pronóstico implica el uso de observaciones pasadas de la variable a pronosticar (velocidad del viento). Para medir la velocidad del viento, las estaciones meteorológicas utilizan dispositivos llamados anemómetros, pero debido a un mal mantenimiento, errores de conexión o desgaste natural, pueden presentar datos falsos o faltantes. En este trabajo, se explora la reconstruión la serie de tiempo mediante Redes Neuronales Artificiales, para posteriormente realizar el pronóstico a corto plazo de la serie de tiempo de la velocidad del viento utilizando el método de k Vecinos más Cercanos. Para probar la metodología propuesta, se utilizaron datos recolectados de diferentes locaciones del estado de Michoacán.

Palabras clave: pronóstico de la velocidad del viento, recostrucción de series de tiempo, RNA, kNN.

Wind Speed Short-term Forecasting from Incomplete Data

Abstract. This paper presents a wind speed short-term forecasting from incomplete time series. This study has been proposed given the interest increasing in a global transition of clean energy production. Being

the forecast of interest variable an important process for the planning, administration and production stages. The forecasting process involves the use of historical data of the variable to be forecast (wind speed). Measuring the wind speed, weather stations use devices called anemometers, but due poor maintenance, connection errors or natural wear, they may present false or missing data. In this work, the reconstruction of the time series using Artificial Neural Networks is explored, and then the short-term forecast of the time series of the wind speed using the method of k Nearest Neighbors is presented. To test the proposed methodology, we used data collected from different locations in the state of Michoacan.

Keywords: wind speed forecast, time series reconstruction, ANN, kNN.

1. Introducción

La energía desempeña un papel de suma importancia en las actividades humanas. Así, la producción de energía se ha convertido en una de las principales cuestiones económicas y medioambientales en todo el mundo [15]. Además, la demanda de energía ha aumentado constantemente con el tiempo, causando daños al medio ambiente al producirla. Según el censo realizado por Observ'ER & Foundation Énergies pour le Monde en 2012, la producción de energía dominante es la producida a través de combustibles fósiles, la cual abarca más de dos tercios de la producción total de energía (68,1 %) [15]. Este tipo de energía, además de ser no renovable, genera daños ambientales irreversibles a nuestro planeta, debido a las grandes cantidades de dióxido de carbono (CO2) emitidas a la atmósfera, lo que aumenta los gases de efecto invernadero.

Un mecanismo que busca contrarrestar los daños ambientales y reducir los costos de producción de energía es el uso de energía renovable, también llamada energía limpia. Se encuentra una amplia gama de opciones dentro de las energías renovables como: solar, eólica, biomasa, geotérmica, residuos no renovables e hidroeléctrica, entre otras. De lo anterior, la energía solar y eólica han presentado un crecimiento anual global entre 2002 y 2012 de 50.6 % y 26.1 %, respectivamente [15].

Teniendo en cuenta el gran y continuo crecimiento de estas energías renovables, se han encontrado muchos desafíos (satisfacer la demanda, reducir los costes de producción, mejorar la planificación de las plantas energéticas, etc.). Para llegar a una plena adopción de este tipo de energía es necesario encontrar soluciones eficaces a estos desafíos. Una de las necesidades principales es hacer predicciones de las variables involucradas en la producción de energía. Es decir, tener una estimación de la cantidad de energía que se producirá en los próximos minutos, horas, días o meses [18]. Algunas de las variables involucradas en la producción de energía limpia podrían ser las siguientes: velocidad del viento, temperatura, humedad, presión atmosferica, niveles de presas, causales de rios, etc. En este trabajo, se trata específicamente en el pronóstico de velocidad del viento. La predicción de estas variables puede realizarse a corto, mediano y largo plazo. No hay un marco de tiempo definido para cada uno de los períodos [10].

Sin embargo, regularmente se utilizan previsiones a mediano y largo plazo para la planificación táctica y estratégica, respectivamente. La predicción a corto plazo se utiliza a nivel operativo y su periodicidad puede ser en escala de tiempo de minutos, horas, o días [10,5].

La capacidad de predecir la velocidad del viento es esencial para la integración correcta de la energía eólica en los sistemas eléctricos. Según Barber et al. [2], la importancia de los pronósticos de viento para la industria de energía eólica se deriva de tres hechos: 1) La potencia agregada producida y consumida en todo el sistema eléctrico debe estar casi en equilibrio en todo momento para garantizar la fiabilidad y la seguridad del suministro. 2) La potencia de un parque eólico es muy variable ya que depende fuertemente de la velocidad y dirección del viento. y 3) No existe un mecanismo eficiente y rentable de almacenamiento de energía eléctrica.

Recientemente, una estimación para el año 2030 menciona que un pronóstico perfecto será valorado en 3 *billones* de dólares anuales [16], para el sistema de energía de Estados Unidos. La estimación fue realizada por el Departamento de Energía [2]. Por lo tanto, la previsión de la velocidad del viento sigue desempeñando un papel importante en la tarea de suministro de energía [5].

En la literatura se encontró que han habido diferentes enfoques utilizados para esta tarea [4,13,9,1]. Chang presenta [5] un estudio que categoriza diferentes enfoques que se ocupan del problema de pronóstico que establece la Persistencia, Física, Estadística, Correlaciones Espaciales, Inteligencia Artificial y Métodos Híbridos. El método Persistence (o Naïve) es un método de pronóstico básico, donde $Y_{t+\Delta} = Y_t$ [20]. Este método particular se utiliza a menudo como un método basal. Los métodos físicos son predicciones numéricas del tiempo desarrolladas por los meteorólogos para una predicción del tiempo a gran escala [12]. Los métodos estadísticos (ME) utilizan datos históricos para encontrar relaciones dentro de la serie de tiempo de velocidad del viento. Dentro de los métodos estadísticos, tenemos el Auto Regresivo (AR), Auto Regresivo de Media Móvil (ARMA), y el Auto Regresivo Integrado de Media Móvil (ARIMA), entre otros [8,7,6]. Los métodos de Inteligencia Artificial (MIA) utilizan Redes Neuronales Artificiales (RNA) [7,3], Máquinas de Soporte Vectorial (MSV) [21], k-Vecinos Más Cercanos (kNN) [19], entre otros como métodos de predicción. Los MIA mejoran en algunos casos los resultados obtenidos utilizando métodos estadísticos. Los métodos híbridos son una combinación de métodos de predicción [7] (por ejemplo, la combinación de métodos estadísticos y de inteligencia artificial).

Los ME o MIA usualmente realizan el pronóstico usando datos históricos. Pero, si el conjunto de datos presenta huecos en la representación de sus observaciones, el ME puede tener problemas para realizar un pronóstico adecuado. En realidad, las series de tiempo de viento presentan datos faltantes debido a problemas con los sensores (por ejemplo, problemas de comunicación, desgaste natural, falta de mantenimiento, etc.). La velocidad del viento es irregular e intermitente [11], lo que sumado al hecho de que las series temporales se encuentran incompletas, la tarea de pronóstico a corto plazo es aún más difícil.

En este artículo, enfrenta del problema de la reconstrucción de la series de

tiempo de la velocidad del viento. Se propone una metodología que se ocupa de este problema, la cual puede trabajar fuera de línea y en tiempo real. La propuesta presentada comienza caracterizando el comportamiento temporal de la serie de tiempo, creando una base de datos de patrones. A continuación, se modela la serie de tiempo utilizando un método híbrido. Este método utiliza una RNA como un enfoque de predicción y Algoritmos Genéticos (AG) para definir la mejor topología de la RNA. Una vez entrenada a la RNA, se procede a reconstruir la serie de tiempo de velocidad del viento. Con estas series de tiempo reconstruídas se procede a realizar pronóstico a corto plazo de la velocidad del viento utilizando el algoritmo de k-Vecinos Más Cercanos (kNN). El resto del trabajo está organizado de la siguiente manera: La Sección 2 describe la metodología propuesta para la realización de la predicción a corto plazo, en donde uno de los pasos previos al pronóstico es el de reconstruir una serie temporal de velocidad del viento, la Sección 3 presenta los experimentos realizados con nuestra propuesta y finalmente las conclusiones de este trabajo se presentan en la Sección 4.

2. Predicción de la velocidad del viento

El problema del pronóstico de series de tiempo de la velocidad del viento no es algo nuevo. Es una tarea complicada ya que hay una amplia gama de patrones de comportamiento diferentes en los datos a modelar. Cada serie de tiempo cumple con distintas características y al modelar series de tiempo de viento se observa distintos comportamientos, por lo que es necesario modelar cada una de ellas. Con anterioridad, distintos métodos de pronóstico se han implementado con el objetivo de realizar pronósticos [5]. De estos métodos sobresalen los que emplean técnicas de inteligencia artificial como las RNA [7,3], dado que en muchos de los casos este método de pronóstico proporciona mejores resultados que los obtenidos mediante los ME.

Aunado a que el problema de realizar pronósticos de series de tiempo es naturalmente complicado; si añade el problema de que la serie temporal podría encontrarse incompleta, el problema incrementa su complejidad, aunque, hoy en día hay una serie de algoritmos y técnicas que ofrece una pronóstico eficaz [5,20]. Sin embargo, en la mayor parte de la literatura, el problema de la predicción de la velocidad del viento se aborda suponiendo que la serie temporal está completa (sin ningún hueco). Sin embargo, este no es el caso en realidad. Regularmente, hay problemas con sensores que dan como resultado datos faltantes dentro de la serie de tiempo de velocidad del viento. Para hacer frente a este problema utilizando métodos estadísticos como ARIMA o Holt-Winters se convierte en un trabajo no muy sencillo de realizar, o incluso imposible. Por lo tanto, en este artículo trata el problema de la reconstrucción de las series de tiempo de velocidad del viento como una subtarea previa a la realización del pronóstico a corto plazo y realiza una comparación entre los pronósticos obtenidos con series de tiempo reconstruídas y los obtenidos con las series de tiempo sin reconstruír. El proceso completo de pronóstico se observar en la Figura 1.

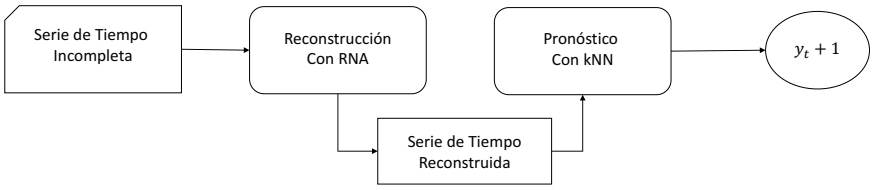


Fig. 1. Proceso de reconstrucción y pronóstico de la serie de tiempo de la velocidad del viento.

2.1. Reconstrucción de series de tiempo

Como ha sido mencionado con anterioridad, nos hemos visto en la necesidad de reconstruir las observaciones faltantes en la serie de tiempo. Para lograr esto es necesario realizar una serie de pasos los cuales se pueden observar en la Figura 2. Una vez completada la reconstrucción de datos se obtiene una nueva serie de tiempo; esta nueva serie de tiempo reconstruída es utilizada para la realización del pronóstico. Un ejemplo de reconstrucción de datos se puede observar en la figura 3.

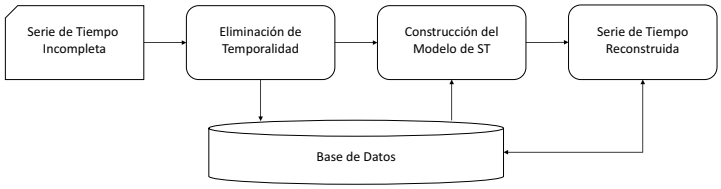


Fig. 2. Diagrama de flujo del proceso de reconstrucción de la serie de tiempo de la velocidad del viento

Para hacerle frente a los huecos dentro de la serie de tiempo, se propone la remoción de la temporalidad. En lugar de tener una ventana deslizante de tamaño m, el proceso de eliminación de temporalidad crea una base de datos en la que se registran m características de las series temporales, asociadas con su respectivo valor de pronóstico. Con la base de datos creada, se comienza a construir el modelo de serie de tiempo (CTME). Este proceso utiliza una RNA para modelar la serie de tiempo y Algoritmos Genéticos para definir la arquitectura de la RNA [8,7,6]. Una vez definida la estructura óptima y realizado el proceso de entrenamiento de RNA, se comienza el proceso de reconstrucción

de datos. La Figura 2 muestra el diagrama de flujo del proceso de reconstrucción de la serie temporal de velocidad del viento.

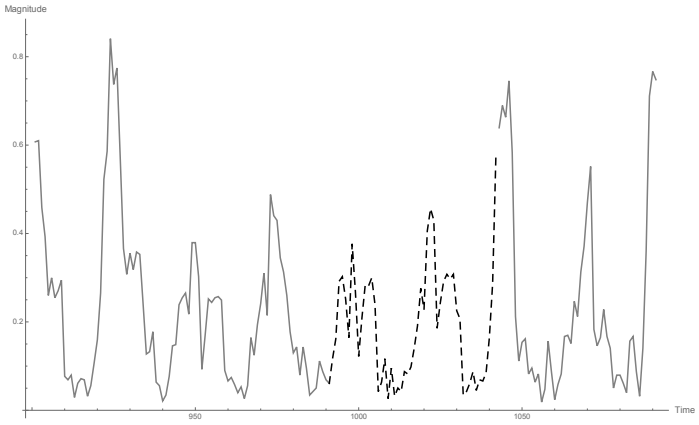


Fig. 3. Ejemplo de reconstrucción de datos en serie de tiempo.

2.1.1 Proceso de eliminación de la temporalidad

Para tratar el problema de los huecos producidos por datos faltantes en las series de tiempo, se propone la creación de una base de datos. Regularmente, para el problema de pronóstico, hay una ventana de tamaño m que se desliza a través de la serie de tiempo. Pero debido a los agujeros dentro de la serie de tiempo, esta ventana ya no puede ser utilizada de la misma manera. Es por ello que se propone crear una base de datos con los patrones temporales dentro de la serie de tiempo de velocidad del viento. La base de datos registra las características de m (lecturas anteriores de la serie de tiempo) y las asocia con sus respectivos valores de pronóstico. Después de crear la base de datos, se eliminan todos los registros donde se encuentra ausencia de datos. Al convertir la serie de tiempo a una base de datos, se puede dejar de lado la dimensión de tiempo.

La Tabla 1 muestra el resultado de crear la base de datos y extraer las w características asociadas con una salida esperada. Cada registro dentro de la base de datos representa las últimas m observaciones en un tiempo definido (t), asociado a la salida esperada y_{t+1} (para propósitos de entrenamiento). De la Tabla 1 se observa que los registros $m+2$, $m+3$, $m+4$, $m+5$, $m+6$, $m+7$ presentan un hueco (ϕ) en su interior. Estos registros se eliminan. Con la base de datos definida (sin los huecos), se puede comenzar a buscar una RNA que modela el comportamiento de la serie del tiempo de la velocidad del viento.

2.1.2 Construcción del modelo de series de tiempo

Tabla 1. Transformación de la serie temporal en un problema de clasificación.

	entradas (W_t)					$\hat{y}_{t+1} = P(W_t)$
m	y_1,	y_2,	a_3,	...	y_m	y_{m+1}
$m+1$	y_2,	y_3,	y_4,	...	y_{m+1}	y_{m+2}
$m+2$	y_3,	y_4,	y_5,	...	y_{m+2}	ϕ
$m+3$	y_4,	y_5,	y_6,	...	ϕ	y_{m+4}
$m+4$	y_5,	y_6,	y_7,	...	y_{m+4}	y_{m+5}
$m+5$	y_6,	y_7,	ϕ,	...	y_{m+5}	y_{m+6}
$m+6$	y_7,	ϕ,	y_9,	...	y_{m+6}	y_{m+7}
$m+7$	ϕ,	y_9,	y_{10},	...	y_{m+7}	y_{m+8}
$m+8$	y_9,	y_{10},	y_{11},	...	y_{m+8}	y_{m+9}
$m+9$	y_{10},	y_{11},	y_{12},	...	y_{m+9},	y_{m+10},
		...			...	
$n-1$	y_{n-m},	y_{n-m+1},	y_{n-m+2},	...	y_{n-1}	y_n

En trabajos recientes, se ha demostrado que las RNA presentan grandes cualidades de clasificación [5]. Una vez definida la base de datos, se aborda el problema de regresión que permite crear un modelo capaz de representar el comportamiento de la serie de tiempo de la velocidad del viento.

En este artículo, la arquitectura propuesta de la RNA es un Feedforward Multilayer Perceptron, entrenado por métodos basados en gradiente. La Figura 4 muestra la arquitectura de la RNA propuesta en este trabajo. Una RNA, como un aproximador universal puede aprender cualquier función dada. Un conjunto de m observaciones pasadas se consideran como los datos de entrada, la capa oculta consta de h neuronas ocultas, la capa de salida corresponde al valor de pronóstico $\hat{y}_{t+1}$; se utiliza un sigmoide como función de activación. La RNA que modela la serie de tiempo de velocidad del viento $\hat{y}_{t+1}$ puede definirse como:

$$\begin{aligned} \hat{y}_{t+i} &= f_1(\sum_{l}^{w+1} w_l x_l), \\ x_l &= f_2(\sum_{j}^{m} w_{lj} y_{t-l_j}), \end{aligned} \tag{1}$$

donde f_1 y f_2 son las funciones de activación, y w son los coeficientes (también conocidos pesos de sinápticos).

Para modelar el comportamiento de la serie de tiempo de velocidad del viento, se necesita proporcionar un modelo preciso. Para definir la exactitud del modelo óptimo, se utilizo la medida estadística de la media del error al cuadrado (MSE), que se define como:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2. \tag{2}$$

Esta medida se minimizará durante el proceso de entrenamiento.

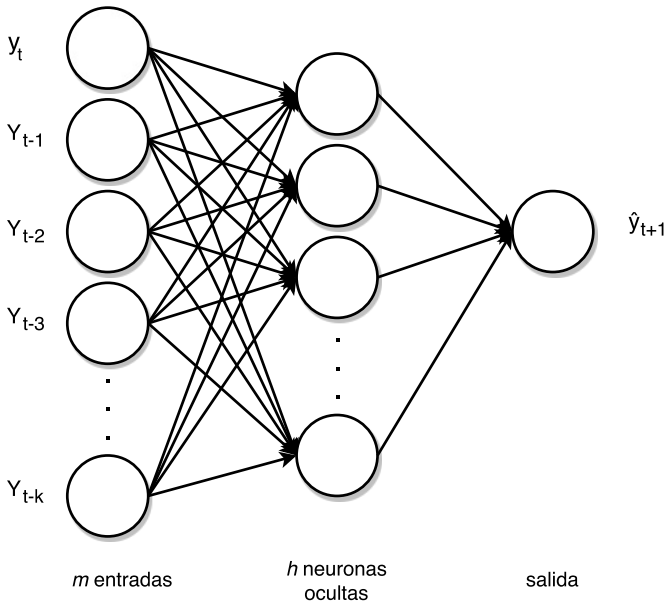


Fig. 4. Arquitectura de RNA con $m+1$ observaciones pasadas, h neuronas en su capa oculta y una única salida ($\hat{y}_{t+1}$).

Algoritmos genéticos (AG) son una técnica de optimización inspirada en el principio de la evolución de Darwin. Esto es, imita una versión simplista del proceso de evolución biológica, que consiste en crear una población de individuos, donde cada individuo representa una solución prospectiva del problema que se está resolviendo. AG modifica esta población utilizando operadores genéticos: selección, mutación y recombinación [17].

Determinar la mejor arquitectura de RNA y el algoritmo de entrenamiento es un problema de optimización. Esto es, definen la arquitectura de la RNA determinando el número de entradas y el número de neuronas en la capa oculta. Este es un paso crítico en el proceso de entrenamiento que generalmente se realiza por ensayo y error. En trabajos anteriores, el uso de AG ha sido probado dando excelentes resultados [8]. Cada individuo (el cromosoma) en el AG se define como un vector binario que está codificando el número de entradas y el número de neuronas ocultas de la RNA. El algoritmo empleado de AG para encontrar la mejor topología de RNA se describe paso a paso el Algoritmo 1.

El AG comienza generando una población inicial, a partir de esta población se evalúan todos sus individuos. El proceso continua aplicando selección cruza, mutación a los individuos de la población. Este procedimiento se repite hasta que se alcanza un criterio de convergencia. El pseudocódigo de este proceso lo encontramos en el Algoritmo 1. El algoritmo genético proporciona básicamente como resultado el número de entradas, neuronas en la capa oculta y algoritmo de entrenamiento. Después de definir la topología de la RNA, se comienza un

proceso de refinamiento. El refinamiento es el mismo proceso, simplemente se da la oportunidad de un entrenamiento más extenso con la topología de la red ya definida [7,3].

Algorithm 1 Encontrando la mejor topología para la red usando AG

```
1: procedure AG
2:     Generar una población inicial
3:     Computar la función de evaluación de cada individuo
4:     while NOT Terminado do
5:     //Producir nueva generación
6:     for Tamaño Población/2 do
7:         BEGIN //ciclo reproductivo
8:         Seleccionar dos individuos de la anterior generación, para el cruce
9:         (probabilidad de selección proporcional a la función de evaluación del
10:        individuo).
11:        Cruzar con cierta probabilidad los dos individuos obteniendo dos
12:        descendientes.
13:        Mutar los dos descendientes con cierta probabilidad.
14:        Computar la función de evaluación de los dos descendientes mutados.
15:        Insertar los dos descendientes mutados en la nueva generación.
16:        Seleccionar los np mejores individuos.
17:        END //ciclo reproductivo
18:    if La población ha convergido then Terminado := True
19:    return mejor individuo
```

2.1.3 Reconstrucción de datos

Para el proceso de reconstrucción, se utiliza la RNA que modela el comportamiento de la serie de tiempo de velocidad del viento. Usando este modelo, se puede comenzar con el proceso de reconstrucción. Este proceso comienza con un vector de observaciones ordenadas que se evalúan secuencialmente. Cuando se encuentra un hueco o dato faltante en la serie temporal, comienza el proceso de datos de reconstrucción. Este proceso utiliza las m últimas observaciones del vector utilizado como entrada. La salida de la RNA ya entrenada en los pasos anteriores se utiliza para llenar el vacío. El proceso continúa hasta que se alcanza el final del vector de observación. La Figura 5 muestra el diagrama de flujo del proceso de reconstrucción.

2.1.4 Pronóstico a corto plazo mediante Vecinos más Cercanos (kNN)

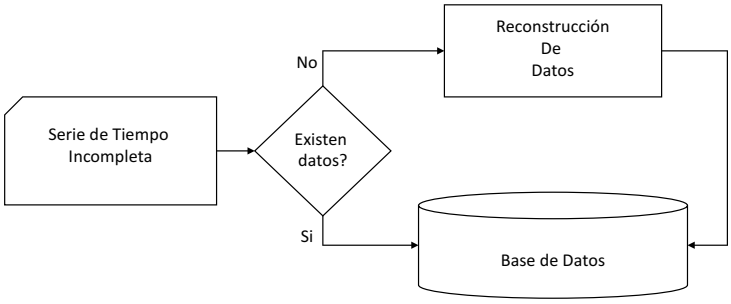


Fig. 5. Proceso de reconstrucción de datos faltantes

Para el pronóstico a corto plazo de la velocidad del viento, se utilizan las series reconstruidas de acuerdo a los pasos definidos previamente. Para el pronóstico se propone utilizar el algoritmo de k Vecinos más Cercanos (kNN). Este algoritmo es una técnica de clasificación y regresión basada en la similaridad de los individuos de una población. La premisa de esta técnica indica que los miembros de una población coexisten rodeados de individuos similares que tienen propiedades similares. Esta idea sencilla es la regla de aprendizaje del los clasificador y regresores kNN. Como se muestra en la Figura 6 donde se busca la similaridad de los individuos más cercanos.

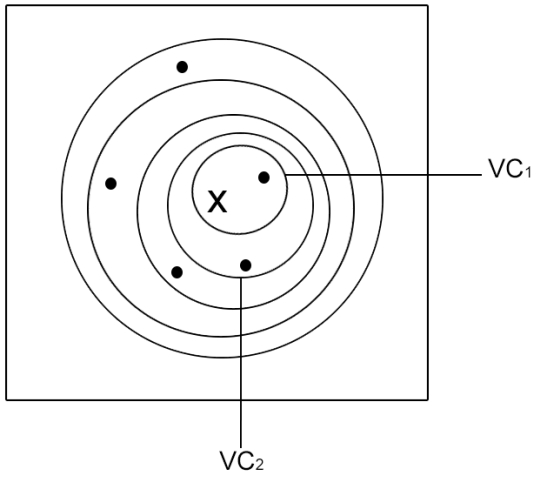


Fig. 6. Ejemplo de clasificación basada en la similaridad de los individuos kNN.

Un algoritmo kNN particular, se caracteriza por cuestiones tales como el número de vecinos, el tipo de distancia utilizada, etc. El método usado en este

documento, para realizar el pronostico de y_{t+1}, es necesario buscar la similitud del individuo x en la base de datos definida en la Tabla 1. Donde el individuo x es definido como las observaciones pasadas de la serie de tiempo en un tiempo t ($x = [y_{t-w}, y_{t-w+1}, ..., y_t]$). Después de evaluar las distancias de los individuos de la base de datos, se toman los k vecinos más cercanos y el pronostico resultante es el promedio de la columna y_{t+1} de la Tabla 1. La definición matemática de nuestro modelo de kNN se puede observar a continuación:

$$\hat{y}_{t+1} = \frac{1}{k} \sum_{i=1}^{k} P(VC(i, y_t)). \tag{3}$$

La distancia euclidiana de dos individuos se define por la Ec. (4):

$$d(r, s) = \sqrt{\sum_{j=1}^{m} (r_j - s_j)^2}. \tag{4}$$

La ventana que termina en el tiempo t, denotada por w_t es la secuencia o vector de las m mediciones hasta t como se indica en la Ec. (5):

$$w_t = [y_{t-m+1}, y_t] \in \mathbb{R}^m. \tag{5}$$

La definición recursiva del conjunto de k vecinos más cercanos es:

$$VC(k, x) = argmin(d(w, x)|w \in W - V_{k-1}(x)), \tag{6}$$

donde:

$$\begin{aligned} V_k(x) &= \bigcup_{j=0}^{k} VC_j(k, x), \\ V(0, x) &= \phi. \end{aligned} \tag{7}$$

Para la determinación de los números adecuados de k y w se realizó una exploración en un rango de valores lo suficientemente amplio para establecer las combinaciones que permitan llegar a un valor de error menor. En este caso se realizaron experimentos en donde se iteraron estos valores desde $k = 1$ hasta $k = 100$ y desde $w = 1$ hasta $w = 100$. Un ejemplo gráfico de esta exploración se observa en la Figura 7, en donde el eje vertical indica la magnitud de los errores.

3. Resultados

Para probar la aptitud de nuestra propuesta, se realizaron experimentos con series de tiempo de la velocidad del viento procedentes de anemómetros ubicados en seis diferentes lugares dentro del estado de Michoacán, México, con registros a intervalos horarios. Dado que los sensores son susceptibles a fallas de funciones y errores de lectura, se tiene el problema de la obtención de series de tiempo con datos faltantes.

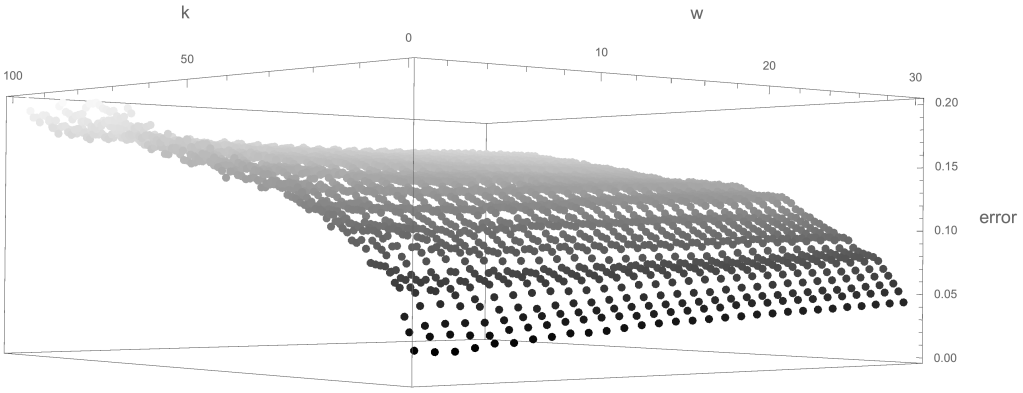


Fig. 7. Exploración de valores óptimos de k y w para el pronóstico mediante kNN.

Se realizaron varios experimentos aplicando el enfoque del método de reconstrucción de la serie de tiempo descrito en la Figura 2. Cada experimento comienza eliminando la temporalidad de la serie temporal creando una base de datos siguiendo el procedimiento descrito en la sección anterior. Una vez obtenida la base de datos, definimos una RNA que modela el comportamiento de la serie de tiempo de velocidad del viento.

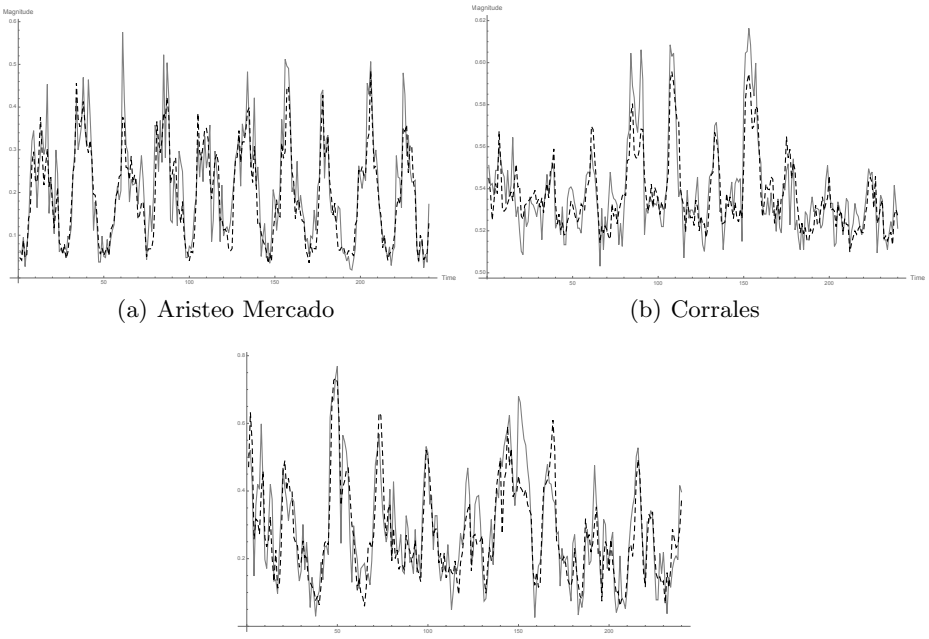
La topología obtenida en esta etapa se utiliza en un proceso de refinamiento, que sólo aumenta el número de épocas, en este caso, se usó hasta 3,000 épocas.

Una vez realizada la reconstrucción y obtenida la nueva serie de tiempo reconstruida; ésta es utilizada como entrada para el proceso de pronóstico mediante el algoritmo de kNN. Realizamos experimentos con seis series de tiempo de distintas estaciones de viento. Paralelamente se realizó el mismo proceso de pronóstico con las series de tiempo sin reconstruir. La aptitud de los pronósticos realizados se muestra en la Tabla 2.

Tabla 2. Comparativa de errores definidos por la medida estadística MSE obtenidos en los pronósticos utilizando RNA y kNN en series sin reconstruir y reconstruidas.

MSE en Pronóstico RNA vs kNN						
	Sin reconstruir			Reconstruida		
	RNA	kNN	Naïve	RNA	KNN	Naïve
1 La Palma	0.0078	**0.0028**	0.0067	0.0079	0.0034	0.0086
2 La Piedad	0.0076	**0.0047**	0.0073	0.0077	0.0052	0.0081
3 Markazuza	0.0010	0.0008	0.0011	0.0010	**0.0001**	0.0005
4 Aristeo Mercado	0.0071	0.0037	0.0084	0.0073	**0.0037**	0.0087
5 Corrales	0.0022	0.0008	0.0012	0.0037	**0.0001**	0.0006
6 Melchor Ocampo	0.0144	**0.0073**	0.0112	0.0137	0.0082	0.0115

Un ejemplo gráfico del proceso de reconstrucción se muestra en la Figura 3. En la Figura 3, la línea continua representa los datos reales, y los punteados representa la reconstrucción. Una vez realizado la reconstrución se procedio a realizar el pronostico mediante el uso de kNN. En la Figura 8 muestra un ejemplo de pronóstico de tres series de tiempo correspondientes a Aristeo Mercado, Corrales, y Melchor Ocampo. Para esta Figura 8, la linea continua representa los datos reales y la punteada representa al pronostico.



(a) Aristeo Mercado (b) Corrales

(c) Melchor Ocampo

Fig. 8. Resultado del pronóstico kNN de las estaciones Aristeo Mercado, Corrales, Melchor Ocampo, con serie reconstruida por RNA.

4. Conclusiones

Este artículo ha propuesto un enfoque de reconstrucción de series de tiempo basado en una metodología neural-evolutiva híbrida. Esta metodología permite determinar las arquitecturas óptimas de RNA para modelar series temporales incompletas y posteriormente para poder reconstruirlas. Los experimentos se realizaron considerando múltiples series temporales con datos faltantes. Se generaron modelos y reconstrucción de datos. Se realizó una comparación entre

los pronósticos obtenidos a partir de series de tiempo reconstruidas y series de tiempo con datos faltantes. Se determinó que en algunos casos la reconstrucción puede mejorar la eficiencia del pronóstico pero en otros caso no ocurría lo mismo. Se requiere de realizar un análisis estadístico para determinar la superioridad de un método sobre otro. Los experimentos se realizaron utilizando la plataforma Python utilizando la biblioteca Neurolab [14] para las implementaciones de RNA.

Referencias

1. Alexiadis, M., Dokopoulos, P., Sahsamanoglou, H., Manousaridis, I.: Short-term forecasting of wind speed and related electrical power. Solar Energy 63(1), 61–68 (1998)
2. Barber, C., Bockhorst, J., Roebber, P.: Auto-regressive hmm inference with incomplete data for short-horizon wind forecasting. In: Advances in Neural Information Processing Systems. pp. 136–144 (2010)
3. Barbounis, T.G., Theocharis, J.B., Alexiadis, M.C., Dokopoulos, P.S.: Long-term wind speed and power forecasting using local recurrent neural network models. IEEE Transactions on Energy Conversion 21(1), 273–284 (2006)
4. Cassola, F., Burlando, M.: Wind speed and wind energy forecast through kalman filtering of numerical weather prediction model output. Applied energy 99, 154–166 (2012)
5. Chang, W.Y.: A literature review of wind forecasting methods. Journal of Power and Energy Engineering 2(04), 161 (2014)
6. Erdem, E., Shi, J.: Arma based approaches for forecasting the tuple of wind speed and direction. Applied Energy 88(4), 1405–1414 (2011)
7. Flores, J.J., Graff, M., Rodriguez, H.: Evolutive design of arma and ann models for time series forecasting. Renewable Energy 44, 225–230 (2012)
8. Flores, J.J., Loaeza, R., Rodríguez, H., Cadenas, E.: Wind speed forecasting using a hybrid neural-evolutive approach. In: MICAI 2009: Advances in Artificial Intelligence, pp. 600–609. Springer (2009)
9. Foley, A.M., Leahy, P.G., Marvuglia, A., McKeogh, E.J.: Current methods and advances in forecasting of wind power generation. Renewable Energy 37(1), 1–8 (2012)
10. Giacomoni, M., Kanta, L., Zechman, E.: Complex adaptive systems approach to simulate the sustainability of water resources and urbanization. Journal of Water Resources Planning and Management 139(5), 554–564 (2013)
11. Hayashi, M., Kermanshahi, B.: Application of artificial neural network for wind speed prediction and determination of wind power generation output. In: Proceedings of ICEE. pp. 12–15 (2001)
12. Lei, M., Shiyan, L., Chuanwen, J., Hongling, L., Yan, Z.: A review on the forecasting of wind speed and generated power. Renewable and Sustainable Energy Reviews 13(4), 915–920 (2009)
13. Liu, H., Tian, H., Pan, D., Li, Y.: Forecasting models for wind speed using wavelet, wavelet packet, time series and artificial neural networks. Applied Energy 107, 191–208 (2013)
14. Neurolab: Neurolab a python plugin. Available: https://pypi.python.org/pypi/neurolab [Online]
15. Observ, E.: Worldwide electricity production from renewable energy sources. Ninth inventory (2012)

16. Piwko, D., Jordan, G.: The economic value of day-ahead wind forecasts for power grid operations. In: 2010 UWIG Workshop on Wind Forecasting (2010)
17. Rangel, H.R., Puig, V., Farias, R.L., Flores, J.J.: Short-term demand forecast using a bank of neural network models trained using genetic algorithms for the optimal management of drinking water networks. Journal of Hydroinformatics (2016), http://jh.iwaponline.com/content/early/2016/08/31/hydro.2016.199
18. Rodriguez Rangel, H., Puig, V., Farias, R.L., Flores, J.J.: Short-term demand forecast using a bank of neural network models trained using genetic algorithms for the optimal management of drinking water networks. Journal of Hydroinformatics (2016)
19. Yesilbudak, M., Sagiroglu, S., Colak, I.: A new approach to very short term wind speed prediction using k-nearest neighbor classification. Energy Conversion and Management 69, 77–86 (2013)
20. Zhao, X., Wang, S., Li, T.: Review of evaluation criteria and main methods of wind power forecasting. Energy Procedia 12, 761–769 (2011)
21. Zhou, J., Shi, J., Li, G.: Fine tuning support vector machines for short-term wind speed forecasting. Energy Conversion and Management 52(4), 1990–1998 (2011)

ISSN 1870-4069

Pronóstico del tipo de cambio USD/MXN con redes neuronales de retropropagación

Francisco D. Meneses-Bautista, Matías Alvarado

Centro de Investigación y de Estudios Avanzados del IPN,
Departamento de Computación,
Ciudad de México, México

fmeneses@computacion.cs.cinvestav.mx, matias@cs.cinvestav.mx

Resumen. El análisis de las series de tiempo permite caracterizar un fenómeno e incluso predecir, con un cierto grado de precisión, su comportamiento a futuro. La posibilidad de anticipar los movimientos de los mercados resulta sumamente atractiva para los responsables de la toma de decisiones, tanto en la iniciativa privada como en el sector público. En esta investigación se aborda la regresión de series de tiempo del tipo de cambio dólar estadounidense/peso mexicano, empleando un modelo generado por redes neuronales artificiales de retropropagación. Los resultados obtenidos ratifican empíricamente las ventajas de la utilización de las redes neuronales en el análisis y pronóstico de series de tiempo financieras, producto de las capacidades de aproximación de funciones y generalización que presentan dichas redes.

Palabras clave: pronóstico de tipos de cambio, redes neuronales artificiales, series de tiempo.

Forecasting USD/MXN Exchange Rate with Backpropagation Neural Networks

Abstract. The analysis of time series allows to characterize a phenomenon and even to predict, under a certain degree of precision, its future behavior. The possibility of anticipating market movements is extremely attractive for decision-makers, both in the private sector and in the public sector. This research addresses the regression of time series of the US dollar / Mexican peso exchange rate, using a model generated by backpropagation artificial neural networks. The results obtained ratify empirically the advantages of the use of neural networks in the analysis and forecast of financial time series, as a result of the capabilities of function approximation and generalization that these networks present.

Keywords: exchange rates forecasting, artificial neural networks, time series.

1. Introducción

Conocer con antelación el estado futuro de ciertas variables financieras permite obtener el mayor beneficio en un entorno de volatilidad. En las áreas de la economía y las finanzas, los especialistas han usado tradicionalmente diferentes herramientas estadísticas y econométricas de análisis de series de tiempo, con la finalidad de comprender mejor la mecánica inherente a fenómenos económicos complejos. Se han desarrollado métodos analíticos que pretenden explicar la manera en la que la interacción entre las variables independientes da forma al comportamiento observado (análisis fundamental). Sin embargo, también existen otros métodos que no buscan explicar las relaciones causales entre los factores y el fenómeno que producen, sino que sitúan el foco de interés en las propias series de tiempo y su gráfica (análisis técnico). Desde esta última perspectiva, la serie de tiempo contiene la suficiente información sobre cómo cambia el fenómeno a través del tiempo, de tal forma que es posible utilizarla para realizar regresiones y pronósticos. Son muy utilizados los modelos *ARIMA* (en especial, la metodología de Box-Jenkins [1]), los modelos de heterocedasticidad condicional autoregresiva *ARCH* y *GARCH*, y el modelo del *Paseo Aleatorio* [2].

Sin embargo, en las últimas décadas, el auge de las tecnologías de la información ha permitido desarrollar sistemas autoadaptativos de cómputo, como las redes neuronales artificiales, que son capaces de hallar, de manera automática, las correlaciones subyacentes en medio de grandes conjuntos de datos aparentemente inconexos. Esta habilidad es de gran valor para el análisis de regresión, puesto que permite determinar en qué grado influye un conjunto de variables independientes en el comportamiento de un fenómeno dado. Además, sometiendo a la red neuronal a un proceso de entrenamiento adecuado, es posible obtener un modelo de cómputo implícito que no sólo reproduzca los datos de la serie de tiempo, sino que también pueda ser utilizado para extrapolar puntos futuros de la serie. En este trabajo, se emplearon redes neuronales entrenadas con el algoritmo de retropropagación para analizar y pronosticar el comportamiento del tipo de cambio dólar estadounidense/peso mexicano (USD/MXN). El objetivo de los experimentos fue evaluar la capacidad predictiva de las redes neuronales de retropropagación, probando diferentes arquitecturas y conjuntos de datos: series de tiempo del tipo de cambio del peso mexicano respecto a varias divisas, tomadas en distintos periodos históricos.

El campo de regresión de series de tiempo con redes neuronales ha sido abordado en varias investigaciones y desde una amplia variedad de perspectivas. Zhang et al. [8] realizaron experimentos sobre el pronóstico del tipo de cambio GBP/USD con redes neuronales. Analizan cómo responden diferentes diseños de arquitecturas, cambiando tanto el tamaño de los vectores de entrada, como el conjunto de objetos de entrenamiento. Al final, concluyen que el número de nodos de entrada tiene un impacto importante en su modelo y que el índice de desempeño de la red (ya sea la raíz del error cuadrático medio *RMSE* o el porcentaje medio de error absoluto *MAPE*) sirve también para determinar el tamaño del conjunto de objetos de entrenamiento usado. Al final determinan que la red neuronal tiene una mayor capacidad predictiva que el modelo del

Paseo Aleatorio, sobre todo, cuando se trata de periodos mensuales, puesto que en periodos más prolongados, ambos métodos obtienen resultados muy similares.

En el trabajo de Yao et al. [7], se consigue hacer una aproximación de tipos de cambio con tasas de acierto superiores al 70 por ciento, empleando redes neuronales de retropropagación (*BPNN*). Los autores desarrollan un modelo para pronosticar las variaciones en los tipos de cambio, que incorpora algunas variables de tipo cualitativo. Además, complementan su investigación implementando un simulador de inversiones.

En años más recientes, Saini et al. [6] realizaron una comparación entre la capacidad de predicción de tipos de cambio que tienen las redes *BPNN* y redes neuronales recurrentes (*RNN, Recurrent Neural Network*). Calcularon independientemente la regresión de cuatro series de tiempo, de divisas distintas. Los experimentos se hicieron sobre una red *BPNN* y una *RNN*, ambas con arquitectura fija. Los parámetros del experimento fueron el factor de aprendizaje, la cantidad de épocas de entrenamiento y el *momentum*. Sus resultados confirman la habilidad de ambos tipos de redes neuronales para aproximar satisfactoriamente las series de tiempo. Sin embargo, las redes *RNN* tuvieron, en general, un mejor desempeño en sus experimentos.

En el trabajo de Galeshchuk [4] se analiza el efecto que se genera sobre la predicción al utilizar series de tiempo con diferentes intervalos de muestreo. El modelo de pronóstico utilizado en sus experimentos es un perceptrón multicapa, y se alimenta con series de tiempo con muestras de frecuencia diaria, mensual y cuatrimestral. Los resultados empíricos obtenidos indican que la red entrega una mejor predicción cuando el intervalo solicitado es a corto plazo.

2. Pronóstico con redes neuronales

Una red neuronal artificial es un modelo de cómputo autoadaptativo y bioinspirado. En ella, se trata de imitar la gran cantidad de interconexiones que presentan las células nerviosas, de manera que el arreglo global de células individuales simples alcanza niveles de procesamiento muy elevados, brindándole al organismo la capacidad de enfrentarse a problemas muy complejos, con gran eficacia y precisión.

Las neuronas artificiales son las unidades básicas de procesamiento en las redes neuronales artificiales. Cada neurona artificial efectúa una evaluación de los valores que recibe como entrada, a través de una *función de activación*. Suele elegirse como función de activación alguna que sea de tipo no lineal y continua. Las neuronas artificiales se unen entre sí por medio de enlaces sinápticos, ponderados por un conjunto de *pesos*. Se llama *capa* a la agrupación de neuronas que se encuentran en el mismo nivel de profundidad en la red (Figura 1). Las señales de salida de una capa alimentan a todas las neuronas de la capa siguiente. La función de salida de la red neuronal es una composición de las funciones de activación de todas las neuronas en la red.

El problema del *aprendizaje* de la red neuronal consiste en encontrar un conjunto adecuado de pesos, que permita a la red producir la salida esperada.

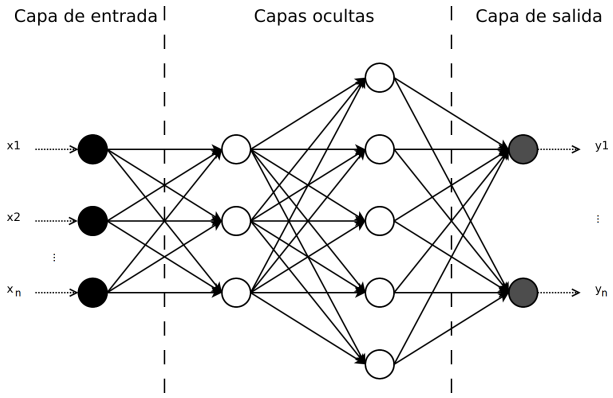


Fig. 1. Red neuronal artificial de cuatro capas.

En el *entrenamiento* por retropropagación, se busca minimizar el error global de la red, expresado como la diferencia entre la salida producida por la red y la salida esperada, ante un conjunto conocido de patrones de entrada. El proceso de retropropagación calcula la aportación al error global proveniente de cada neurona en la capa de salida, reajusta el conjunto de pesos de la capa de salida de acuerdo a un *factor de aprendizaje* y al gradiente de la función de error, y continúa con la capa anterior (aquella que está directamente conectada con la capa de salida). Este procedimiento sigue con cada capa de la red hasta llegar a la capa de entrada. Tras cada *época* del entrenamiento, el error de la red va reduciéndose hasta llegar al mínimo. Después de un entrenamiento exitoso, la red neuronal es capaz de reproducir las salidas esperadas con un nivel de error despreciable, de manera que puede ser empleada como un aproximador universal de funciones [3,5].

Las habilidades de las redes neuronales de generalización y de aproximación de funciones son muy importantes para las tareas de regresión y pronóstico con series de tiempo. La red neuronal entrenada puede concebirse como un modelo de regresión implícito, que reproduce las observaciones de la serie de tiempo al ser alimentado con un conjunto de variables independientes. Entonces, para utilizar este modelo en el pronóstico, se debe tomar como entrada valores de las variables independientes hasta el momento t y entrenar a la red para que produzca el valor de la serie de tiempo en el momento $t+1$ como salida.

3. Desarrollo experimental

3.1. Variables independientes y conjunto de datos

Sea $F_t = f(\boldsymbol{x_t})$ el valor de la función f en el momento t, y $\boldsymbol{x_t}$ el vector cuyas componentes $(x_t^1, x_t^2, ..., x_t^n)$ representan al conjunto de las n variables independientes que describen al fenómeno, en el momento t.

El modelo propuesto en esta investigación establece que la función pronóstico g, calculada mediante redes neuronales, determina las correlaciones que permiten inferir el siguiente valor de la serie de tiempo (F_{t+1}), dadas las variables independientes $x_t^1, x_t^2, ..., x_t^n$, junto al valor de la función F_t. El valor siguiente de la serie de tiempo F_{t+1}, obtenido de la función g así definida, queda expresado como

$$F_{t+1} = g(\boldsymbol{x_t}, F_t) = g(x_t^1, x_t^2, ..., x_t^n, F_t).$$

La decisión sobre cuáles variables integrarán los datos de entrada puede realizarse a través de un análisis causal o estadístico. Sin embargo, el enfoque adoptado en este trabajo consistió en probar diferentes combinaciones de series de tiempo como variables independientes. Al finalizar el entrenamiento de las redes neuronales, puede calcularse un índice de error y asociarlo a cada conjunto de variables independientes. De esta manera, el mejor conjunto para realizar el pronóstico es aquél que produce el menor grado de error durante la validación.

Los datos empleados en los experimentos fueron obtenidos del sitio web del Banco de México[1]. En el sitio se hayan disponibles valores diarios del tipo de cambio peso/dólar y los precios de las divisas que integran la canasta del DEG (Derecho Especial de Giro), a saber, el dólar estadounidense, el euro, la libra esterlina, el yen japonés y el yuan chino. Para el desarrollo de los experimentos fueron elegidas como variables independientes cuatro de las divisas mencionadas: el dólar estadounidense, el euro, la libra esterlina y el yen japonés. El yuan chino fue descartado debido a la escasez de los datos disponibles[2]. A partir de estas cuatro series de tiempo, se generaron varios conjuntos de vectores de entrenamiento. Los objetos que integran los conjuntos de entrenamiento pueden describirse como vectores de tipo:

$$\boldsymbol{x_t} = (x_t^{\mathrm{USD/MXN}}, x_t^{\mathrm{EUR/MXN}}, x_t^{\mathrm{GBP/MXN}}, x_t^{\mathrm{JPY/MXN}}), \quad \text{con} \quad \boldsymbol{x_t} \in \mathbb{R}^4,$$

donde

- $x_t^{\mathrm{USD/MXN}} \in \mathbb{R}$, es el valor del tipo de cambio dólar estadounidense / peso mexicano, en el instante t,
- $x_t^{\mathrm{EUR/MXN}} \in \mathbb{R}$, es el valor del tipo de cambio euro / peso mexicano, en el instante t,
- $x_t^{\mathrm{GBP/MXN}} \in \mathbb{R}$, es el valor del tipo de cambio libra esterlina / peso mexicano, en el instante t,
- $x_t^{\mathrm{JPY/MXN}} \in \mathbb{R}$, es el valor del tipo de cambio yen japonés / peso mexicano, en el instante t.

Se probaron cinco diferentes conjuntos de objetos de entrenamiento. El primero de ellos está formado por los vectores descritos, con cuatro componentes. Los otros conjuntos de objetos de entrenamiento contienen vectores cuyas componentes son una combinación de sólo tres de las cuatro divisas elegidas anteriormente

[1] `http://www.banxico.org.mx`

[2] Sólo se tienen registros del yuan chino desde octubre de 2016 (año en que se integró al DEG).

(Tabla 1). La hipótesis inicial sobre los conjuntos de entrenamiento fue que, dada una red neuronal con una arquitectura determinada, los datos de las cuatro series de tiempo permiten obtener los mejores resultados en el pronóstico, ya que al descartar cualquiera de las series, se afecta el modelo calculado por la red.

Tabla 1. Componentes de los vectores en los conjuntos de entrenamiento.

Conjunto	Variables independientes
00deyl	USD/MXN, EUR/MXN, GBP/MXN, JPY/MXN
01eyl	EUR/MXN, GBP/MXN, JPY/MXN
02dyl	USD/MXN, GBP/MXN, JPY/MXN
03del	USD/MXN, EUR/MXN, JPY/MXN
04dey	USD/MXN, EUR/MXN, GBP/MXN

El conjunto de objetivos para el entrenamiento está constituido por series de tiempo con valores $y_t = x_{t+1}^{\mathrm{USD/MXN}} \in \mathbb{R}$, es decir, el tipo de cambio dólar estadounidense/peso mexicano en el instante $t + 1$.

Los datos de las series de tiempo empleadas se componen de muestras diarias que corresponden al periodo entre el 2 de Enero de 2015 y el 24 de Marzo de 2017. De la totalidad de estos datos, se formaron tres subconjuntos de entrenamiento con distinta cantidad de objetos: el primero con datos en un periodo semestral; el segundo, anual y el último, a dos años (Tabla 2). Para la validación (pronóstico) se eligieron, en los tres casos, series de tiempo con 30 muestras (equivalentes, aproximadamente, a un periodo de mes y medio), a partir del día inmediatamente posterior al último objeto en cada serie de entrenamiento.

Tabla 2. Detalle de los periodos elegidos para entrenamiento y pronóstico.

	Entrenamiento			Pronóstico		
Periodo	Desde	Hasta	Objetos	Desde	Hasta	Objetos
Semestral	01-Jul-2016	31-Ene-2017	149	01-Feb-2017	15-Mar-2017	30
Anual	04-Ene-2016	30-Dic-2016	252	02-Ene-2017	13-Feb-2017	30
Dos años	02-Ene-2015	08-Feb-2017	530	09-Feb-2017	23-Mar-2017	30

Los valores de las series de tiempo recibieron un procesamiento previo para mejorar el desempeño durante el entrenamiento de las redes neuronales. Se realizó un escalamiento en todos los valores de las series, de tal manera que a cada muestra se le hace corresponder un valor real en el intervalo $[0, 1]$, muy

apropiado para usarse en las redes de retropropagación con función de activación sigmoidal logística.

En el caso particular del yen japonés, la serie de tiempo fue sometida, en primer lugar, a un escalamiento, por un factor de 100, debido que el valor nominal del yen, desde el año 2015, ha estado alrededor de los 15 centavos de peso mexicano. Posteriormente, todas las series de tiempo fueron transformadas de acuerdo al procedimiento que se describe a continuación.

Sea F_t el valor de una serie de tiempo en el momento t, y sean $F_{\min}$ y $F_{\max}$ dos valores constantes, extremos del intervalo $[F_{\min}, F_{\max}]$, y tales que $F_{\min}$ es menor o igual que todas las observaciones que componen a las series de tiempo y, de manera similar, $F_{\max}$ es mayor o igual que todos las muestras de dichas series. Entonces, el valor escalado $F_{N(t)}$ que representa al valor F_t de la serie de tiempo original, puede ser expresado de esta manera:

$$F_{N(t)} = \frac{F_t - F_{\min}}{F_{\max} - F_{\min}}, \qquad 0 \leq F_{N(t)} \leq 1.$$

Los valores elegidos para delimitar el intervalo $[F_{\min}, F_{\max}]$, anteriormente descrito, fueron $F_{\min} = 10.00$ y $F_{\max} = 30.00$. El valor de $F_{\min}$ se determinó al observar que el valor mínimo en las series de tiempo corresponde al precio que presentó el yen el día 6 de Marzo de 2015, con un valor nominal de 0.1221 pesos, que sujeto al escalamiento inicial de la serie de tiempo del yen, equivale a 12.12 pesos. El valor $F_{\min}$ elegido brinda una holgura de 2 pesos, aproximadamente. En el caso del valor $F_{\max}$ elegido, éste supera por casi 3 pesos al valor máximo presente en las series de tiempo, correspondiente al valor nominal de la libra el día 12 de Febrero de 2016, cuando se cotizó en 27.6716 pesos. El intervalo holgado escogido hace que sea posible realizar pronósticos que rebasen los límites mínimo y máximo de los valores de las series de tiempo.

3.2. Diseño de la red neuronal de retropropagación

Denótese como N a la función de pronóstico obtenida por la red neuronal después del entrenamiento. Entonces, la definición de la función N en términos de la funciones de regresión f y de pronóstico g, anteriormente mencionadas, es

$$N(\boldsymbol{x_t}) = g(\boldsymbol{x_t}, F_t) = f(\boldsymbol{x_{t+1}}) = y_t,$$

que, expresada en términos de las cuatro variables independientes utilizadas, resulta

$$N(\boldsymbol{x_t}) = N(x_t^{\text{EUR/MXN}}, x_t^{\text{GBP/MXN}}, x_t^{\text{JPY/MXN}}, x_t^{\text{USD/MXN}}) = y_t = x_{t+1}^{\text{USD/MXN}}.$$

Con el fin de encontrar una arquitectura que modelara apropiadamente la función N, se ensayaron redes con distinta cantidad de capas y neuronas (Figura 2). Para todas las arquitecturas de red probadas, el número de entradas fue tres o cuatro, dependiendo del conjunto de variables independientes utilizado. La capa de salida siempre fue construida con una única neurona.

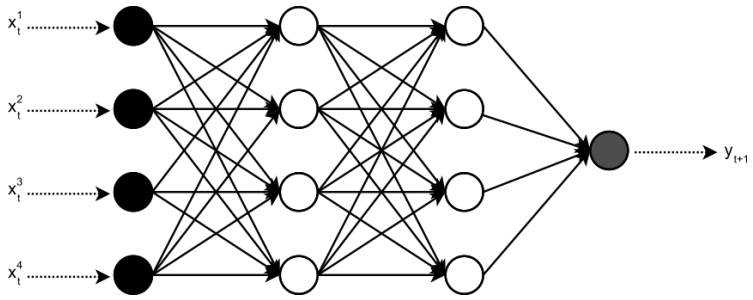


Fig. 2. Red de retropropagación de arquitectura a4441.

Se probaron en total doce arquitecturas para cada caso de pronóstico, con una o dos capas ocultas, y cada capa contó con dos, cuatro u ocho neuronas (Tabla 3). Se empleó la función sigmoidal logística como función de activación en todas las neuronas de la red ($\text{logsig}(x) = \frac{1}{1+e^{-x}}$).

Tabla 3. Arquitecturas probadas en los experimentos.

Arquitecturas		Neuronas por capa oculta	
3 entradas	4 entradas	1ª capa	2ª capa
a321	a421	2	0
a341	a441	4	0
a381	a421	8	0
a3221	a4221	2	2
a3241	a4241	2	4
a3281	a4281	2	8
a3421	a4421	4	2
a3441	a4441	4	4
a3481	a4481	4	8
a3821	a4821	8	2
a3841	a4841	8	4
a3881	a4881	8	8

3.3. Entrenamiento

El proceso de entrenamiento fue de tipo on-line: para cada patrón de entrenamiento evaluado por la red, se calculó el error cuadrático medio (MSE) como estimador del error global de la red, y con este índice de error, se efectúo el ajuste de pesos por retropropagación. El factor de aprendizaje fue fijado a η =0.4. El proceso fue restringido a 100,000 épocas de entrenamiento, y el error mínimo

deseado fue de $RMSE$=0.001 (raíz del error cuadrático medio acumulado en una época). Se registró el grado de error obtenido por cada arquitectura, tanto en la regresión de la serie de tiempo de entrenamiento, como en la de pronóstico.

4. Resultados

En la Tabla 4 se muestran los resultados generales de las etapas de entrenamiento y de regresión. Para ambas etapas, se presenta el *RMSE* promedio obtenido, por todas las arquitecturas y periodos de entrenamiento. Cabe mencionar que el error mínimo deseado para el entrenamiento no fue alcanzado por ninguna arquitectura (el *RMSE* mínimo obtenido fue de 0.00453514, con la arquitectura *a4881* (4 entradas, 8 neuronas en la primera capa oculta, 8 neuronas en la segunda capa oculta y una salida), con el conjunto de variables *00deyl* (dólar-euro-yen-libra) y con datos de periodo anual).

Tabla 4. RMSE promedio del entrenamiento y la regresión.

Variables	Entrenamiento	Regresión
00deyl	0.0061432136	0.0096681008
01eyl	0.0092615111	0.0168266333
02dyl	0.0063488836	0.0094228061
03del	0.0062851569	0.0097144117
04dey	0.0063465556	0.0101616786

De los resultados obtenidos en la etapa de pronóstico fue posible identificar dos casos distintos: el pronóstico de 30 muestras, con un mayor grado de error, y el pronóstico de 1 muestra, con el menor grado de error. En la Tabla 5 se muestra el *RMSE* en el prónostico, empleando las series de tiempo de validación completas (con las 30 muestras), para cada conjunto de variables utilizado. También se incluyen los valores mínimo, maxímo y promedio en cada caso. En la Tabla 6 se pueden observar los mismos datos, esta vez, para el pronóstico de la primera muestra en las series de tiempo de validación, es decir, el pronóstico al siguiente día).

4.1. Análisis

De los datos mostrados en la Tabla 5, se puede destacar la siguiente información sobre el pronóstico de 30 muestras:

- El mejor pronóstico se presentó al usar las variables *02dyl* (dólar-yen-libra), la arquitectura *a341* (3 entradas, 4 neuronas en una capa oculta y 1 salida), con un periodo de un año en las series de entrenamiento y obteniendo un *RMSE* de 0.00842705 (Figura 3).

Tabla 5. RMSE del pronóstico de 30 muestras.

RMSE por cada periodo

	Menor RMSE			Mayor RMSE		
Variables	Arquitectura	Periodo	RMSE	Arquitectura	Periodo	RMSE
00deyl	a481	1y	0.00846566	a4881	6m	0.0139388
01eyl	a341	2y	0.0146082	a341	1y	0.0789643
02dyl	a341	1y	0.00842705	a381	2y	0.0129769
03del	a3841	1y	0.00853617	a3821	2y	0.0164238
04dey	a341	1y	0.00908241	a3881	6m	0.013041

RMSE promedio de todos los periodos

	Menor RMSE		Mayor RMSE	
Variables	Arquitectura	RMSE	Arquitectura	RMSE
00deyl	a4821	0.0124279333	a4221	0.0106122233
01eyl	a3281	0.0178648667	a341	0.0382287333
02dyl	a3221	0.0106841667	a3821	0.0115879667
03del	a341	0.0106401867	a3821	0.0123929067
04dey	a3241	0.0098830933	a3821	0.0110833733

- El pronóstico con el mayor grado de error fue producido por la red con arquitectura *a341* (3 entradas, 4 neuronas en una capa oculta y 1 salida), con el conjunto de variables *01eyl* (euro-yen-libra), con datos de entrenamiento de un año, y con *RMSE* de 0.0789643 (Figura 4).
- Considerando el valor promedio de los tres periodos de entrenamiento, el mejor caso de pronóstico se obtiene de la red con arquitectura *a3241* (3 entradas, 2 neuronas en la primera capa oculta, 4 neuronas en la segunda capa oculta y 1 salida), usando como el conjunto de variables *04dey* (dólar-euro-yen), con valor *RMSE* de 0.0098830933.

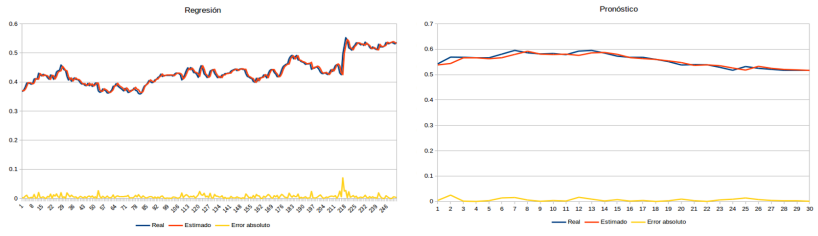


Fig. 3. Regresión y pronóstico a341(02dyl).

Tabla 6. RMSE del pronóstico de 1 muestra (al día siguiente).

RMSE por cada periodo

	Menor RMSE			Mayor RMSE		
Variables	Arquitectura	Periodo	RMSE	Arquitectura	Periodo	RMSE
00deyl	a4881	2y	0.000442617	a4821	6m	0.0139143
01eyl	a3221	2y	0.000981482	a3841	1y	0.037046
02dyl	a3841	2y	0.000979935	a3221	1y	0.025883
03del	a3441	2y	0.00257343	a3821	6m	0.00879175
04dey	a3421	2y	0.003174	a3841	6m	0.0156557

RMSE promedio de todos los periodos

	Menor RMSE		Mayor RMSE	
Variables	Arquitectura	RMSE	Arquitectura	RMSE
00deyl	a4881	0.0061538857	a4221	0.0076484267
01eyl	a3441	0.00536915	a3881	0.0164226433
02dyl	a341	0.00756548	a3221	0.0122008333
03del	a3481	0.0052086567	a3821	0.00672563
04dey	a321	0.0079934933	a3821	0.0091191733

Por otra parte, en la Tabla 6, para el caso del pronóstico a 1 día, se observa que:

- El mejor pronóstico se presentó al usar las variables *00deyl* (dólar-euro-yen-libra), la arquitectura *a4881* (4 entradas, 8 neuronas en la primera capa oculta, 8 neuronas en la segunda capa oculta y 1 salida), con un periodo de 2 años en las series de entrenamiento y obteniendo un *RMSE* de 0.000442617 (Figura 5).
- El pronóstico con el mayor grado de error fue producido por la red con arquitectura *a3841* (3 entradas, 8 neuronas en la primera capa oculta, 4 neuronas en la segunda capa oculta y 1 salida), con el conjunto de variables

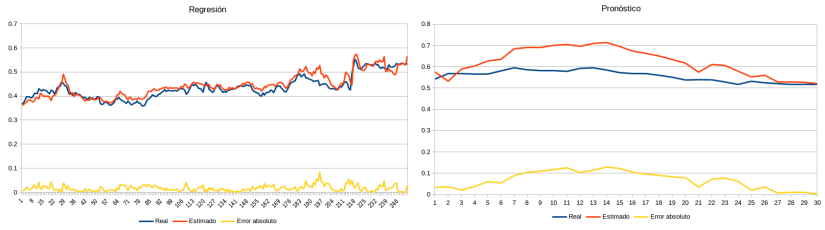


Fig. 4. Regresión y pronóstico a341(01eyl).

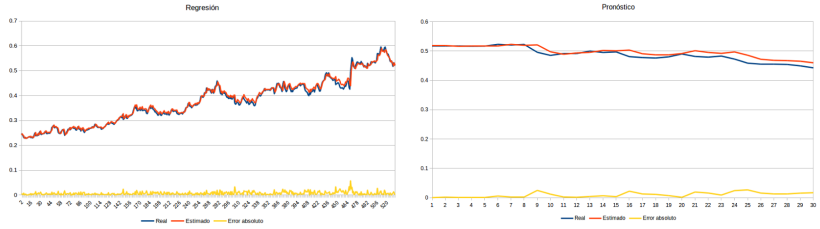


Fig. 5. Regresión y pronóstico a4881(00deyl).

01eyl (euro-yen-libra), con datos de entrenamiento de un año, y con *RMSE* de 0.037046 (Figura 6).

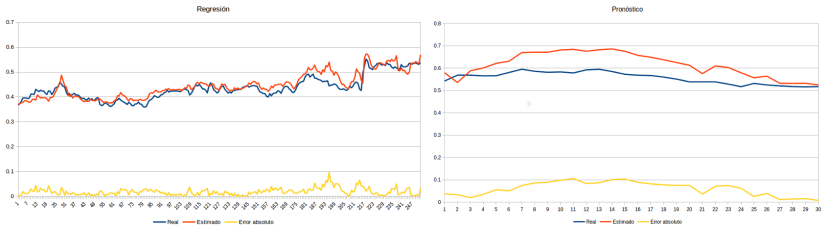


Fig. 6. Regresión y pronóstico a3841(01eyl).

- Considerando el valor promedio de los tres periodos de entrenamiento, el mejor pronóstico se obtiene de la red con arquitectura *a3481* (3 entradas, 4 neuronas en la primera capa oculta, 8 neuronas en la segunda y 1 salida), usando como el conjunto de variables *03del* (dólar-euro-libra), con valor *RMSE* de 0.0052086567.

Al evaluar el pronóstico a 30 muestras calculado por las diferentes redes, se encontró que la mejor predicción fue producida por la arquitectura *a341* (3 entradas, 4 neuronas en una capa oculta y 1 salida). Sin embargo, el grado de error obtenido es equivalente, en pesos mexicanos, a una variación de $(F_{\max} - F_{\min})(RMSE)$=(20.0)(0.00842705)=0.168541, es decir, aproximadamente 17 centavos de peso mexicano. Este elevado grado de error puede ser atribuido al modelo de pronóstico propuesto, el cual está diseñado para pronosticar el valor de la variable USD/MXN únicamente al día siguiente. Además, es necesario notar que la misma arquitectura reportó el mayor grado de error con el conjunto de variables *01eyl* (euro-yen-libra), por lo que, con estos datos, no se puede llegar a una afirmación concluyente sobre la capacidad predictiva de esta arquitectura.

Un razonamiento similar se sigue al evaluar los conjuntos de entrenamiento. Los resultados obtenidos sugieren que el mejor conjunto de variables independientes es el conjunto *02dyl* (dólar-yen-libra), con datos dentro de un periodo

anual. Esta evidencia empírica rechaza, inicialmente, la hipótesis establecida en la sección 3.1. Sin embargo, al observar los resultados obtenidos, no se puede identificar una tendencia relevante que permita argumentar firmemente a favor o en contra de ese conjunto de variables, dado que, usando la arquitectura equivalente *a441* (4 entradas, 4 neuronas en una capa oculta y 1 salida) y el mismo periodo anual en los datos de entrenamiento, pero con el conjunto de datos *00deyl* (dólar-euro-yen-libra), se obtiene un *RMSE* de 0.00846566, es decir, una diferencia de 0.00003861, equivalente a 0.07722 centavos de peso mexicano.

En cuanto al pronóstico a 1 día, es notable la disminución en el grado de error obtenido, con respecto al pronóstico a 30 días, pasando de 0.00842705 a 0.000442617, en los mejores pronósticos para cada caso. Esta disminución del grado de error puede explicarse por la manera en que fueron entrenadas las redes neuronales: la serie de tiempo modelada por la red corresponde al valor del tipo de cambio dólar estadounidense/peso mexicano al día siguiente. Además, al comparar el *RMSE* promedio obtenido, puede afirmarse que el pronóstico a 1 día (con *RMSE* promedio de 0.0052086567) supera en calidad al pronóstico de 30 observaciones (con *RMSE* promedio de 0.0098830933).

Los resultados del pronóstico a 1 día indican que la arquitectura con mejores capacidades predictivas es *a4881* (4 entradas, 8 neuronas en la primera capa oculta, 8 neuronas en la segunda capa oculta y 1 salida), que es también, la arquitectura más compleja de las que se probaron. El *RMSE* de 0.000442617 (equivalente a un error absoluto de 0.885234 centavos de peso mexicano), obtenido por esta arquitectura, supera al siguiente mejor pronóstico, que se consiguió con las variables *02dyl*(dólar-yen-libra) y la arquitectura *a3841*, con datos en periodo de 2 años y *RMSE* de 0.000979935. La diferencia entre los dos pronósticos fue de 0.000537318, es decir, más del doble del *RMSE* obtenido en el primer caso. Al comparar los conjuntos de entrenamiento, el mejor pronóstico se consiguió con el conjunto *00deyl* (dólar-euro-yen-libra), apoyando la hipótesis de la sección 3.1.

Del análisis de los pronósticos con mayor grado de error, es importante hacer notar que, tanto para el caso de una muestra como para 30, el conjunto de datos *01eyl*(euro-yen-libra), con datos en periodo de un año, se ve asociado con esos pronósticos de baja calidad. Dicho conjunto es el único que no incluye la variable USD/MXN. Este resultado puede indicar que la información que aporta el valor del tipo de cambio USD/MXN en un determinado día es escencial para calcular el pronóstico al día siguiente y, por lo tanto, no debe ser descartada.

5. Conclusiones

Los experimentos realizados prueban la habilidad que tienen las redes neuronales de retropropagación para modelar y reproducir de manera adecuada la serie de tiempo que describe la variación del tipo de cambio USD/MXN. En el pronóstico, las redes neuronales calcularon predicciones de mejor calidad en el corto plazo (al día siguiente) que en el caso de 30 muestras a futuro. Este deterioro en la calidad de los pronósticos puede haberse debido al modelo empleado, que sólo se enfoca en realizar la predicción al día siguiente. Los datos

empleados durante el entrenamiento de las redes fueron adaptados para concordar completamente con dicho modelo, y esto puede representar una dificultad para que se alcance un nivel de generalización aceptable que permita predecir correctamente varias muestras. Será necesario continuar con las investigaciones para determinar si una extensión en el modelo puede mejorar la capacidad de predicción en plazos mayores.

Como parte del análisis de las variables empleadas, los resultados del pronóstico a un día apoyan de decisión de utilizar de los cuatro tipos de cambio en el modelo, sin descartar ninguno, así como la importancia de incluir el mismo tipo de cambio peso/dólar para mejorar la predicción. Sin embargo, este resultado está limitado a las varibles elegidas en esta investigación y sería interesante analizar el efecto del uso de un conjunto ampliado de variables independientes, que no necesariamente se restrinja a tipos de cambio.

El pronóstico a un día alcanzó un error absoluto menor a un centavo, en el mejor de los casos. Si bien, se puede considerar que es un grado de error aceptable, es necesario evaluar el modelo con más periodos de prueba, con la finalidad de observar si la calidad del pronóstico se mantiene a través del tiempo, ante las condiciones cambiantes que se presentan cotidianamente en los mercados de divisas.

Agradecimientos. Agradecemos al Consejo Nacional de Ciencia y tecnología y al Centro de Investigación y de Estudios Avanzados del IPN por el apoyo recibido para la realización de este trabajo de investigación.

Referencias

1. Box, G., Jenkins, G.: Time series analysis: forecasting and control. Holden-Day series in time series analysis, Holden-Day (1970)
2. Brockwell, P.J., Davis, R.A.: Introduction to time series and forecasting, pp. 5–12. Springer International Publishing, 3rd edn. (2016)
3. Cybenko, G.: Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems 2(4), 303–314 (1989)
4. Galeshchuk, S.: Neural networks performance in exchange rate prediction. Neurocomputing 172, 446–452 (2016)
5. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Networks 2(5), 359 – 366 (1989)
6. Saini, S.S., Parkhe, O., Khadtare, T.: Analysis of feedforward and recurrent neural network in forecasting foreign exchange rate. Imperial Journal of Interdisciplinary Research 2(6) (2016)
7. Yao, J., Tan, C.L.: A case study on using neural networks to perform technical forecasting of forex. Neurocomputing 34(1-4), 79 – 98 (2000)
8. Zhang, G., Hu, M.Y.: Neural network forecasting of the british pound/us dollar exchange rate. Omega 26(4), 495–506 (1998)

ISSN 1870-4069

Algoritmo de desactivación de estaciones base para reducir el consumo de energía en redes celulares

Zury Jeheili Santiago Manzano[1], Anabel Martínez Vargas[1], Ángel G. Andrade[2]

[1] Universidad Politécnica de Pachuca, Zempoala, Hidalgo, México

[2] Universidad Autónoma de Baja California, Facultad de Ingeniería, Mexicali, Baja California, México

zury_santiago@micorreo.upp.edu.mx, anabel.martinez@upp.edu.mx, aandrade@uabc.edu.mx

Resumen. Los teléfonos inteligentes se han convertido en dispositivos indispensables para las actividades cotidianas que realiza el ser humano. Su proliferación ha incrementado la infraestructura celular en un esfuerzo de los operadores móviles por satisfacer la demanda. Dado que la planificación de las redes celulares se realiza considerando la demanda en horas pico, se despliegan un gran número de estaciones base, manteniendo el mismo número de estaciones base aun cuando la intensidad de tráfico se reduzca. Esta estrategia ha traído consigo que se eleven los niveles de energía en las redes celulares, afectando no solo a sus gastos operativos sino también contribuyendo con las emisiones de carbono a la atmósfera. En el presente trabajo se muestra un algoritmo de desactivación de estaciones base para ahorrar energía en una red celular considerando la reasignación de usuarios móviles. Para encontrar el conjunto de estaciones base que deberán desactivarse y seguir ofreciendo los servicios demandados, se utiliza un algoritmo genético con modelo de población estacionario.

Palabras clave: algoritmo genético, redes verdes, LTE, ahorro de energía.

Base Station Switching On/Off Algorithm to Reduce Energy Consumption in Cellular Networks

Abstract. Smartphones have become essential devices for the daily activities that the human being performs. Its proliferation has increased the cellular infrastructure in an effort of the mobile operators to satisfy the demand. Due to the cellular network planning is performed considering the peak hour, a large number of base stations are deployed, maintaining the same number of base stations even in low traffic states. This strategy led to high levels of energy in cellular networks, affecting their operating expenditure and contributing to CO_2 emissions. The present work shows a base station switching on/off algorithm to save energy in a cellular network considering the reassignment of mobile users.

In order to find out the base stations to switch off, a genetic algorithm with steady-state population model is used.

Keywords: genetic algorithm, green networks, LTE, energy saving.

1. Introducción

Los teléfonos inteligentes se han convertido en dispositivos indispensables para las actividades cotidianas que realiza el ser humano. Se estima que para el año 2020, el 72% de la población mundial contará con un dispositivo móvil [1]. La planificación de las redes celulares se realiza considerando el peor escenario, es decir, la demanda en horas pico [2]. Esto implica que se despliegue un gran número de estaciones base para atender estados de tráfico altos, sin embargo, este mismo número de estaciones base se mantiene aun cuando se reduzca la intensidad del tráfico. Lo anterior constituye un crecimiento en los niveles de consumo de energía de los operadores, elevando su gasto operativo. Así mismo, el alto consumo de energía resulta en que las redes celulares contribuyan con el 0.2% de emisiones de carbono del total del 2% que la industria de las Tecnologías de la Información y Comunicaciones (TIC) emiten a la atmósfera [3]. Ante estas problemáticas, una solución propuesta son las llamadas redes verdes que como su nombre lo sugiere son redes celulares amigables con el medio ambiente. Basadas en el hecho de que la estación base (BS) es el elemento que más energía consume (aproximadamente el 57% de la energía total consumida por una red celular), los esfuerzos se han concentrado en técnicas para mejorar su eficiencia energética manteniendo los requerimientos de Calidad de Servicio (QoS) de los usuarios móviles [3]. Estas técnicas se centran en dos aspectos: 1) soluciones de hardware para mejorar el consumo de energía de los componentes de la BS (amplificadores de potencia, procesadores de señales digitales, sistemas de enfriamiento, cables alimentadores); y 2) administración inteligente de los elementos de la red basada en las variaciones de carga de tráfico. La técnica de apagado y encendido de las BS pertenecen a esta última categoría, logrando reducir entre 12-40% del consumo de energía en una red celular [4].

En la literatura se han reportado diversos trabajos que proponen enfoques para activar y desactivar las BS. Por ejemplo, en [2] se propone una planeación de redes verdes LTE (Long-Term Evolution) para minimizar las BS activas y definir los patrones de apagado basándose en el cambio de las condiciones de tráfico en la red. Los aspectos que considera son la cobertura de las BS, el número máximo de usuarios móviles a los que una BS puede servir y la probabilidad de interrupción. Para redes pequeñas y medianas, el problema es transformado a uno de programación lineal entera y se resuelve con un algoritmo de ramificación y poda. Para redes de mayor tamaño se propone una solución heurística. Es conocido que la eficiencia de los algoritmos de ramificación y poda depende de los algoritmos usados para cada uno de estos procesos, para evitar una repetida ramificación, sin poda.

En [5], bajo la premisa de que apagar las BS en un orden especifico y no precisamente comenzando por las BS de menor carga permite que un menor número de BS permanezcan activas, los autores proponen un enfoque para minimizar el número

de BS activas. Presenta cuatro restricciones para la desactivación, que son: ancho de banda de la BS, BS desactivadas, cobertura y probabilidad de interrupción. Se aplica un algoritmo genético y se realiza la reasignación de usuarios móviles, aunque no está explícitamente indicado como se lleva a cabo.

De igual forma, en [6] se aplica un algoritmo de sociedad de arañas para resolver el problema de apagado y encendido de las BS en el que se minimiza el número de las BS activas. Para apagar una BS, el algoritmo penaliza de acuerdo a las BS vecinas disponibles. Si las BS vecinas a las que se desea apagar tienen recursos disponibles para recibir a más usuarios móviles el valor de aptitud de esta BS será menor, en caso contrario, será superior. Al disminuir la aptitud la solución será considerada más adecuada para el problema. El algoritmo de sociedad de arañas se modificó para resolver el problema en espacio discreto, ya que originalmente está diseñado para resolver problemas de optimización en espacio continuo. En [7] se argumenta que transformar un algoritmo de optimización que ha sido diseñado para resolver problemas en espacio continuo, a uno en espacio discreto, afecta su desempeño.

Así también en [8] se presenta un esquema que considera la operación dinámica de las BS, así como el problema de asociación de usuarios. En este trabajo se modela el problema de operación de BS dinámicas (apagar o encender una BS considerando el tráfico de carga) como una función de aptitud que minimiza el consumo de energía y el nivel de flujo de datos. Se derivan dos problemas: 1) La asignación de los usuarios móviles, para lo que se debe definir una política que garantiza que los usuarios móviles se asocien con las BS de una manera energéticamente eficiente, teniendo en cuenta el equilibrio de carga; y, 2) El apagado de las BS que se resuelve mediante un algoritmo voraz. Sin embargo, un algoritmo voraz analiza todas las posibles soluciones al problema lo que es muy costoso computacionalmente.

Un algoritmo que decide el apagado y encendido de las BS en una red heterogénea (red celular y red inalámbrica de área local) se propone en [9]. Se define una función de costo que minimiza el consumo de energía permitiendo un mayor número de accesos a la red. Aplican dos algoritmos codiciosos: el primero basado en la función de costo y el segundo basado en la densidad de puntos de acceso dentro de la cobertura de cada BS (apaga las BS con mayor número de usuarios móviles asociados). No obstante, el uso de algoritmos codiciosos tiene un costo computacional muy alto [10].

En [3] se propone redimensionar una red verde LTE, determinando el número mínimo de BS activas dada una determinada carga de tráfico con restricciones de QoS. La reducción de energía se representa por el número de las BS activas. Se aplica un algoritmo genético con elitismo para resolver el problema. El elitismo en algunos problemas lleva a una convergencia prematura, así como a un desempeño pobre en problemas que tienen una alta dimensionalidad.

Por último, el trabajo presentado en [11] reduce el consumo de energía en una red celular mediante un algoritmo de apagado de BS estableciendo que las BS se desactiven una a la vez, puesto que así se afecta mínimamente la carga de las demás BS; cada que se apague una BS se evaluará el incremento de carga en las BS vecinas. Para ello, toma en cuenta el tipo de región (urbana, metropolitana, etc.), ubicación de las BS y su cobertura. Se propone un algoritmo secuencial llamado SWES para reducir el tráfico

de señalización evaluando el impacto que tendrá en la red el apagado de determinada BS.

En este trabajo se propone un algoritmo para reducir el consumo de energía en una red verde LTE mediante el apagado y encendido de las BS. La mínima cantidad de BS activas estarán en función de la carga de tráfico actual (número de usuarios móviles activos). Es un problema de optimización NP-duro [2] cuyo objetivo es minimizar el número de estaciones bases activas para reducir el consumo de energía. La motivación para usar este enfoque viene de las fluctuaciones de tráfico que se dan a lo largo de un día, donde cargas bajas de tráfico no necesitan mantener todas las estaciones bases activas. El algoritmo propuesto considera la reasignación de usuarios una vez que una estación base es desactivada.

Los recursos de las BS están disponibles en un instante de tiempo, entonces el algoritmo evalúa el escenario de red en ese momento, desactivando algunas BS. Al desactivar una BS, busca reasignar sus usuarios móviles a las BS vecinas. Enseguida, analiza si con la selección de BS activas y reasignación es suficiente para atender al 99% de los usuarios móviles presentes en ese instante de tiempo. Con ello se garantizará que con las BS que aún permanecen activas se seguirán ofreciendo los servicios a los usuarios móviles.

Para la búsqueda de una solución se utiliza un algoritmo genético (AG) con modelo de población estacionario [12]. Un AG permite tratar instancias del problema de gran tamaño, encontrando soluciones satisfactorias en un tiempo razonable [10]. Durante el proceso de búsqueda de una solución, un AG imita el proceso natural de supervivencia del más fuerte. Evoluciona una población de soluciones aplicando iterativamente operaciones de selección, cruzamiento y mutación. Una vez que el proceso iterativo concluye, el individuo más apto es la solución del problema. A diferencia de un AG con modelo de población generacional donde la población de individuos es reemplazada por los hijos en cada generación, un AG con modelo de población estacionario remplaza solo una parte de la población de individuos. Esto reduce el tiempo de búsqueda de una solución.

Aunque algunas de las propuestas que se describieron anteriormente tratan indirectamente la complejidad del problema transformándolo en otro cuya solución óptima es menos difícil de encontrar, la presente propuesta lo trata directamente usando un AG. Por otro lado, los AG están diseñados para trabajar en espacios discretos, por lo cual tampoco es necesaria alguna transformación en su proceso de búsqueda de una solución. La implementación del AG con modelo de población estacionario, reduce el tiempo de búsqueda de una solución ya que no procesa a toda la población sino a una fracción de ella. En el modelo de optimización se considera explícitamente la reasignación de usuarios, ya que al desactivar una BS se debe observar la afectación que esto traería para sus usuarios móviles asociados. La reasignación de usuarios introduce mayor complejidad al problema, ya que no solo el algoritmo propuesto debe decidir qué conjunto de BS se deben desactivar sino también que usuarios móviles se deben asociar a cada BS que permanece activa. Esto con el fin, de que con las BS que permanezcan activas se sigan ofreciendo el mayor número de servicios a los usuarios móviles activos. Al igual que el trabajo en [3], la reducción de energía se representa por el número de las BS activas.

El artículo está organizado de la siguiente forma: la Sección 2 define el escenario de simulación y el modelo de optimización. La Sección 3 describe el algoritmo genético para la desactivación de las BS y reasignación de usuarios móviles. La Sección 4 muestra los experimentos realizados y los resultados obtenidos. Finalmente, la Sección 5 concluye el presente trabajo.

2. Formulación del problema

Un AG inicia con una población que es un conjunto de soluciones candidatas (individuos) al problema, las cuales se representan como una cadena (binaria, entera o flotante). A partir de algunos individuos de la población (padres) se crean nuevas soluciones (hijos). El objetivo de generar nuevos individuos es eventualmente encontrar mejores soluciones al problema. Cuando se aplica un AG se debe definir una función objetivo, que es una función matemática con la que se evalúa a cada solución candidata (individuo) para determinar qué tan apta es para resolver el problema. Otro elemento del AG son las restricciones que establecen si una solución es factible para el problema que se aborda [13].

En un AG que usa un modelo de población estacionario [13], se crean n nuevos individuos, donde n es menor al tamaño de la población. Dichos individuos son insertados a la población en el lugar de sus padres. En el AG de modelo de población estacionario que se usa en este trabajo, se crean dos hijos a partir de dos padres y se hace una comparativa de las aptitudes de los padres e hijos para que los dos individuos más aptos de los cuatro evaluados tomen el lugar de los padres [12].

En la Figura 1 se muestra el escenario de evaluación para este trabajo. Se observa que la red celular está compuesta por un número de estaciones base y de usuarios móviles desplegados sobre un área bidimensional con coordenadas cartesianas aleatorias que siguen una distribución uniforme para cada estación base (BS_j) y cada usuario móvil (UT_i). Para diferenciar la ubicación de estos dos componentes, una BS_j usa la notación (x_j, y_j), por otra parte, un UT_i usa (u_i, v_i). El total de estaciones base y usuarios móviles en la red en un instante de tiempo se denotan con la letra J e I respectivamente. Las BS pueden ser de dos tipos macro-celdas y femto-celdas. Los usuarios móviles asignados a una BS están delimitados dentro de su radio de cobertura D (en la Figura 1, la BS y sus respectivos usuarios móviles son del mismo color). El radio de cobertura de una femto-celdas siempre será menor que el radio de cobertura de una macro-celda. Cuando una BS se apaga (ver BS_4) y alguno de sus usuarios móviles no se pudo reasignar a una nueva BS este se considera un usuario móvil sin servicio (ver UT_{10}).

La distancia euclidiana entre una BS_j y un UT_i se denota como $d_{j,i}$ y se calcula aplicando la ecuación (1):

$$d_{j,i} = \sqrt{(u_i - x_j)^2 + (v_i - y_j)^2}. \quad (1)$$

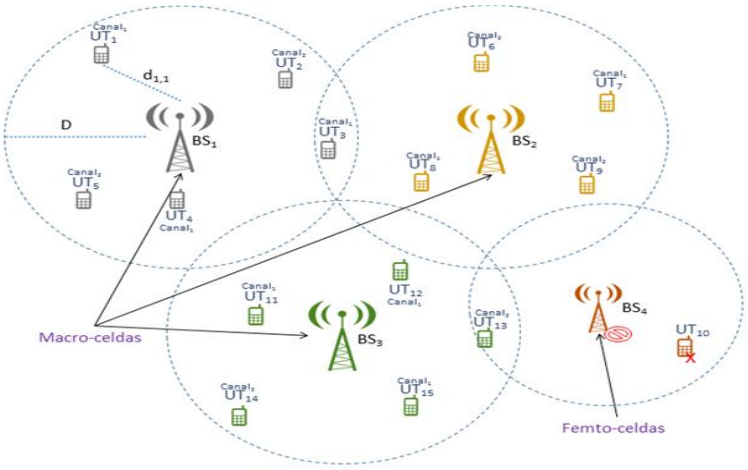


Fig. 1. Escenario de evaluación.

Cada BS_j brinda servicio a varios UT_i simultáneamente, para conocer esta relación se construyó la tabla *RBU* (Relación Bases Usuarios) donde las filas representan a la BS_j y las columnas a los UT_i. Si $RBU_{i,j} = 1$, la BS_j da servicio al usuario UT_i; en caso contrario, no hay relación entre el UT_i y la BS_j. Tomando como referencia el escenario de la Figura 1, se construyó la Tabla *RBU* mostrada en la Tabla 1. De esta forma, es posible observar cuando una *BS* está apagada, como en el caso de la BS_4 (fila 4) pues solo hay ceros en sus casillas. También es posible establecer cuales *UT* no están asociados a alguna *BS*, como es el caso del UT_{10} el cual solo tiene valores iguales a cero en cada BS_j.

Tabla 1. *RBU*.

Estaciones Base \ Usuarios	UT_1	UT_2	UT_3	UT_4	UT_5	UT_6	UT_7	UT_8	UT_9	UT_{10}	UT_{11}	UT_{12}	UT_{13}	UT_{14}	UT_{15}
BS_1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
BS_2	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
BS_3	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
BS_4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Una BS solo puede asignar una cantidad de canales *C* y dar servicio a un determinado número de usuarios móviles. *MUT* es el número máximo de usuarios móviles a los que puede servir una BS. Las macro-celdas pueden servir a más usuarios móviles que las femto-celdas.

El vector binario *SBS* representa a una solución candidata (individuo) del AG. Su longitud será igual al valor de *J*. La BS_j está encendida si el elemento SBS_j tiene un valor de 1 y apagada en caso contrario. En la Figura 2 se muestran algunos ejemplos de individuos y los escenarios a los que hacen referencia. Por otro lado, el vector *CU* de

longitud igual a *I* contiene el identificador del canal que cada UT_i tiene asignado por la BS que le da servicio. Los elementos en *CU* podrán tomar un valor desde 1 hasta *C*.

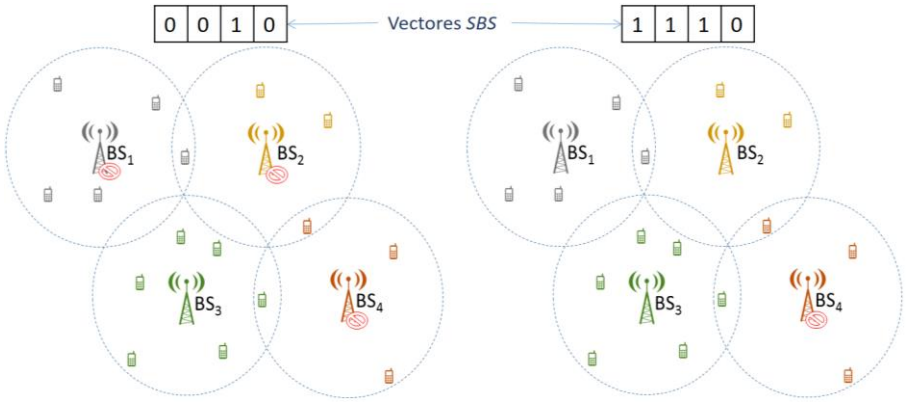


Fig. 2. Ejemplos del vector *SBS* y sus escenarios correspondientes.

Con el fin de obtener el número mínimo de *BS* activas que den servicio a la mayoría de los usuarios móviles en la red, se define la siguiente función objetivo:

$$Minimizar \sum_{j=1}^{J} SBS_j \,. \tag{2}$$

Como se muestra en (2), la solución es un vector *SBS* que determine el menor número de BS encendidas. Por otro lado, para que una solución se considere factible al problema, debe de cumplir las siguientes restricciones:

$$\sum_{j=1}^{J} RBU_{\,j,i} \leq 1, \qquad \forall i, \tag{3}$$

$$On_j = 1, \tag{4}$$

$$d_{j,i} \leq D, \forall i, \tag{5}$$

$$\sum_{i=1}^{I} RBU_{\,j,i} \leq MUT_j, \quad \forall j, \tag{6}$$

$$SNR_{j,i} \geq \alpha, \qquad \forall i, \tag{7}$$

$$PI_k \leq (1-\beta)J, \qquad \forall k. \tag{8}$$

La restricción (3) limita a que un UT_i sea servido por una sola BS_j. En la restricción (4) se establece que solo las BS encendidas podrán brindar servicio a los UT; esto es $ON_j = 1$ si $\sum_{i=1}^{I} RBU > 0$ *y* $SBS_j = 1$, en caso contrario, $ON_j = 0$. La distancia entre una BS_j y un UT_i al que se brinde servicio se limita en la restricción (5), donde esta no puede

ser mayor al umbral de radio de cobertura *D* de la BS en análisis. En (6) se obliga a una BS a no superar el máximo número de los usuarios móviles a los que puede dar servicio. La interferencia percibida por un UT_i se representa con el valor de la relación señal a interferencia (*SNR*) y debe ser mayor o igual a un umbral α como se establece en (7). Por último, el porcentaje de los usuarios móviles sin servicio al apagar algunas BS debe ser menor al umbral β cumpliendo con (8), este porcentaje representa la probabilidad de interrupción *PI*.

3. Algoritmo genético para desactivación de las BS y reasignación de usuarios móviles

El procedimiento basado en un AG para desactivar las BS considerando la reasignación de usuarios móviles se muestra a continuación:

```
INICIO
1.CONSTRUIR escenario inicial;
2.INICIALIZAR población con soluciones candidatas aleatorias;
3.EVALUAR cada solución candidata;
4. REPETIR HASTA (CONDICIÓN DE PARO sea satisfecha) HACER
  5.SELECCIONAR padres;
  6.CRUZAR pareja de padres;
  7.MUTAR los hijos resultantes;
  8.EVALUAR hijos;
  9.SELECCIONAR los dos mejores individuos;
10. FIN REPETIR HASTA
11.BUSCAR mejor individuo de la población;
FIN
```

Para construir el escenario inicial (PASO 1), se despliegan aleatoriamente las *BS* y los usuarios móviles dentro del área de cobertura. Enseguida, se especifica la cantidad de macro-celdas y femto-celdas que se incluirán la red. Se toma a cada uno de los usuarios móviles para ser asignados, un UT_i es asignado a una BS_j de acuerdo a: (1) si el total de los usuarios móviles a los que en ese momento da servicio la BS_j es menor a *MUT* y (2) si la distancia $d_{j,i}$ es menor o igual a *D*. Para elegir el canal asignado a cada UT_i se genera un número aleatorio en el rango de [1, *C*] y este valor se guarda en el vector *CU* en la posición *i*.

El tamaño de la población se denota con *NS*. Se inicializa aleatoriamente al *k-ésimo* vector *SBS* con valores binarios (PASO 2).

En el PASO 3 se evalúa al *k-ésimo* vector *SBS* y su aptitud se representa por A_k. Con base al escenario inicial, se reasignan los usuarios móviles que estuvieron asociados a las BS apagadas en cada solución. Para ello, se realizan las siguientes acciones:

1. Identificar las BS_j apagadas, es decir, los elementos del vector *SBS* donde $SBS_j = 0$.
2. Reasignar a los usuarios móviles asociados a una BS apagada a una BS activa, cumpliendo con las condiciones mostradas en las ecuaciones (3) – (7).

3. Calcular PI_k aplicando la ecuación (8). PI_k es el porcentaje de los usuarios móviles que no fueron reasignados a una BS activa o aquellos usuarios móviles sin servicio en la solución *k*. Un UT_i se considera sin servicio si todas las celdas de la columna *i* en la tabla *RBU* tienen un valor igual a 0.
4. Evaluar (2), contando los elementos del vector *SBS* donde $SBS_j = 1$, es decir, las BS_j activas.

Una solución puede contar con todas las BS encendidas o ninguna, sin embargo, soluciones con pocas BS encendidas o ninguna dejaría a muchos usuarios sin servicio, lo que se traduce en un mayor valor de *PI*. Por tanto, si $PI_k > \beta$, la solución *k* será penalizada, es decir, su valor de A_k se incrementará para que al comparar su aptitud con la aptitud de una solución factible esta quede en desventaja (las soluciones con menor aptitud son elegidas como padres). Una solución *k* factible puede tener un valor máximo de $A_k=J$ (todas las *BS* activas) por tanto a las soluciones no factibles o penalizadas se les sumará (J+1) a su aptitud, con el fin de marcar una diferencia clara entre las soluciones factibles y no factibles. De esta manera se logra que el AG obtenga soluciones con pocas BS activas y que también cumplan con el umbral de *PI* definido.

Cada solución *k* cuenta con una tabla *RBU* con la asignación de usuarios móviles a las *BS* activas, los UT_i sin servicio (si existen) y las BS_j apagadas. Para calcular el valor de SNR_j, al reasignar un UT_i a una BS_j, se identifica el canal menos utilizado dentro de la BS_j, y se aplica la ecuación (9):

$$SNR_{j,i} = PR_{j,i} - IT_{j,i}, \tag{9}$$

$$PR_{j,i} = PT - PL_{j,i}, \tag{10}$$

$$PL_{j,i} = A + B\log\left(d_{j,i}\right) + N, \tag{11}$$

donde $PR_{j,i}$ es la potencia recibida en el UT_i. $IT_{j,i}$ es la interferencia total de los usuarios móviles que usan el mismo canal en la BS_j. En la ecuación (10), *PT* es la potencia de transmisión de las BS. El valor de pérdidas por trayectoria $PL_{j,i}$ se determina a partir de (11) donde *A* y *B* son valores constantes, *N* es un valor aleatorio distribución Gaussiana con una varianza de σ^2 que hace referencia al tipo de área donde se ubica la red (es mayor para una metrópolis y menor en espacios abiertos) [14]. En el presente trabajo se consideraron los siguientes valores $A = 50$, $B = 40$, y $\sigma^2 = 10$.

Seleccionar a los padres por la técnica de torneo [12] es el PASO 5 del algoritmo. El número de participantes en el torneo es de dos individuos.

Enseguida en el PASO 6 se genera un número decimal aleatorio en un rango de [0,1] el cual es comparado con la probabilidad de cruzamiento *PC*. Si dicho número es menor o igual a *PC* se procede a generar dos nuevos individuos con una combinación de los bits o elementos del vector *SBS* de ambos padres. De manera más específica se utiliza el cruce por dos puntos [13].

Para el PASO 7 algunos bits de los hijos se mutan, un bit mutado toma su valor inverso (1 se convierte en 0 y 0 se convierte en 1). Para decidir que bits se mutaran, por cada elemento de los nuevos vectores *SBS* se genera un número decimal aleatorio en

un rango de [0,1]. Si este valor es menor a la probabilidad de mutación (*PrM)* el bit es cambiado.

Una vez que se han mutado los hijos, en el PASO 8 se procede a evaluar estas soluciones o vectores *SBS*. Esto es, se realiza la reasignación y cálculo del valor de A_k para los dos nuevos individuos aplicando las cuatro acciones mencionadas en el PASO 3.

En el PASO 9 se compara el valor de A_k de los dos padres y de los dos hijos. Los dos individuos con mejor valor de aptitud A_k se insertan en la población en el lugar que ocupaban los padres.

En el PASO 10, se considera un ciclo al proceso de seleccionar a los padres, cruzarlos, mutar a los nuevos individuos y remplazar a los padres con los mejores individuos.

Existen distintas condiciones de paro, por ejemplo, si en el problema se conoce la solución óptima se puede forzar al algoritmo a realizar los ciclos que sean necesarios para encontrar dicha solución o una muy próxima a ella. En el caso de un AG con modelo de población estacionario, una condición de paro puede ser realizar los ciclos necesarios para que todos los individuos de la población se reemplacen por lo menos una vez por sus hijos. Debido a que no se conoce la solución óptima en el problema planteado y el cambiar a todos los individuos podría requerir de un número demasiado grande de ciclos, la condición de paro del algoritmo mencionada en el PASO 4 es realizar un determinado número de ciclos.

Por último, en el PASO 11 se busca en la población la solución con menor valor de A_k. En caso de que existan dos o más soluciones con el menor número de *BS* encendidas, se selecciona a la solución con menor probabilidad de interrupción *PI*.

4. Resultados

Se realizaron una serie de experimentos considerando los siguientes casos (i) una red homogénea (solo macro-celdas) y (ii) una red heterogénea (macro-celdas y femto-celdas). Este último caso se incluyó ya que como se ha reportado en [4], las femto-celdas sirven como soporte a las macro-celdas en estados de tráfico altos y se desactivan en caso contrario.

Tabla 2. Parámetros de simulación.

Descripción	**Valor**
Potencia de transmisión	-40 dB
Umbral de la relación señal interferencia (α)	3 dB
Área de cobertura	25 km2
Radio de cobertura (*D*) de las BS macro/femto	2 km / 1 km
Número de canales por BS	8 canales
Umbral de la probabilidad de interrupción (β)	1 %
Número máximo de usuarios móviles por macro-celda	150
Número máximo de usuarios móviles por femto- celda	75

Se simula una red con 20 BS, divididas en 10 macro-celdas y 10 femto-celdas en redes heterogéneas. En todos los casos, se fue incrementando la cantidad de usuarios móviles activos en la red. Los parámetros de simulación considerados para la evaluación se muestran en la Tabla 2.

Tabla 3. Descripción de los experimentos.

Experimento	Número de usuarios móviles	Número de macro-celdas	Número de femto-celdas
Red Homogénea			
1	500	20	-
2	200	20	-
3	100	20	-
Red Heterogénea			
4	500	10	10
5	200	10	10
6	100	10	10

Para el AG se consideró una población de 50 individuos con una PC=0.6 y PrM=0.001, además se realizan 2000 ciclos por cada ejecución del algoritmo. En la Tabla 3 se especifican las características de los 6 experimentos realizados. Cada experimento consta de 100 ejecuciones, en cada ejecución la posición de los usuarios móviles se modificó aleatoriamente y se obtuvo una solución *k*.

En los seis experimentos se observó que se reasignan los usuarios móviles de las BS apagadas a sus BS vecinas, y en la mayoría de las ejecuciones se mantiene con servicio al 99% de los usuarios móviles presentes en la red. De igual forma el conjunto de las BS encendidas está distribuido en el área de cobertura de la red, es decir, las BS activas en las soluciones ofrecidas por el algoritmo no se encuentran en una sola región del área de cobertura o demasiado cerca una de otra. En el experimento 4 se mantuvieron encendidas BS muy cercanas solo cuando son heterogéneas, es decir, una femto-celda y una macro-celda.

En el experimento 1 se observó que, si una sola BS cubre una determinada área y no existen BS vecinas que puedan dar servicio a la mayoría de sus usuarios móviles, esta BS se mantiene encendida. Las BS encendidas no se seleccionan solo por el número de usuarios móviles a los que están sirviendo, sino por el número de usuarios móviles a los que puedan recibir de sus BS vecinas. Es decir, un BS se mantiene encendida no solo por tener muchos usuarios móviles relacionados en el escenario inicial, más bien se mantiene encendida porque al realizar la reasignación se logró incrementar considerablemente este valor.

En el experimento 4 se generó una mayor cantidad de tráfico en la red (mayor número de usuarios móviles). Se observa que una femto-celda da servicio a los usuarios móviles que se encuentran dentro del radio de cobertura de una macro-celda, es decir, una femto-celda da servicio a algunos usuarios móviles de la macro-celda.

En la Tabla 4, se reporta la solución con mejor aptitud obtenida durante las 100 ejecuciones de cada experimento. Así mismo, se reporta la peor solución obtenida y el promedio de las 100 ejecuciones. Para todos los experimentos realizados se observó

que en promedio el número de BS activas se redujo considerablemente como se observa en la Tabla 4, de las 20 BS desplegadas inicialmente, solo se mantuvieron encendidas 10 BS o menos.

Tabla 4. BS encendidas para la mejor y peor corrida de cada experimento.

Experimento	BS encendidas/*PI* (mejor ejecución)	BS encendidas (peor ejecución)	Promedio de BS encendidas
1	6 BS / 0.6 %	14 BS	9
2	5 BS / 1.0 %	11 BS	7
3	5 BS / 0.0 %	11 BS	7
4	8 BS / 0.2 %	15 BS	10
5	6 BS / 0.5 %	13 BS	8
6	5 BS / 1.0 %	14 BS	8

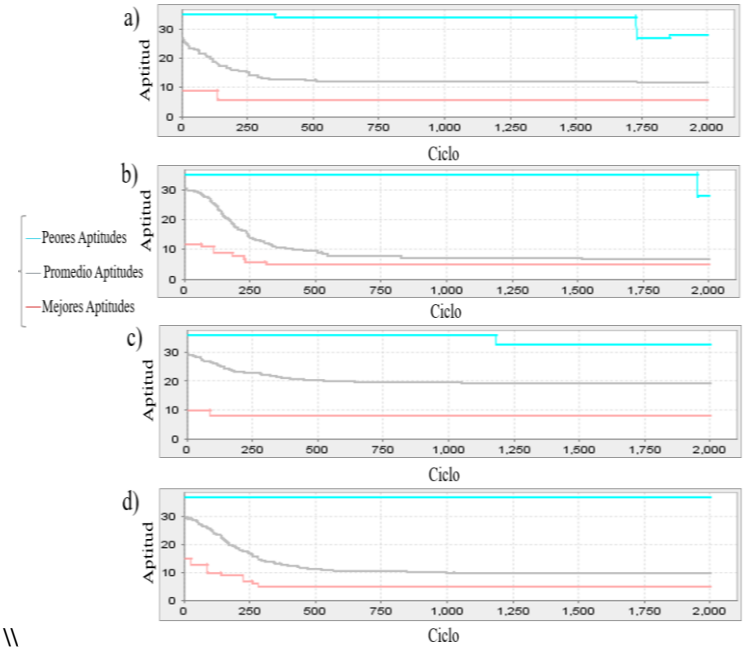


Fig. 3. Convergencia del algoritmo: a) experimento 1, b) experimento 3, c) experimento 4, d) experimento 6.

Al disminuir el número de usuarios móviles también se reduce el número de BS activas para ambos tipos redes. Por ejemplo, en redes homogéneas al reducir de 500 a 200 usuarios móviles, el número de BS encendidas disminuye de 6 a 5. En el caso de redes heterogéneas se reduce de 8 a 5 BS encendidas (ver Tabla 4). En el experimento 5 se logran apagar más BS en la mejor corrida que en el experimento 2, puesto que la

mayoría de las BS apagadas eran femto-celdas. Entonces el algoritmo decide apagar una BS con menores capacidades (radio de cobertura y *MUT*).

Con respecto al número de usuarios móviles reasignados no se observa diferencias significativas al comparar los dos tipos de redes. Sin embargo, cuando hay 100 usuarios móviles en la red, en promedio se resignan a 36 de ellos; por otra parte 75 de los 200 usuarios móviles presentes en los experimentos 2 y 5 fueron reasignados; y para los experimentos con 500 usuarios móviles en promedio se reasignaron 165 usuarios móviles. Entonces al apagar las BS se reasigna a casi el 50% de los usuarios móviles presentes en la red.

En el experimento 4, de las 12 BS apagadas, 8 eran femto-celdas, mientras que para el experimento 5 todas las femto-celdas fueron apagadas. Entonces para redes heterogéneas cuando el número de usuarios móviles disminuye, las BS apagadas son femto-celdas.

En el 100% de las ejecuciones del algoritmo en redes homogéneas se obtienen soluciones donde se cumple con $PI \leq 1\%$, en cambio, en redes heterogéneas solo en un 70-90% de las ejecuciones se obtiene soluciones que cumplen dicha relación. Entonces es más complejo realizar el apagado de BS en redes heterogéneas.

En la Figura 3 se muestra la convergencia de la mejor solución encontrada en los experimentos 1, 3, 4 y 6. Además se muestra el promedio y la peor aptitud por ciclo. Se observa que conforme aumenta el número de usuarios móviles es más difícil para el algoritmo evolucionar o converger. Esto se debe a que el número de soluciones válidas disminuye dado que las soluciones están estrechamente ligadas al número de usuarios móviles presentes en el área.

5. Conclusiones y trabajo futuro

En el presente trabajo se abordó el problema de desactivación de BS implementando un algoritmo genético con modelo de población estacionario. Esto con el objetivo de minimizar el número de BS encendidas considerando la reasignación de usuarios móviles, para redes homogéneas y heterogéneas.

Conforme se incrementa la cantidad de usuarios móviles en la red, es más complejo encontrar una solución. Lo anterior se debe al proceso de reasignación de usuarios móviles incluido en este trabajo. Por este motivo, es justificable utilizar la técnica de AG con población estacionaria, ya que en cada ciclo de evaluación solo considera a dos padres y a dos hijos, a diferencia de un AG con población generacional.

Los experimentos realizados con diversa densidad de tráfico (número de usuarios móviles activos) y distintos tipos de red muestran que:

- Se puede apagar hasta un 50% del número de BS encendidas cuando los estados de tráfico son bajos tanto en redes homogéneas como heterogéneas.
- Después de apagar las BS se reasignaron entre 33-38% de los usuarios móviles presentes en la red, por este motivo la reasignación de usuarios móviles debe ser considerado un factor de impacto en soluciones que involucren desactivación de BS en redes celulares.

- Es posible dar servicio a casi el 100% de los usuarios móviles aun si se apagan ciertas BS de una red.
- En redes heterogéneas el algoritmo decide desactivar primero a las BS femto-celdas cuando el tráfico es menor, lo que coincide con lo reportado en la literatura, en la que se menciona que su uso es más beneficioso cuando la red macro-celular presenta estados de tráfico alto.

Como trabajo futuro se plantea un análisis de sensibilidad con el umbral de la relación señal-interferencia para estimar hasta qué punto se puede garantizar QoS para los usuarios móviles.

Referencias

1. Cisco VNI Complete Forecast. Disponible en: http://www.cisco.com/c/dam/assets/sol/service-provider/vni-complete-forecast/vnisp.html
2. El-Beaino, W., Al-Kanj, L., El-Hajj, A. M., Dawy, Z.: Optimized joint cell planning and BS on/off switching for LTE networks. 12(16), pp. 1537–1555 (2015)
3. Azzam, S. M., Elshabrawy, T.: Re-Dimensioning Number of Active eNodeBs for Green LTE Networks Using Genetic Algorithms. In: Proceedings of European Wireless 2015, 21th European Wireless Conference, pp. 1–6 (2015)
4. Alsharif, M. H., Nordin, R., Ismail, M.: Survey of Green Radio Communications Networks: Techniques and Recent Advances. J. Comput. Netw. Commun. 2013, e453893, doi:10.1155/2013/453893 (2013)
5. Akram, B. S., Furkan, A., Halim, Y.: A genetic algorithm based cell switch-off scheme for energy saving in dense cell deployments. IEEE Globecom Workshops (2012)
6. James, J. Q., Yu, V. O. K. L.: Base Station Switching Problem for Green Cellular Networks with Social Spider Algorithm. In: IEEE Congress on Evolutionary Computation (CEC), Beijing (2014)
7. Pampara, G.: Angle modulated population based algorithms to solve binary problems. Disponible en: http://repository.up.ac.za/handle/2263/22801 (2012)
8. Kyuho, S., Hongseok, K., Bhaskar, K., Yung, Y.: Base Station Operation and User Association Mechanisms for Energy-Delay Tradeoffs in Green Cellular Networks. 29, 1525–1536 (2011)
9. Seonwook, K., Sunghyun, C., Byeong, G. L.: A Joint Algorithm for Base Station Operation and User Association in Heterogeneous Networks. IEEE Communications Letters, 17, pp. 1552–1555 (2013)
10. Talbi, E. G.: Metaheuristics: From Design to Implementation. John Wiley & Sons (2009)
11. Oh, E., Son, K., Krishnamachari, B.: Dynamic Base Station Switching-On/Off Strategies for Green Cellular Networks. IEEE Trans. Wirel. Commun. 12, pp. 2126–2136 (2013)
12. Neal-Holtschulte, M. M.: Should every man be an island? In: GECCO, Amsterdam, The Netherlands (2013)
13. Eiben, A. E., Smith, J. E.: Introduction to Evolutionary Computing. Springer Berlin Heidelberg (2015)
14. Stephen, B., Wicker, B. K.: Experimental analysis of local search algorithms for optimal base station local (2000)

ISSN 1870-4069

Classification of Cervical Cancer Using Assembled Algorithms in Microscopic Images of Papanicolaou

Obrayan H. Gómez[1], Eddy Sánchez-DelaCruz[1], A. Paulina de la Mata[2]

[1] Technological Institute of Misantla, Veracruz, Mexico

[2] Department of Chemistry, University of Alberta, Edmonton, Canada

esanchezd@itsm.edu.mx**

Abstract. In Mexico cervical cancer is the second leading cause of death from malignant neoplasms in women, but this mortality rate has been reduced in recent years thanks to early detection programs as the pap smear test, which is aimed at finding pre-cancerous abnormalities in cells that cover the cervix. The pap smear test is an efficient medical test, but it presents problems at the moment of interpretation under the microscope, due to the large number of cells in the sample and others external factors. In order to solve this disadvantage, computational techniques are used to support the samples classification. In this research we propose to use assembled algorithms to construct a classifier. The database used is from Herlev University Hospital, the data were formulated as a binary classification problem. The results of the experiments (exhaustive search) show that using the combinations of algorithms Bagging+MultilayerPerceptron and AdaBoostM1+LMT is obtained a high percentage of correctly classified instances, 95.74%.

Keywords: cervical cancer, pap test, classification, assembled algorithms.

1 Introduction

Cancer is a set of related diseases that have a common characteristic: abnormal cellular growth, where cells begin to divide without stopping and to invade adjacent tissues. Due to human body is made up of millions cells, cancer can appear anywhere in body [1]. In growth cycle, a cell born of division of stem cell, once cell gets old or is damaged, it dies, and another cell replaces it to start cycle again. However, cancer cells grow abnormally and survive to continue dividing.

** Corresponding author is E. Sánchez-DelaCruz.

Cervical cancer begins when cells that cover cervix start to grow uncontrollably. The cervix is composed of two parts: endocervix and exocervix covered by two different cell types, glandular and squamous, respectively. The area where these two parts join is called the transformation zone, most of the cervical cancer begins in cells of this zone.

Pap test is a method for early detection of cervical cancer (make it by experts), where samples of cervix cells are observed through microscope in order to find abnormalities cells. However an accurate analysis of the hundreds of thousands of cells in each sample is not possible for human eye. Each sample examined may contain about 300,000 cells. In the literature related to pap smear [2,3,4,7,10,13,17] we can identify that the classification of the samples has been approached from a computational perspective.

This article is organized as follows: the motivation of research is mentioned in section two, previous work are described in section three, section four refers to the methodology applied, the experiments and analysis of results are described in section five and six respectively, finally in section seven the conclusion and perspectives are presented.

2 Motivation

Two important points that motivated this research were: 1) According [16] in Mexico there is a high rate of cervical cancer in women, and it ranks second mortality place by cancer in the country. 2) Difficult task of analyzing pap smear samples and degree of uncertainty in determining the stage of sample.

2.1 Problematic

Globally, cervical cancer is one of the main causes of health problems in the population and the seventh most frequent neoplasm in the world population and the fourth most frequent among women with an estimated 528,000 new cases diagnosed annually. It is also an important cause of death for a malignant tumor in women with 266,000 annual deaths, 87% of which occur in underdeveloped countries [16]. In Mexico, there are also high rates of patients with this disease with annual occurrence of 13,960 cases, with an incidence of 23.3 cases per 100,000 women. The states with the highest mortality from cervical cancer are Morelos (18.6), Chiapas (17.2) and Veracruz (16.4) [16].

In most patients with cervical cancer, a high level of accuracy is not obtained when determining the stage of the disease. This problem has been approached from a computational perspective [2,3,4,7,10,13,17], however, based on the analysis of the state of the art, it has been identified that: the percentages of correct classification of stages of the disease can still be improved.

2.2 Proposed Solution

In this research, we propose to identify the phases of cervical cancer using assembled algorithms to minimize error margin, because with this approach

has been reached competitive percentages in breast cancer detection [5] and categorization of neurodegenerative diseases [15,14] . In section four it will be discuss more in detail of the methodology that is followed to solve the problem.

3 Previous Works

Jantzen et al., published in 2005 a dataset of Herlev University Hospital[c] which contain a total of 20 characteristics, this dataset was used for testing and comparing their own neural network with linear activation functions for minimized the error [7].

Cortes et al., propose a method of optimization by swarming of particles (OSP) for segmentation of images of microscopic papanicolaou tests and the identification of abnormal characteristics in the cells of the samples, as result a comparative table is shown where is obtained a better segmentation comparing against to the Darwinian DOSP and FODOSP models, with a classification rate of 77.5% [13].

Sharma et al., performed a classification of the stages of cancer in images of papanicolaou samples. The dataset used was collected from Fortis Hospital Mohali, Punjab, India. Images segmentation was made for detect contours and to detect nuclei and cytoplasm of cells, once isolated they extracted morphological characteristics as area, perimeter, extension, and nucleus ratio with cytoplasm. Subsequently, the KNN algorithm was used as classier, obtaining 84.3% of accuracy [17].

Bora et al., propose an intelligent system that automates the categorization of papanicolaou samples. The system was evaluated on two generated databases by Ayursundra Healthcare Pvt. Ltd. and Dr B. Borooah Cancer Institute, Guwahati, India, as well as the database built by Herlev University Hospital. For the classification they use assembled methods combining three individual classifiers selected: Multilayer Perceptron, Random Forest and Least Square Support Vector Machine (SVM). They obtained an accuracy of 98.11 % and an accuracy of 98.3 % at the extended level and a 99.11% at the cell level using generated databases, for the Herlev database, they obtained 96.51 % (2 classes) and 91.71 % (3 classes) accuracy [2].

Marinakis et al., propose a meta-heuristic algorithm to cells classification with dataset of Herlev University Hospital. They applied a genetic algorithm for the selection of the most representative numerical characteristics of the image. The genetic algorithm was combined with three classifiers based on the nearest neighbor: 1-nearest neighbor (1NN), k-nearest neighbor (KNN) and wk-nearest neighbor (WKNN). As a result they obtained a classification of 98.14 % with 1NN, 97.61% with KNN and 97.39 % with WKNN for 2 classes, for 7 classes they obtained a 96.95 % rank with 1NN, 96.73% with KNN and 96.94 % with WKNN [10].

Camargo et al., to discriminate normal cells from abnormal cells, propose a classification method based on global MPEG-7 color and texture descriptors:

[c] http://mde-lab.aegean.gr/downloads

Color Layout and Scalable Color. They use the Herlev University Hospital database, where, unlike the aforementioned methods that use the morphological characteristics of the cells, the proposed method uses the color space and texture information of the nucleus and cytoplasm. The classification algorithms used were KNN and SVM [3].

Chankong et al., worked in segmentation and classification of papanicolaou samples using three different datasets: ERUDIT, LCH and Herlev. For the images segmentation they used a fuzzy C-means algorithm. For classification were implemented a Bayesian classifier, discriminant linear analysis, KNN, artificial neural network (ANN) and SVM. The best-performing algorithm was ANN with 96.20% for four classes and 97.83% for two classes with ERUDIT dataset, 93.78% for seven classes and 99.2% for two classes with Herlev dataset, 95.00% for four classes and 97.00% for two classes with LCH dataset [4].

4 Materials and Methods

A classifier model is associated with pattern recognition [12]. A brief description of methodology is given below (see Fig. 1):

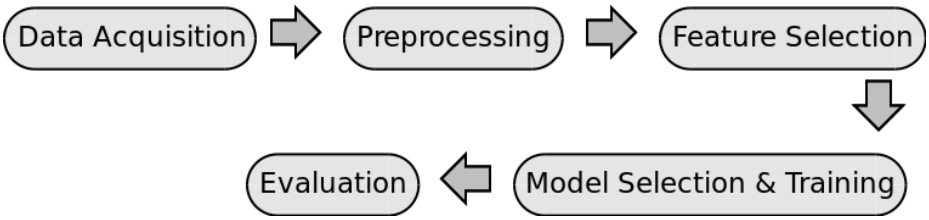


Fig. 1. Proposed methodology based on [12].

- Data acquisition. The pap smear database was constructed by Herlev University Hospital. It consists of a collection of 917 images, 242 normal cells and 675 abnormal cells, Table 1 shows a database description. Each image is described by 20 typical numerical characteristics in cellular measurements (see Table 4).
- Preprocessing. The purpose of this section is to apply a treatment to the data acquired which include: Images conversion to grayscale and transformation of the data to csv format.
- Attribute selection. The goal is to reduce the dimensionality of data, reduce the total number of features without losing important information. The proposed methods for characteristics reduction are: Correlation-Based Feature Selection (CFS), Chi Squared, Consistency, Information Gain and Symmetrical uncertainty.

Table 1. Contents of the database. Taken from [7].

Class	Category	Cell type	Cell count
1	Normal	Superficial squamous epithelial	74
2	Normal	Intermediate squamous epithelial	70
3	Normal	Columnar epithelial	98
			242
4	Abnormal	Mild squamous non-keratinizing dysplasia	182
5	Abnormal	Moderate squamous non-keratinizing dysplasia	146
6	Abnormal	Severe squamous non-keratinizing dysplasia	197
7	Abnormal	Squamous cell carcinoma in situ intermediate	150
			675
			917

- Classifier. There is a wide variety of Machine Learning algorithms, this are some categories available in WEKA software: Bayes, Functions, Trees, Lazy, Rules, Meta (assembled) and Miscellaneous.
- Evaluation. The dataset is splitted in two parts, one for training and second for test the classifier.

The last four methodology phases are described in detail in sections five and six.

5 Experiments

The experiments were performed in a computer with: OS Windows 10 Home Single Language, Intel(R) Core(TM) i3-5005U CPU 2.00 GHz 2.00 GHz, RAM 8.00 GB, HD 1.00 TB, 64-bit operating system and x64 processor. In addition, to execute the algorithms described in this section, we used the programming language R v.3.3.3 and the software WEKA v.3.8.

5.1 Attribute Selection

Based on [6],[8],[9],[11],[18], five filtering methods were applied for feature reduction:

- CFS: evaluates the value of an attributes set considering the ability to predict the class and the correlation between attributes. It seeks to maximize the correlation between classes and minimize the correlation between attributes.
- Chi Squared: this method is used to determine if there is a significant association between two variables.
- Consistency: the method looks for smallest subset of features that discriminate best from the original dataset.

Table 2. Summary of the 20 characteristics in the database. Based on [7].

Column	Feature	Name
1	Nucleus area	Narea
2	Cytoplasm area	Carea
3	N/C ratio	N/C
4	Nucleus brightness	Ncol
5	Cytoplasm brightness	Ccol
6	Nucleus shortest diameter	Nshort
7	Nucleus longest diameter	Nlong
8	Nucleus elongation	Nelong
9	Nucleus roundness	Nround
10	Cytoplasm shortest diameter	Cshort
11	Cytoplasm longest diameter	Clong
12	Cytoplasm elongation	Celong
13	Cytoplasm roundness	Cround
14	Nucleus perimeter	Nperim
15	Cytoplasm perimeter	Cperim
16	Nucleus position	Npos
17	Maxima in nucleus	Nmax
18	Minima in nucleus	Nmin
19	Maxima in cytoplasm	Cmax
20	Minima in cytoplasm	Cmin

– Information Gain: this method indicates the information amount of an attribute to predict the class when it is present or absent in dataset. That is, it measures the difference of information between cases in which the value of attribute is known and where the value is unknown.
– Symmetrical Uncertainty: this method calculates the correlation between two attributes, it can be said that it is information gain normalized.

In order by relevance Table 3 lists the attributes resulting from each applied filter. The CFS algorithm obtained only one relevant attribute, with the Consistency algorithm we obtained 19 attributes, the Chi Squared, Information Gain and Symmetrical Uncertainty algorithms were generated 20 representative attributes for each one.

The weight column was added to give numeric value to the occurrence of each attribute according position, the attribute best positioned have value of 20. Next, in order to obtain the total weight, for example, of KC attribute that was the best attribute in four cases and the third in last case, the weights were added ($20 + 20 + 18 + 20 + 20 = 98$). Table 4 shows the total weight of each attribute.

Figure 2 provides the calculated weight of each attribute. When divided the data average it is possible to be observed that excel seven attributes, in addition that best threes have a significant separation from each other, which leads us to suppose that these attributes are best to discriminate the classes.

Table 3. Attributes resulting from each filtering.

Weight	CFS	Chi Squared	Consistency	Information Gain	Symmetrical Uncertainty
20	K.C	K.C	KerneA	K.C	K.C
19		CytoA	CytoA	CytoA	CytoA
18		CytoMax	K.C	CytoMax	CytoMax
17		KernePeri	KerneYcol	CytoMin	CytoMin
16		CytoMin	CytoYcol	CytoShort	CytoShort
15		KerneLong	KerneShort	CytoPeri	CytoPeri
14		CytoShort	KerneLong	CytoLong	KerneA
13		CytoPeri	KerneElong	KerneA	KernePeri
12		KerneMax	KerneRund	KernePeri	KerneLong
11		KerneShort	CytoShort	KerneLong	CytoLong
10		KerneMin	CytoLong	KerneMax	KerneMax
9		CytoLong	CytoElong	KerneShort	KerneShort
8		CytoRund	CytoRund	KerneMin	KerneMin
7		KerneA	KernePeri	CytoRund	CytoRund
6		KerneYcol	CytoPeri	KerneYcol	KerneYcol
5		KerneRund	KernePos	KernePos	KerneRund
4		KernePos	KerneMin	KerneRund	KernePos
3		KerneElong	CytoMax	KerneElong	KerneElong
2		CytoYcol	CytoMin	CytoYcol	CytoYcol
1		CytoElong		CytoElong	CytoElong

5.2 Classification

Binary classification was contemplated: normal and abnormal (see Table 1). For training and test sets, criterion of 2/3 and 1/3 was used as shown in Table 5.

An exhaustive search was performed using the 20 attributes described in Table 4 to determine which meta-classifier generate the best results. The procedure to be followed was to compare each meta-classifier algorithm with the rest of algorithms belonging to each of the available categories.

6 Results and Analysis

After the algorithm assembly experiments by exhaustive search, we obtained the best results with Bagging+MultilayerPerceptron and AdaBoostM1+LMT, 95.74% for both. Confusion matrix for each one are showed in Table 6 and Table 7.

In confusion matrix of Bagging+MultilayerPerceptron classifier we can see that there is a confusion of 8 and 5 instances. On the other hand, 63 and 229 instances were correctly classified. For case of AdaBoostM1+LMT classifier we can see that there is a confusion of 9 and 4 instances. On the other hand, 62 and 230 instances were correctly classified. These results represents a high degree of reliability.

Table 4. Total weight of attributes in order of relevance in all filters.

Attribute	K.C	CytoA	CytoMax	CytoShort	KerneA
Weight	98	76	57	57	54
Attribute	KerneLong	CytoMin	KernePeri	CytoPeri	CytoLong
Weight	52	52	49	49	44
Attribute	KerneShort	KerneYcol	KerneMax	CytoRund	KerneMin
Weight	44	35	32	30	30
Attribute	KerneRund	CytoYcol	KerneElong	KernePos	CytoElong
Weight	26	22	22	18	12

Table 5. Training and test sets.

Class	Training	Test	Total
Normal	162	80	242
Abnormal	450	225	675
Total	612	305	917

Even though the percentages are identical, the confusion matrix of each of them indicates that Bagging+MultilayerPerceptron better classifies the data Normal class and AdaBoostM1+LMT better classifies the data of Abnormal class.

7 Conclusion and Future Work

After an exhaustive search was determined that classifiers Bagging+MultilayerPerceptron and AdaBoostM1+LMT were bests, both with 95.74% of correct classification using 20 attributes and treating the problem as a binary classification. Comparing these results against results obtained in previous works (that use the database constructed by Herlev University Hospital in binary classification) the classifiers proposed in this work far exceeded the results obtained in [13] that consists of 77.5% precision. However, [2] obtained 96.51%, [10] 98.14% and [4] 99.2%. It should be emphasized that the classifiers proposed in this article use 20 attributes of dataset, we suppose that with reducing characteristics number

Table 6. Confusion Matrix: Bagging+MultilayerPerceptron.

a	b	Classified as
63	8	a= Normal
5	229	b= Abnormal

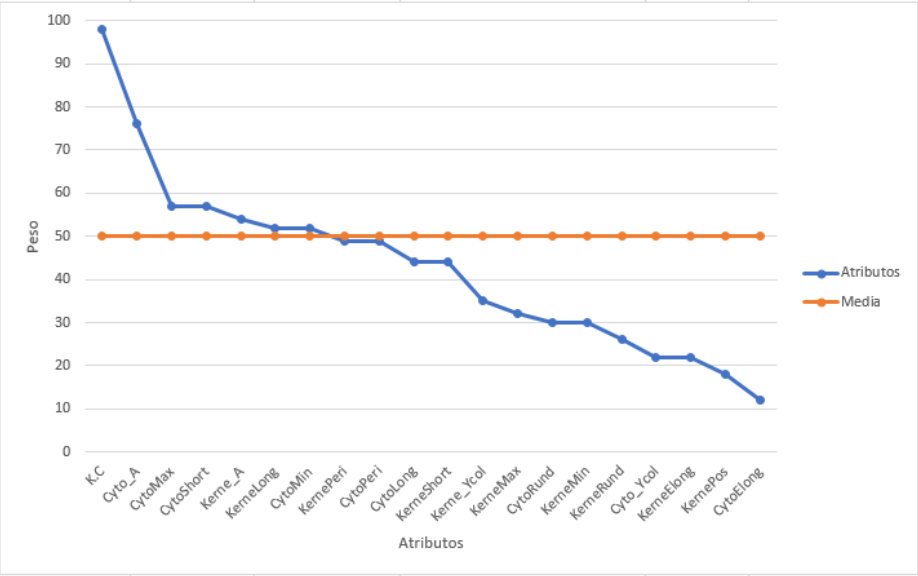


Fig. 2. Attributes with respect to the mean of the data.

Table 7. Confusion Matrix: AdaBoostM1+LMT.

a	b	Classified as
62	9	a= Normal
4	**230**	b= Abnormal

we will achieve competitive or superior percentages to the reported in previous work.

For the immediate continuation of the investigation it is intended:

- To apply the classifiers Bagging+MultilayerPerceptron and AdaBoostM1+LMT using seven attributes selected and to compare results with this study.
- Apply the same methodology of this study for a multi-class classification (with all seven classes), ie, 1) select best attributes and 2) to find ideal algorithm combination, firstly using all attributes and next using best attributes.

References

1. American Cancer Society: What is cervical cancer? https://www.cancer.org/cancer/cervical-cancer/about/what-is-cervical-cancer.html (2016)
2. Bora, K., Chowdhury, M., Mahanta, L.B., Kundu, M.K., Das, A.K.: Automated classification of pap smear images to detect cervical dysplasia. Computer Methods and Programs in Biomedicine 138, 31–47 (2017)

3. Camargo, L.H., Diaz, G., Romero, E.: Pap smear cell image classification using global mpeg-7 descriptors. Diagnostic Pathology 8(1), S38 (2013)
4. Chankong, T., Theera-Umpon, N., Auephanwiriyakul, S.: Automatic cervical cell segmentation and classification in pap smears. Computer Methods and Programs in Biomedicine 113(2), 539–556 (2014)
5. de la Cruz, E.S., Alpuín-Jiménez, H., Domínguez, H.d.J.O., Parra, P.P.: Sdca: System to detect cancerous abnormalities. In: LA-NMR. pp. 115–122 (2011)
6. Hernández-Torruco, J., Canul-Reich, J., Frausto-Solís, J., Méndez-Castillo, J.J.: Feature selection for better identification of subtypes of guillain-barré syndrome. Computational and Mathematical Methods in Medicine 2014, 9 (2014)
7. Jantzen, J., Norup, J., Dounias, G., Bjerregaard, B.: Pap-smear Benchmark Data For Pattern Classification, pp. 1–9. NiSIS (2005)
8. Liu, H., Li, J., Wong, L.: A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. Genome informatics. 13, 51–60 (2002)
9. Liu, Y., Schumann, M.: Data mining feature selection for credit scoring models. Journal of the Operational Research Society 56(9), 1099–1108 (2005)
10. Marinakis, Y., Dounias, G., Jantzen, J.: Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification. Computers in Biology and Medicine 39(1), 69–78 (2009)
11. Pitt, E., Nayak, R.: The use of various data mining and feature selection methods in the analysis of a population survey dataset. In: Proceedings of the 2Nd International Workshop on Integrating Artificial Intelligence and Data Mining - Volume 84. pp. 83–93. AIDM '07, Australian Computer Society, Inc., Darlinghurst, Australia, Australia (2007)
12. Polikar, R.: Pattern Recognition. John Wiley & Sons, Inc. (2006)
13. Ramos, E.C., Huerta, E.B., Caporal, R.M., Cruz, J.F.R., Hernández, J.C.H.: Segmentación e identificación de características anormales de células obtenidas de imágenes microscópicas de cáncer de cérvix, utilizando el método de optimización por enjambre de partículas (pso). Revista Espacio ITH 3(2), 16–20 (2013)
14. Sánchez-Delacruz, E., Acosta-Escalante, F., Boll-Woehrlen, C., Álvarez-Rodríguez, F.J., Hernández-Nolasco, A., Wister, M.A., Pancardo, P.: Categorización de enfermedades neurodegenerativas a partir de biomarcadores de la marcha. Komputer Sapiens 2, 17–20 (May-August 2015)
15. Sánchez-Delacruz, E., Acosta-Escalante, F., Wister, M.A., Hernández-Nolasco, J.A., Pancardo, P., Méndez-Castillo, J.J.: Gait Recognition in the Classification of Neurodegenerative Diseases, pp. 128–135. Springer International Publishing, Cham (2014), http://dx.doi.org/10.1007/978-3-319-13102-3_23
16. Secretaría de Salud: Cáncer de cuello uterino. https://www.gob.mx/salud/acciones-y-programas/cancer-de-cuello-uterino (2015)
17. Sharma, M., Singh, S.K., Agrawal, P., Madaan, V.: Classification of clinical dataset of cervical cancer using knn. Indian Journal of Science and Technology 9(28), 1–5 (2016)
18. Zheng, Z., Wu, X., Srihari, R.: Feature selection for text categorization on imbalanced data. SIGKDD Explor. Newsl. 6(1), 80–89 (Jun 2004)

ISSN 1870-4069

Evaluación de parámetros de encuesta de ingreso del CENEVAL para alumnos candidatos a ingresar al nivel superior, caso de estudio ITP

David Gonzalez-Marron,Angelica Enciso-Gonzalez, Ana Karen Hernandez-Gonzalez, David Gutierrez-Franco, Brenda Guizar-Barrera, Alejandro Marquez-Callejas

Instituto Tecnológico de Pachuca, Pachuca, Hidalgo, México

{dgonzalez, a_enciso}@itpachuca.edu.mx, {karen_heez,davidgf_1993}@hotmail.com, {b.soffgb.comalejandromarqueztec}@gmail.com

Resumen. Los estudiantes candidatos a ingresar a las universidades, requieren efectuar un examen de ingreso donde se verifiquen sus conocimientos adquiridos en el nivel académico anterior a fin de poder desempeñarse adecuadamente en sus estudios superiores, uno de los principales exámenes realizados es el EXANI-II, el cual es aplicado por CENEVAL, para la realización de este examen es necesario llenar una encuesta con datos generales que proporcionan datos socioeconómicos de cada uno de los candidatos, en este trabajo se presenta un análisis de las preguntas utilizadas en esta encuesta junto con los resultados obtenidos en los exámenes, con el fin de validar si hay algunos parámetros socioeconómicos que permitan predecir el factor de éxito en la realización de exámenes de ingreso al Instituto Tecnológico de Pachuca.

Palabras clave: minería de datos, ETL, educación.

Evaluation of CENEVAL Admission Surveyed Parameters for Students that are Candidates to Enter the Higher Education, ITP Study Case

Abstract. Students who are candidates to enter universities, require an entrance examination to verify their knowledge acquired at the previous academic level in order to be able to perform adequately in their higher studies, one of the main tests carried out is the EXANI-II, the Which is applied by CENEVAL, who demands to fill out a survey with general data that provide socioeconomic data of each of the candidates, in this work an analysis of the questions used in this survey together with the results obtained in the exams, are used to validate if there are some socioeconomic parameters that allow predicting the success factor in the entrance exams to the Instituto Tecnologico de Pachuca.

Keywords: data mining, ETL, education.

1. Introducción

En México el Centro Nacional de Evaluación para la Educación Superior (CENEVAL) es el órgano encargado de diseñar y aplicar los instrumentos de evaluación orientados al ingreso y egreso del nivel superior, realiza además el análisis y la difusión de los resultados obtenidos en estas evaluaciones. En este documento se analiza el cuestionario de contexto que se realiza al momento de la solicitud del examen de ingreso a educación superior EXANI-II, el cual es un examen estandarizado que está dirigido a sustentantes que han concluido los estudios de bachillerato y aspiran a ingresar al nivel superior en institutos, colegios y universidades que han contratado los servicios del CENEVAL. Este examen tiene el propósito de establecer una valoración global que permita determinar quiénes son los aspirantes con mayor probabilidad de éxito en el nivel superior. El examen se compone de dos pruebas; EXANI-II Admisión, que evalúa aptitudes y competencias disciplinares predictivas del desempeño, su propósito es detectar el potencial de los aspirantes para cursar con éxito el primer año del nivel educativo al que desean ingresar, apoyando a la toma de decisiones de las instituciones educativas sobre el ingreso a los programas académicos que ofertan, y el EXANI-II Diagnóstico, que mide las competencias disciplinares esenciales que deben dominar los estudiantes para ingresar al programa de educación superior que han elegido. Dado su carácter diagnóstico, la institución usuaria tiene la prerrogativa de incluir o no esta prueba en su proceso de selección.

Tabla 1. Histórico de aspirantes y aceptados.

Plan de Estudios	2013		2014		2015		2016	
	Aspirantes	Aceptados	Aspirantes	Aceptados	Aspirantes	Aceptados	Aspirantes	Aceptados
Arquitectura	373	80	356	80	357	80	342	80
Administración	84	40	63	40	72	36	64	40
Civil	344	80	328	80	396	80	390	80
Diseño Industrial	0	0	0	0	71	40	63	40
Eléctrica	92	40	84	70	96	80	91	80
Industrial	171	80	168	80	247	80	195	80
IGE	96	40	74	40	107	60	106	80
ITIC´s	29	25	52	40	51	35	38	28
Mecánica	160	80	143	80	141	76	127	77
Química	138	80	129	80	138	80	123	80
Sistemas Computacionales	215	80	167	80	168	80	183	80

Los resultados de este examen se pueden ubicar en 3 categorías, estas son: *elemental*, *satisfactorio* y *sobresaliente*, que permiten identificar los factores que inciden en el desempeño de los sustentantes del examen EXANI-II y que están basados en el índice CENEVAL, cuya escala abarca desde los 700 puntos (calificación más baja) hasta los

1300 puntos (calificación más alta). La categoría 1 o *elemental* corresponde a la obtención de 700 a 899 puntos del índice CENEVAL. La categoría 2 o *satisfactoria* va de 900 a 1099 puntos del índice CENEVAL y la categoría 3 o *sobresaliente*, se asigna a los sustentantes que obtienen entre 1100 y 1300 puntos del índice CENEVAL [1]

En este artículo se analizan los atributos socioeconómicos del cuestionario de contexto del EXANI-II, aplicados a 3266 aspirantes a ingresar al Instituto Tecnológico de Pachuca en dos años diferentes, y así tratar de determinar si algunos de estos datos tienen una influencia en el desempeño del examen, se busca identificar si existe un perfil social de los aspirantes aceptados que permita ser reforzado por tutores a través de una tutoría efectiva durante los primeros semestres de estadía de los estudiantes en la institución. El Instituto Tecnológico de Pachuca ofrece 11 programas educativos, los cuáles se muestran en la Tabla 1, pudiendo ser visto que existen programas que tienen más demanda que otros, en la tabla igualmente se muestra el número de aspirantes y aceptados en cuatro diferentes años [2].

Cabe mencionar que la política del Instituto Tecnológico de Pachuca para la selección del EXANI- II se conforma utilizando el siguiente criterio: Examen diagnóstico, al cual se le asigna un valor de 80% y de un 20% para el módulo de especialidad, el criterio para los aspirantes aceptados es que el ICE (Índice de desempeño del CENEVAL) sea de 1000 puntos o más, aunque hay excepciones para aceptar a aspirantes cuyo puntaje es de 920 o mayor, esto con el objeto de formar grupos más nutridos dependiendo del programa educativo seleccionado.

Por la diversidad de los programas educativos es necesario contar con información confiable que permita abatir el índice de deserción, de reprobación y aumentar la eficiencia terminal, como puede ser constatado en las investigaciones llevadas a cabo por investigadores de la Universidad Autónoma de Yucatán [10] y por investigadores del Centro Nacional de Evaluación de la Educación Superior [11].

La estructura del documento está conformada de la siguiente manera, en la sección 2 se describe la encuesta utilizada para este análisis, en la sección 3 se describe el proceso de ETL (extracción, transformación y cargado de los datos), las alternativas para la integración de datos, los comandos utilizados para hacer que las bases de datos de los diferentes exámenes aplicados sean equivalentes, en la sección 4 se explica el algoritmo de Minería de datos empleado en este trabajo y por último en la sección 5 se presentan las conclusiones a las que se llegaron con los resultados de este análisis.

2. Descripción de encuesta realizada

El cuestionario de contexto del EXANI-II incluye factores generales y educativos; el cual se realiza al momento del registro para la obtención de ficha. Éste recopila información socioeconómica, psicológica, personal y académica, estructurada en tres áreas, como se muestra en la Tabla 2.

La información previa del EXANI-II está constituida por 90 atributos originales, los cuales fueron normalizados a 52 para un tipo de exámenes y de 49 para otro tipo de exámenes como se muestra en la Tabla 3, una vez realizada esta operación, se procedió a hacer un filtrado adicional debido a que muchos datos proporcionados por el CENEVAL eran datos redundantes , datos resultados de los exámenes, o datos personales, por lo cual fueron retirados del análisis 13 campos adicionales, los atributos

relacionados con aspectos socioeconómicos fueron englobados en seis constructos como se propone en [12].

Tabla 2. Estructura del cuestionario de contexto del EXANI-II.

Área	Dominio Indicador	Descripción
Personal	Datos Generales	Recopila información general del sustentante, así como si alguno o ambos padres hablan una lengua indígena o dialecto.
	Recursos no cognitivos	Recopila información acerca de la motivación, autocontrol, manejo del tiempo y el compromiso académico.
	Recursos cognitivos	Recopila información sobre la habilidad para realizar actividades de ofimática, el uso de internet y el nivel de dominio del inglés.
	Situación laboral	Recopila información sobre los antecedentes laborales y el trabajo actual.
	Trayectoria académica	Recopila el promedio del bachillerato del sustentante así como si tuvo alguna beca previa.
Escolar	Características de la escuela de procedencia	Identifica el año de egreso, el régimen, modalidad y la institución de origen entre otros.
Social	Familiar	Evalúa a través de la estructura familiar, el capital cultural y económico, así como el involucramiento de los padres en sus estudios.

Tabla 3. Conformación de los constructos encontrados.

	Constructo	Número de variables
1	Habilidad para escribir	6
2	Trabajo bajo presión	7
3	Aspiraciones personales	4
4	Qué tanto lo describe	8
5	Planeación de actividades	6
6	Entorno social	21
	Total	52

A continuación, se describen estos constructos:

Aspiraciones personales: Se valoran aspectos como el máximo nivel de estudios que le gustaría estudiar al sustentante, a que sueldo aspira dentro de 10 años en caso de que se gradúe de una carrera universitaria, en caso de que no se gradúe de una carrera universitaria y en caso de que se gradúe de un posgrado.

Habilidad para escribir: Se conforma de un grupo de valoraciones relacionadas con la habilidad que tiene el sustentante para expresar sus ideas de manera escrita. Involucra su percepción con respecto a sus compañeros, la redacción de su opinión sobre un tema,

la escritura de reportes de una lectura, la redacción de una historia, la escritura de una carta a un familiar y la redacción de una solicitud a una autoridad.

Trabajo bajo presión: Incluye variables que evalúan la forma de respuesta del sustentante a distintas condiciones de evaluación. Por ejemplo, evalúa que tan familiarizado está el estudiante con los exámenes de opción múltiple, si los nervios afectan su desempeño, si el nivel de importancia del examen afecta su desempeño, si el nerviosismo repercute en su velocidad de respuesta, si la presión de tiempo incrementa su nerviosismo, su tipo de respuesta ante preguntas muy difíciles y si ha tenido una experiencia previa al examen en cuestión.

Que tanto lo describe: Se pretende evaluar la actitud y perseverancia del sustentante con valoraciones sobre si se desamina al encontrarse con problemas inesperados, si hay dificultad para mantener la atención en metas que requieren varios meses para ser alcanzadas, si se considera una persona que se esmera, si nuevas metas lo distraen de otras previamente establecidas, entre otras.

Planeación de actividades: Evalúa la organización y cumplimiento de prioridades mediante la valoración de aspectos como la elaboración de una lista de actividades, la planeación de actividades del día, claridad de logros para la semana siguiente, establecimiento de prioridades, cumplimiento de prioridades, realización de cosas que intervienen con sus prioridades.

Entorno social: Evalúa la situación socioeconómica del sustentante mediante la evaluación del nivel de estudios de los padres, la existencia de una figura que oriente académicamente al sustentante, si el sustentante cuenta con servicios como teléfono, lavadora, internet, tabletas electrónicas, horno de microondas, televisión de paga, número de televisores, número de reproductores de DVD, número de computadoras, número de autos, número de baños completos de la casa, entre otras.

Tabla 4. Áreas clasificadas.

Arquitectura	Administración	Ingenierías
Índice de pensamiento matemático	Índice de pensamiento matemático	Índice de pensamiento matemático
Índice de pensamiento analítico	Índice de pensamiento analítico	Índice de pensamiento analítico
Índice de escritura de la lengua	Índice de escritura de la lengua	Índice de escritura de la lengua
Índice de comprensión lectora	Índice de comprensión lectora	Índice de comprensión lectora
	Dictamen de Estadística	Dictamen de Física
		Dictamen de Matemáticas
Dictamen de Ingles	Dictamen de Ingles	Dictamen de Ingles

Con el fin de identificar qué factores socioeconómicos impactan en el desempeño de los sustentantes del EXANI-II, se realizó un análisis con datos provistos por el departamento de Desarrollo Académico del Instituto Tecnológico de Pachuca,

correspondientes a dos años diferentes y aplicada a 3266 aspirantes, de los cuales solo 3105 de éstos contaban con información completa para poder realizar el análisis.

Los datos proporcionados por los 11 programas educativos se clasificaron para su análisis en tres áreas diferentes, ya que por su naturaleza presentaban diferencias fuertes que impactan en el análisis, en la Tabla 4 se muestran los índices de relevancia para cada programa.

3. Realización del proceso de ETL

Para la realización del proceso ETL, se clasificaron los datos en 3 diferentes rubros, ya que son tres áreas que se evalúan con diferentes parámetros, como: (Arquitectura, Administración e Ingenierías) buscándose datos comunes para cada una de éstas áreas, El análisis de cada uno de los campos y de cada uno de los archivos consistió en indagar qué significado tiene cada uno para los objetivos planificados en el proyecto, solo se consideraron los campos de mayor relevancia y que con la aplicación de minería de datos podrían ofrecernos los resultados relevantes. Los campos que contienen datos completamente equivalentes en las 3 áreas seleccionadas son los siguientes y se muestran en la Tabla 5. [3]

Tabla 5. Datos comunes en encuestas.

Campo	Representación
Carrera	identifica la carrera de los aspirantes
Sexo	sexo del aspirante
LI_MAD	lengua indígena de la madre
LI_PAD	lengua indígena del padre
CVE_PROC	clave de la institución de procedencia
PROM_BAC	promedio de bachillerato
BECA_NAC	beca de necesidad económica
HRS_TRAB	horas a la semana de trabajo
ESCO_MAD	escolaridad de la madre
ESCO_PAD	escolaridad del padre
SER_INTE	servicio de internet
SER_CABL	servicio de televisión por cable
BIEN_PC	número de computadoras que hay en casa
ICNE	índice de desempeño CENEVAL
IM	índice de desempeño en matemáticas
ILM	índice de desempeño en lógica matemática
IE	índice de desempeño en español
ICL	índice de desempeño en comprensión lectora
DFIS*	dictamen de desempeño en física
DMAT	dictamen de desempeño en matemáticas
DING	dictamen de desempeño en ingles
DEST**	dictamen de desempeño en estadística
Aceptado	muestra si el aspirante fue aceptado o no

*Solo aplica para las bases de datos de ingenierías
**Solo aplica para las bases de datos de administración

Existen 23 atributos comunes en las encuestas analizadas, siendo innecesario realizar alguna transformación en ellos, sin embargo la importancia de estos datos no se pudo

validar hasta el momento de hacer su análisis. En la Tabla 6 se muestran las diferencias existentes entre dos tipos de encuestas realizadas en diferentes años. Con el fin de poder realizar un análisis equivalente, se estandarizan los campos utilizados y los dictámenes obtenidos, utilizando el número 1 para resultados satisfactorios y el número 2 para resultados no satisfactorios.

Tabla 6. Análisis de atributos equivalentes de dos años diferentes.

Encuesta 20**			Encuesta 20**		
Arquitectura	IRLM	índice de razonamiento lógico matemático	Arquitectura	IPMA	índice de pensamiento matemático
	IMAT	índice de matemáticas		IPAN	índice de pensamiento analítico
	IRV	índice de razonamiento verbal		IELE	índice de escritura de la lengua
	IESP	índice de español		ICLE	índice de comprensión lectora
	DDD_MF_ING*	dictamen de ingles		DDD_ML_ING*	dictamen de ingles
Administración	IRLM	índice razonamiento lógico matemático	Administración	IPMA	índice de pensamiento matemático
	IMAT	índice de matemáticas		IPAN	índice de pensamiento analítico
	IRV	índice de razonamiento verbal		IELE	índice de escritura de la lengua
	IESP	índice de español		ICLE	índice de comprensión lectora
	DDD_MF_EST*	índice de estadística		DDD_ML_EST	dictamen de estadística
	DDD_MF_ING*	dictamen de ingles		DDD_ML_ING*	dictamen de ingles
Ingenierías	IRLM	índice razonamiento lógico matemático	Ingenierías	IPMA	índice de pensamiento matemático
	IMAT	índice de matemáticas		IPAN	índice de pensamiento analítico
	IRV	índice de razonamiento verbal		IELE	índice de escritura de la lengua
	IESP	índice de español		ICLE	índice de comprensión lectora
	DDD_MF_ING*	dictamen de ingles		DDD_ML_ING*	dictamen de ingles
	DDD_MG_FIS*	dictamen de física		DDD_ML_FIS*	dictamen de física
	DDD_MG_MAT	dictamen de matemáticas		DDD_ML_MAT*	dictamen de matemáticas

En la Tabla 7 se muestra la equivalencia de campos para diferentes años y el nombre estandarizado utilizado para uniformizar los atributos [3].

Tabla 7. Estandarización de Atributos equivalentes.

Carrera	Campo _año1	Campo_año2	Nombre Estandarizado
Todas	IRML	IPAN	ILM: Índice lógico matemático
Todas	IMAT	IPMA	IM: Índice de matemáticas
Todas	IRV	ICLE	ICL: Índice de comprensión lectora
Todas	IESP	IELE	IE: Índice de español
Todas	DDD_MF_ING	DDD_ML_ING	DING: Índice de inglés
Administración	DDD_MF_EST	DDD_ML_EST	DEST: Índice de Estadística
Ingeniería	DDD_MG_FIS	DDD_ML_FIS	DFI: Índice de Física
Ingeniería	DDD_MG_MAT	DDD_ML_MAT	DMAT: Índice de Matemáticas

Es necesario realizar diferentes transformaciones a otros campos, a fin de generar archivos compatibles que sea posible utilizar para hacer un análisis con diferentes algoritmos de minería, algunos de estos cambios realizados para estandarización se describen en la Tabla 8.

Tabla 8. Estandarización de Atributos equivalentes.

CAMPO	Actividades de estandarización realizadas
Promedio	Uso de 9 rangos diferentes
Licenciatura de los padres	Uso de 2 valores
Horas trabajadas	Uso de un rango de valores
Escolaridad de los padres	Uso de un rango de valores
Becas por desempeño académico, financieras y deportivas	Uso de 2 valores

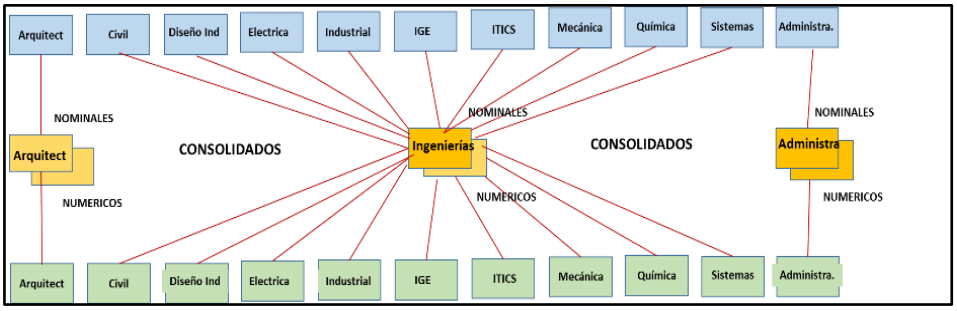


Fig. 1. Proceso de integración de datos utilizando ETL.

Tabla 9. Vista parcial del archivo de salida ARFF.

Viewer

Relation: ADM

PROC	10: PROM_BAC	11: BEC_DAC	12: BEC_NEC	13: BEC_HDA	14: HRS_TRAB	15: ESCO_MAD	16: ESCO_PAD	17: CUAN_LIB	18: SER_TELE	19: SER_LAV	20: SER_REF	21: SER_HOR	22: SER_INTE	23: SER_CABL
nal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
0	5.0	NO	NO	NO	1.0	CARRERA T...	CARRERA T...	3.0	SI	NO	SI	NO	NO	NO
0	4.0	NO	NO	NO	1.0	CARRERA T...	BACHILLER...	4.0	SI	SI	SI	NO	SI	NO
0	7.0	NO	NO	NO	1.0	BACHILLER...	SECUNDARIA	2.0	SI	NO	SI	NO	NO	NO
0	5.0	SI	NO	NO	2.0	BACHILLER...	LICENCIATU...	6.0	SI	NO	SI	NO	NO	SI
0	6.0	NO	NO	NO	1.0	BACHILLER...	LICENCIATU...	6.0	SI	SI	SI	SI	SI	SI
0	3.0	NO	NO	NO	1.0	BACHILLER...	SECUNDARIA	2.0	NO	SI	SI	NO	NO	NO
0	6.0	SI	SI	NO	1.0	BACHILLER...	NO LO SE	7.0	SI	SI	SI	SI	SI	SI
0	3.0	NO	NO	NO	1.0	SECUNDARIA	SECUNDARIA	4.0	NO	SI	SI	NO	NO	NO
0	5.0	NO	SI	NO	1.0	SECUNDARIA	SECUNDARIA	4.0	NO	NO	SI	NO	SI	NO
0	7.0	SI	NO	NO	1.0	SECUNDARIA	NO LO SE	3.0	NO	NO	SI	NO	NO	NO
0	5.0	NO	NO	NO	2.0	SECUNDARIA	SECUNDARIA	2.0	NO	SI	SI	NO	SI	SI
0	5.0	NO	NO	NO	4.0	CARRERA T...	LICENCIATU...	5.0	SI	SI	SI	SI	SI	SI
0	5.0	SI	SI	NO	1.0	LICENCIATU...	LICENCIATU...	4.0	NO	SI	SI	NO	NO	NO
0	5.0	NO	NO	NO	3.0	PRIMARIA	SECUNDARIA	5.0	SI	SI	SI	SI	SI	NO
0	4.0	NO	NO	NO	2.0	SECUNDARIA	SECUNDARIA	2.0	NO	NO	SI	NO	NO	NO
0	7.0	NO	NO	NO	1.0	SECUNDARIA	SECUNDARIA	7.0	NO	NO	SI	NO	NO	NO
0	5.0	NO	NO	NO	2.0	BACHILLER...	BACHILLER...	3.0	SI	SI	SI	SI	SI	SI
0	5.0	NO	SI	NO	4.0	PRIMARIA	NO LO SE	2.0	NO	NO	SI	NO	NO	SI
0	3.0	SI	NO	NO	6.0	SECUNDARIA	PRIMARIA	6.0	SI	SI	NO	NO	SI	NO
0	4.0	SI	NO	SI	6.0	CARRERA T...	PRIMARIA	4.0	SI	NO	SI	NO	NO	SI
0	5.0	NO	NO	NO	2.0	CARRERA T...	CARRERA T...	4.0	SI	NO	SI	NO	NO	NO
0	6.0	NO	NO	NO	2.0	POSGRADO	LICENCIATU...	4.0	SI	SI	SI	NO	SI	SI
0	7.0	NO	SI	NO	1.0	POSGRADO	CARRERA T...	6.0	SI	NO	SI	NO	SI	NO
0	5.0	SI	NO	NO	1.0	LICENCIATU...	CARRERA T...	3.0	SI	SI	SI	NO	NO	NO
0	8.0	SI	NO	NO	1.0	LICENCIATU...	LICENCIATU...	5.0	NO	SI	SI	SI	SI	SI
0	4.0	NO	NO	NO	2.0	LICENCIATU...	POSGRADO	6.0	SI	NO	SI	NO	SI	NO
0	7.0	NO	SI	NO	1.0	CARRERA T...	PRIMARIA	3.0	NO	NO	SI	NO	NO	SI
0	8.0	NO	NO	NO	6.0	CARRERA T...	CARRERA T...	5.0	NO	NO	SI	NO	SI	SI
0	5.0	NO	NO	NO	1.0	BACHILLER...	SECUNDARIA	2.0	NO	SI	SI	SI	NO	NO

Para probar la efectividad de los algoritmos supervisados, se probaron en tipos de datos numéricos y nominales, cuidando en todo momento que los archivos fueran equivalentes a fin de tener una base común de comparación. El proceso de transformación de los datos se realizó utilizando Pentaho [5], permitiendo la transformación de archivos de Excel conformados por 22 archivos en formato CSV (comma-separated values), 11 por cada año analizado y convertidos a 6 archivos ARFF

(Attribute-Relation File Format) 3 nominales y 3 numéricos como puede ser visto en la Figura 1. Los archivos fueron conformados por los atributos que se muestran en la Tabla 9.

4. Realización del proceso de minería

Para la realización del proceso de minería se consideraron los resultados obtenidos de los candidatos en el examen de ingreso CENEVAL de 2 años, ya que se contaba con dicha información a detalle, primeramente se utilizaron algunos algoritmos de clasificación para determinar cuáles son los atributos más influyentes para la determinación de los resultados obtenidos por los candidatos (aceptado o rechazado), utilizando para esto el enfoque de análisis de componentes principales (PCA) [4][7], posteriormente se analizaron que métodos de clasificación proporcionaban mejores resultados. Para la realización del análisis se utilizaron métodos aplicables a datos nominales y métodos para datos numéricos, los tipos de algoritmos utilizados para la evaluación fueron los siguientes: Algoritmos de reglas, algoritmos de árboles, algoritmos de Bayes, algoritmos perezosos y metaalgoritmos [7]. El análisis de componentes principales (PCA) arrojó el siguiente resultado, como puede ser visto en la Tabla 10, donde se muestran los atributos con más relevancia en cada área.

Tabla 10. Análisis de componentes principales PCA para las tres áreas analizadas.

	Campo1	Campo2	Campo3	Campo4	Campo5	Campo6	Campo7	Campo8	Campo9	Campo10
INGENIERIA (2325 Registros)	Trabaja 22.31	Promedio 17.16	Año-Nac 7.01	Vacaciones 5.34	Lug-Proced 4.85	Internet 4.80	Esco-Madre 4.79	Num-libros 4.74	Television 3.85	Esco-Padre 3.74
ARQUITECT (633 Registros)	Trabaja 28.88	Esco-Padre 13.26	Promedio 7.66	Esco-Madre 7.24	Año-Nacim 6.0	Num-libros 5.46	Bec-DepArt 4.25	Automovil 4.12	Tam-Casa 3.89	LenInd-Pad 6.84
ADMON (145 Registros)	Trabaja 30.56	Num-Libros 10.02	Esco-Padre 9.2	Esco-Madre 8.98	Año-Nacim 7.03	LenInd-Mad 6.84	Bec-Acade 6.55	Bec-DepArt 6.26	Serv-Cable 4.35	Promedio 3.25

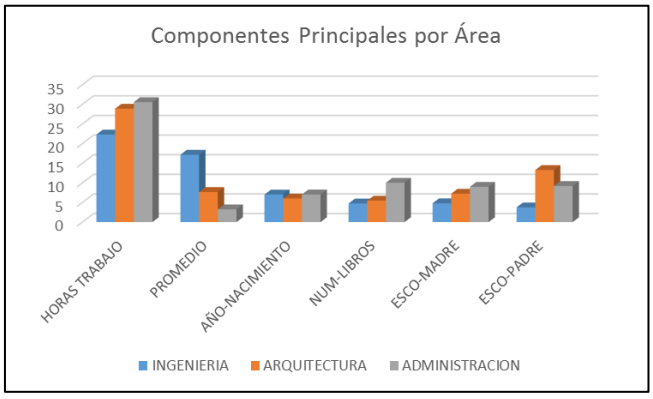


Fig. 2. Contraste de componentes principales por áreas analizadas.

Para la selección de los componentes principales se utilizó la aplicación de BigML [8], seleccionando aquellos atributos que se presentaban en las 3 áreas analizadas, encontrándose coincidencia en 6 atributos, es importante considerar que los más representativos, fueron aquellos pertenecientes al área de ingeniería, ya que el número de registros analizados fueron 2325, contrastados con arquitectura con 635 registros y de únicamente 145 registros pertenecientes a el área de administración.

Aunque el número de registros varía considerablemente, se encontró coincidencia en 6 de los 10 campos más importantes, lo que nos indica que estos campos o atributos son de relevancia para determinar si un alumno es aceptado o rechazado en su examen de ingreso, como puede ser visto en la Figura 2.

El modelo obtenido es presentado en la Figura 3 en forma de árbol, donde puede ser visto que los atributos presentados en la Figura 2 son los que nos permiten hacer una correcta clasificación para el dato objetivo, el cual determina si un alumno aprueba el examen de admisión o no lo hace. A fin de validar la efectividad de la predicción, es generado un nuevo campo denominado "aceptado_prediccion", donde se compara el valor real con el valor calculado a fin de validar la efectividad del modelo utilizado, se seleccionó la métrica denominada como exactitud o accuracy, que está conformada por: (*total_aciertos / total_predicciones*).

Es importante mencionar que existen suficientes casos de aspirantes rechazados y aceptados para tener una muestra representativa de datos y poder realizar un buen análisis.

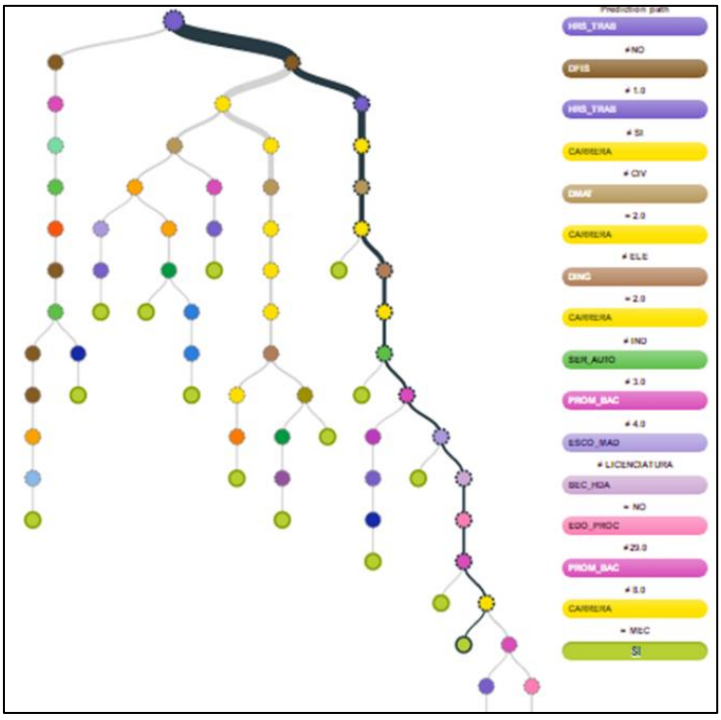


Fig. 3. Modelo de Clasificación obtenido para los datos analizados. [8]

La evaluación del modelo se realizó dividiendo los datos en 2 conjuntos de datos, utilizando 80% de los datos para entrenamiento y el 20% para prueba a fin de evitar un sobreajuste [9], los resultados obtenidos de esta evaluación se presentan en la Figura 4.

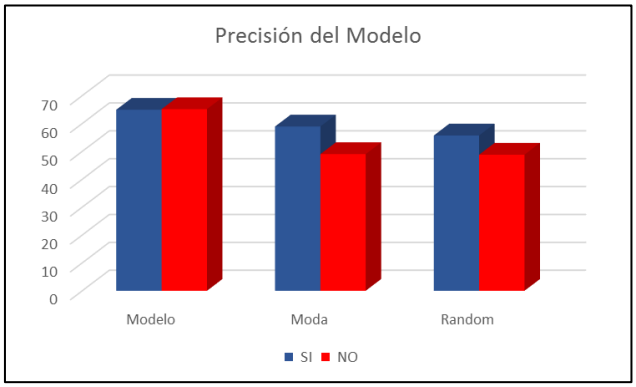


Fig. 4. Evaluación del modelo considerando la exactitud como métrica principal.

En la Figura 5 se muestra el detalle del campo predictor más importante que es el número de horas que trabaja el sustentante, pudiéndose ver que la mayor parte de los candidatos que desean ingresar al Instituto trabajan de 15 a 40 horas semanales.

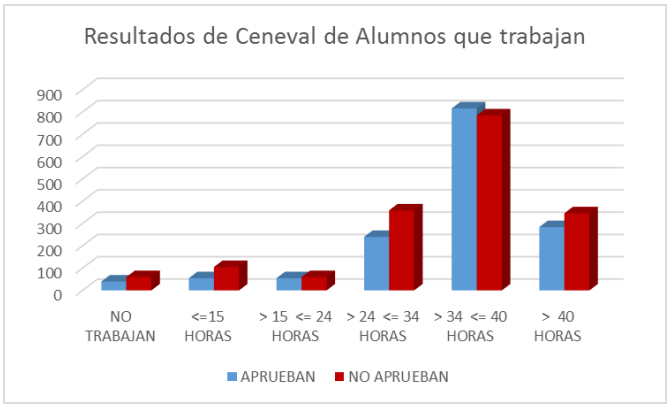


Fig. 5. Desempeño obtenido del campo predictor principal horas trabajadas.

5. Conclusiones

- En base a los resultados obtenidos para el análisis de datos realizado, se puede concluir lo siguiente, el proceso ETL requirió la realización de adecuaciones de los datos, su transformación y cargado en formatos comunes para los diferentes tipos de archivos proporcionados, se englobaron primeramente todas las ingenierías en un archivo común con tipos de datos numéricos y datos nominales, posteriormente se

reunió información de los archivos de las carreras de Arquitectura y Administración para generar un solo archivo global estandarizado.

- Los atributos más relevantes en orden de importancia fueron (horas trabajadas por semana, promedio del bachillerato, año de nacimiento del estudiante, número de libros en casa del estudiante, escolaridad de la madre y escolaridad del padre).
- Los algoritmos utilizados fueron de clasificación para datos nominales y de regresión para datos numéricos, en base a la información proporcionada por el modelo.
- Se pretende generar un plan que permita mejorar la acción de los tutores, donde sea factible, ya que el principal factor considerado fue el de las horas que dedica al trabajo el sustentante, sin embargo este factor seguramente requerirá un apoyo de parte de la institución y de los tutores a fin de no descuidar los estudios de los alumnos una vez que hayan sido aceptados, ya que los alumnos aceptados requieren de tiempo completo para el cumplimiento de sus actividades académicas y frecuentemente este es un factor de reprobación en los primeros semestres de los estudiantes.
- El promedio del bachillerato era algo esperado para poder aprobar el examen del CENEVAL, y efectivamente fue así, lo que valida que el modelo funciona correctamente.
- El número de libros en casa del estudiante fue algo que sorprendió gratamente, ya que indica que, si en la casa hay hábitos de lectura, esto incide en el desempeño de los alumnos.
- La escolaridad del padre y la madre fueron decisivos para un buen desempeño del CENEVAL, ya que en cierta manera refleja cierta presión o acompañamiento de parte de los padres para que los hijos estudien de manera regular.

Referencias

1. CENEVAL: Examenes Nacionales de Ingreso, Disponible en: http://www.ceneval.org.mx/web/guest/exani-ii (2017)
2. ITPACHUCA: Estadística de Ingresos de Alumnos 2013-2016. Departamento de Desarrollo Académico, Informe Interno (2016)
3. Chapa, M., Guizar, B., Franco, D., Hernández, A.: Minería de datos aplicada a encuesta de ingreso del ITP. Reporte de investigación interno (2016)
4. Pérez-López, C.: Minería de datos técnicas y herramientas. Madrid España: Thomson Ediciones Paraninfo S.A. (2008)
5. Rapidminercom.: RapidMiner. Disponible en https://rapidminer.com (2015)
6. Waikatoacnz: Waikatoacnz. Disponible en http://www.cs.waikato.ac.nz/ml/weka/requirements.html (2015)
7. Witten, I., Frank, E., Hall, M.: Data Mining Practical Machine Learning Tool and Techniques. Third Edition, Elsevier (2011)
8. BigML Inc.: Programmatic Machine Learning Application. Disponible en: https://bigml.com (2017)
9. Garcia, D.: Manual de Weka. Disponible en http://sci2s.ugr.es/sites/default/files/files/Teaching/GraduatesCourses/.../weka.pdf (2014)

10. López, I., Echazarreta, C., Pech, S., Gómez, B.: Selección y Permanencia en la Educación Superior: el Caso de la Universidad Autónoma de Yucatán. Revista Iberoamericana de Evaluación Educativa, 3(2), pp. 91–93 (2016)
11. Martínez, J., Herrera, M.: Propiedades psicométricas de la escala de cómputo para el EXANI-II. Revista Electrónica de Investigación Educativa, 16(2), pp. 72–74 (2014)
12. Abreu, J.: Constructos, Variables, Dimensiones, Indicadores & Congruencia. Daena: International Journal of Good Conscience, 7(3), pp. 123–130 (2012)

ISSN 1870-4069

A distance measure for building phylogenetic trees: a first approach

Eunice Ponce-de-Leon-Senti[1], Elva Diaz[1], Hector Guardado-Muro[3], Daniel Cuellar-Garrido[1], Juan José Martinez-Guerra[2], Aurora Torres-Soto[1], Dolores Torres-Soto[4], Arturo Hernandez-Aguirre[5]

[1] Autonomous University of Aguascalientes, Computer Sciences Department, Ags., Mexico

[2] Autonomous University of Aguascalientes, Chemistry Department, Ags., Mexico

[3] Universidad Juárez Autónoma de Tabasco, División Académica de Informática y Sistemas, Tabasco, Mexico

[4] Autonomous University of Aguascalientes, Information Systems Department, Ags., Mexico

[5] CIMAT, A.C., Departamento de Ciencias Computacionales, Guanajuato, Gto., Mexico

eponce@correo.uaa.mx, ediazd@correo.uaa.mx, cuellar_garrido@correo.uaa.mx, atorres@correo.uaa.mx, jjmartin@correo.uaa.mx, hector_guardado_muro@outlook.com, mdtorres@correo.uaa.mx, artha@cimat.mx

Abstract. We propose a distance measure for building phylogenetic trees in comparative genomics. The measure is based on the Bidirectional Best Hit (BBH) concept. The idea behind the measure is, insofar as that two organisms share more Bidirectional Best Hits (BBHs) they are more similar. Although, in general, the sizes of the genomes are different; a similarity measure between two organisms is defined having in the numerator the number of BBHs that exist between them, and in the denominator the semi-sum of the sizes of the genomes of both organisms in order to define a proportion. A distance measure is defined based on the similarity measure. Some restrictions on the number of BBHs between organisms are needed for fulfilling the triangle inequality. We apply different algorithms for building phylogenetic trees reported in literature, using the distance matrix as input, and we obtain suitable phylogenetic trees for a study case of 33 whole proteomes of fungi.

Keywords: distance measure, phylogenetic trees, bidirectional best hits, whole genome or proteome, comparative genomics.

1 Introduction

The many methods for reconstruction phylogenetic trees can be classified into three main categories: parsimony [7], [10], distance [23], [25], [24], and likelihood [8] methods. Our approach is based on the distance methods. The distance methods use the evolutionary distance between operational taxonomic units (in our case, species). Two organisms sharing a recent common ancestor be more similar to each other than two organisms whose last common ancestor was farther. By this reason, it should be possible to infer evolutionary relationships from similarities found between organisms. This is the principle that underlies the distance methods of phylogenetic reconstruction. A distance matrix is generated at each step. The distances represent the dissimilarity between each pair of taxa. The resultant matrix is used to generate a phylogenetic tree. The minimum evolution method [23] use the least-squares method (LSD) in order to give each generated tree a score. Then the tree with the lowest LSD can be found. The unweighted pair group method with arithmetic mean (UPGMA) [25] constructs a tree by identifying the shortest distance in the matrix, clustering those two taxa into a single Operational Taxonomic Unit for use in all subsequent calculations, calculating a new distance matrix, and then repeating these steps. The biggest disadvantage of the UPGMA method is that assumes equal rates of change in each lineage since they diverged from a common ancestor. The neighbor joining method [24] resembles the UPGMA method, but the most important difference is to allow unequal rates of evolution in different branches of the tree.

The concept of bidirectional best hit (BBH) in comparative genomics has had great attention in the last years, that is the case of Wolf and Koonin's paper [27]. They find out experimentally a tight link between orthologs and bidirectional best hits in the case of bacterial and archaeal genomes. Overbeek and his colleagues [19] defined and explained BBHs and used them to detect conserved clusters of genes in order to show that they give evolutionary advantages on individuals and populations. Wolf and Koonin [27] in 2012 said that the ortholog conjecture is the cornerstone of all functional annotation of sequenced genomes and introduced another conjecture, the BBHO (bidirectional best hits (BBH)-orthology equivalence conjecture).

The objective of this paper is to propose a distance, in order to build phylogenetic trees in comparative genomics. To attain this objective the first step is to define an indicator to link the evolutionary theory with the data at hand. In our case we use the bidirectional best hits (BBHs) concept [19], which induces on a set of genes, a reflexive and symmetric binary relation [22], but can not establishes a transitive relation between them. That is, if $\{x_r^i, x_s^j\}$ and $\{x_s^j, x_k^t\}$ are BBHs where x_r^i is a gene that belongs to genome G^i, x_s^j is a gene that belongs to genome G^j and x_k^t is a gene that belongs to genome G^t, then we can not affirm that $\{x_r^i, x_k^t\}$ is a BBH, in general, i.e., there is no guaranty of the existence of this latter pair of genes as a bidirectional best hit. However, we can define a metric based on the BBH concept in order to measure how many genes share two genomes under the BBH's relation. A distance matrix will be used then to

construct phylogenetic trees from knowing distances based methods reported in literature [9].

The case of study, in this paper is a database of 33 whole proteomes of fungi, i.e., completely sequencing proteomes. The fungi are selected because they are very important organisms, that won their place as an object of fundamental research because they affect our daily lives as causative agents of disease, as sources of food, as agents for recycling of biomass, as key ingredients in industrial processes, as essentials tools in medicine (penicillin- that changed the population growth pattern), and as models to study properties of evolution.

Until 2006 most fungal phylogenies had been derived from single gene comparison, or from concatenated alignment of a small number of genes. After 2006 the availability of greater number of data laid the basis to reconstruct phylogenetic trees from whole genomes. Fitzpatrick and his colleagues in 2006 used 42 whole genomes and agglomerative methods to construct a phylogenetic tree [11]. Wang and his colleagues in 2009 [26] had constructed a kingdom-wide fungal phylogenetic tree for 82 sequenced genomes using an alignment free composition method (CV) previously successfully applied to prokaryotik and viral phylogenics [21]. The method is based on whole genomes [20].

The next section discusses the fundamental concepts and notation that establish a relationship between a biological problem, i.e. the building of phylogenies and the mathematical formalism. In the section 3 the properties of the distance measure are proven. The section 4 explains the steps to follow, in order to apply the genomic distance measure. The section 5 gives the results and discussion in the case of study, and finally, the section 6 gives the conclusions and future work.

2 Basic Concepts and Notation

In order to formulate mathematically the biological problem of finding a phylogenetic tree in comparative genomics we have defined some involved concepts from molecular biology. The comparative genomics is a very powerful tool in Molecular Biology [14] and in particular, the phylogenetic reconstruction has become one of the main tools of comparative genomics [16], especially, in the case of the enormous amounts of data generated by several molecular biology methods. A phylogeny is, in general, the evolutionary history of a group of related entities. A genome is the entire genetic constitution of a living organism. We define here the concepts of a whole genome, the genome's size, a BBH, the set of all BBHs between two genomes and the genomic distances matrix. In literature, whole genome is a genome that is sequenced completely. We define a whole genome as a n-tuple of ordered genes, and genome's size as the number of genes of the genome, i.e., n. Henceforth, we will give all definitions using the genome concept, but these can be reformulated mathematically for the proteome concept.

Definition 1. *A whole genome is defined as a n-vector of genes,* $G = (x_1, ..., x_n)$. *The number of genes in the whole genome represents the genome's size, i.e.,* n. *Let* $G^i = (x^i_1, ..., x^i_{n_i})$ *and* $G^j = (x^j_1, ..., x^j_{n_j})$ *be genomes,* n_i *and* n_j *denote the size of genomes* G^i *and* G^j *respectively.*

Henceforward, when we refer to a genome, we are supposing it is a whole-genome.

Definition 2. *For all genomes G^i the number of genes n_i in G^i satisfies that*

$$n_i \in \mathbf{Z}^+ - \{0\} \ . \tag{1}$$

This is trivial for modelling the biological problem because there is no reason to consider a genome G^i, if this genome do not have genes. The values that the number of genes n_i can take, are positives and integers.

Definition 3. *The BBH definition [19] is as follows: Let G^i and G^j be genomes. Two genes, x_r^i , the r-th element of the n_i-tuple G^i, and x_s^j, the s-th element of the n_j-tuple G^j, are called a Bidirectional Best Hit (BBH), if and only if recognizable similarity exists between them (in our case, we use the BLAST similarity score with a E-value [12] lower than* 1.0 x 10^{-5}*), and there is no gene x_k^j of the n_j-tuple G^j that is more similar than x_s^j is to x_r^i, and there is no gene x_k^i of the n_i-tuple G^i that is more similar than x_r^i is to x_s^j. We denote it as, $x_r^i \leftrightarrow_{BBH} x_s^j$.*

Definition 4. *The set of all BBHs between two genomes G^i and G^j is the set*

$$B_{i,j} = \{\{x_r^i, x_s^j\} : x_r^i \leftrightarrow_{BBH} x_s^j\}, \tag{2}$$

and the cardinality of $B_{i,j}$ is the number of BBHs that exist between G^i and G^j, and is denoted by $|B_{i,j}|$.

Definition 5. *For all pair of genomes G^i and G^j the number of BBHs, that is, $|B_{i,j}|$ satisfies that $|B_{i,j}| \in \mathbf{Z}^+$ and it can not exceed the genes number of the shortest genome between both, i.e.,*

$$|B_{i,j}| \leq min\{n_i, n_j\} \ , \tag{3}$$

where n_i and n_j are the genomes size of G^i and G^j respectively.

In the first part of this definition, we have considered that the number of BBHs between two genomes G^i and G^j is a positive number that can be zero. This refers to that do not exist any pair of genes taking one of G^i, and the other of G^j that they are a BBH. Of course, $B_{i,j}$ is a positive number because it is related with the number of genes of G^i and G^j. The second part of this definition about the upper bound of BBHs number between two genomes means that the upper bound of BBHs number is the size of the smallest genome.

Definition 6. *The genomic distances matrix between a set of genomes is a symmetric matrix constructed for all organisms of the study, from a distance measure between all pair of genomes of these organisms.*

3 Proposed Distance Measure

In this part of the chapter, the distance measure for building phylogenetic trees from genomic information is introduced. This distance measure is made from a similarity measure, based on bidirectional best hits [19]. The idea behinds of this is to consider the following premise: two organisms that share more BBHs than other two, should be more similar. To model the real biological problem, the different sizes of the genomes of the involved organisms should be considered in the definition of the measure. The measures defined here are measures between whole genomes.

Definition 7. *A similarity measure δ between G^i and G^j is defined as follows:*

$$\delta(G^i, G^j) = \frac{2|B_{i,j}|}{n_i + n_j}, \tag{4}$$

$|B_{i,j}|$ is the total number of BBHs between G^i and G^j. n_i is the total number of genes from genome G^i, and n_j is the total number of genes from genome G^j.

The similarity measure is standardized with respect to the size of their genomes in order to eliminate the effect of the difference of genomes sizes. The dissimilarity measure is defined as additive inverse of the similarity measure [5] as follows:

Definition 8. *A dissimilarity measure d between G^i and G^j is given by*

$$d(G^i, G^j) = 1 - \delta(G^i, G^j) = 1 - \frac{2|B_{i,j}|}{n_i + n_j}. \tag{5}$$

In the following we demonstrate that $d(G^i, G^j)$ obeys the following four properties, i. e., $d(G^i, G^j) \geq 0$, reflexivity, symmetry and triangle inequality.

Property 1. The dissimilarity measure between two genomes G^i and G^j is a positive number or is equal to zero.

Proof.

$$d(G^i, G^j) = 1 - \delta(G^i, G^j) = 1 - \frac{2|B_{i,j}|}{n_i + n_j} \tag{6}$$

$d(G^i, G^j)$ is a positive number or is equal to zero, if and only if,

$$1 \geq \delta(G^i, G^j) \geq 0, \tag{7}$$

That is

$$1 \geq \frac{2|B_{i,j}|}{n_i + n_j} \geq 0. \tag{8}$$

Since Definition 5

$$|B_{i,j}| \leq min\{n_i, n_j\}, \tag{9}$$

because $|B_{i,j}|$ is the number of BBHs between G^i and G^j, and this number can not overtake the number of genes of the smallest genome because of BBH's

definition. Suppose $n_i \leq n_j$, that is, $|B_{i,j}| \leq n_i$ by Definition 5. The following steps are true:

$2|B_{i,j}| \leq 2n_i = n_i + n_i$ substitute the second n_i by n_j knowing $n_i \leq n_j$,
$2|B_{i,j}| \leq n_i + n_j$ dividing both sides by $n_i + n_j$, where

$$\frac{2|B_{i,j}|}{n_i + n_j} \leq 1. \tag{10}$$

An analogous demonstration is possible as $n_j \leq n_i$.

Demonstrate that $\delta(G^i, G^j) \geq 0$, i.e., $\frac{2|B_{i,j}|}{n_i+n_j} \geq 0$. $|B_{i,j}| \in \mathbf{Z}^+$ by BBH's definition and also $2|B_{i,j}| \in \mathbf{Z}^+$. n_i and $n_j \in \mathbf{Z}^+$. The quotient of positive integer numbers belong to $\mathbf{R}^+$. Finally, $n_i \neq 0$ and $n_j \neq 0$ by Definition 2. Finally

$$0 \leq \frac{2|B_{i,j}|}{n_i + n_j} \leq 1. \tag{11}$$

Property 2. Reflexivity property: For every genome G^i, $d(G^i, G^i) = 0$.

Proof.

$$d(G^i, G^i) = 1 - \delta(G^i, G^i) = 1 - \frac{2|B_{i,i}|}{n_i + n_i} = 1 - \frac{2n_i}{2n_i} = 1 - 1 = 0. \tag{12}$$

$|B_{i,i}| = n_i$ because the number of BBHs of a genome G^i with itself is the same as the number of genes the genome G^i has.

Property 3. Symmetry property: For all pair of genomes, G^i and G^j, $d(G^i, G^j) = d(G^j, G^i)$.

Proof.

$$d(G^i, G^j) = 1 - \delta(G^i, G^j) = 1 - \frac{2|B_{i,j}|}{n_i + n_j} = 1 - \frac{2|B_{j,i}|}{n_j + n_i} = 1 - \delta(G^j, G^i) = d(G^j, G^i). \tag{13}$$

$|B_{i,j}| = |B_{j,i}|$ because the number of BBHs of G^i with G^j is the same as the number of BBHs of G^j with G^i, due the BBH's definition.

Property 4. Triangle inequality property: Let G^i, G^j and G^k be the genomes in a database, and let n_i, n_j, n_k be the genome sizes respectively. Since $n_i, n_j, n_k \in \mathbf{Z}^+ - \{0\}$ and $|B_{i,k}|, |B_{k,j}|, |B_{i,j}| \in \mathbf{Z}^+$ and suppose that the following expressions are fulfilled,

$$|B_{ik}| = \frac{n_i + n_k}{2}\, \alpha \leq min\{n_i, n_k\}, \tag{14}$$

$$|B_{kj}| = \frac{n_k + n_j}{2}\, \beta \leq min\{n_k, n_j\}, \tag{15}$$

$$0 \leq \alpha + \beta \leq 1\ , \tag{16}$$

and then the triangle inequality $d(G^i, G^j) \leq d(G^i, G^k) + d(G^k, G^j)$ is fulfilled. Notice that $\alpha \geq 0$ and $\beta \geq 0$ because in the suppositions (14) and (15), the left side of both equations is a positive number, that is, $|B_{ik}| \in \mathbf{Z}^+$ and $|B_{kj}| \in \mathbf{Z}^+$ as we show in definition 5.

Proof. Since the suppositions in (14) and (15) we have

$$2|B_{ik}| = (n_i + n_k)\ \alpha, \tag{17}$$

$$2|B_{kj}| = (n_k + n_j)\ \beta, \tag{18}$$

that is

$$1 - \frac{2|B_{ik}|}{n_i + n_k} = 1 - \alpha, \tag{19}$$

$$1 - \frac{2|B_{kj}|}{n_k + n_j} = 1 - \beta. \tag{20}$$

Adding (19) and (20),

$$1 - \frac{2|B_{ik}|}{n_i + n_k} + 1 - \frac{2|B_{kj}|}{n_k + n_j} = 2 - \alpha - \beta. \tag{21}$$

Since the supposition (16) we have

$$2 - (\alpha + \beta) \geq 1. \tag{22}$$

Finally, using inequality (11), we obtain that the triangle inequality

$$1 - \frac{2|B_{ik}|}{n_i + n_k} + 1 - \frac{2|B_{kj}|}{n_k + n_j} \geq 1 \geq 1 - \frac{2|B_{ij}|}{n_i + n_j}, \tag{23}$$

i.e.

$$d(G^i, G^j) \leq d(G^i, G^k) + d(G^k, G^j) \tag{24}$$

is fulfilled for all $n_i, n_j, n_k \in \mathbf{Z}^+ - \{0\}$ and $|B_{i,k}|, |B_{k,j}|, |B_{i,j}| \in \mathbf{Z}^+$ and for all G^i, G^j and G^k that satisfy (14), (15) and (16). The property has been demonstrated.

We have demonstrated that the dissimilarity measure d defined in Definition 8 is a distance measure. In the following we refer d as a distance measure.

The suppositions (14), (15) and (16) made for the proof of triangle inequality property means that the number of BBHs for any two pairs (i, k) and (k, j) of organisms will be always less than the size of the shortest genome, in a factor α and a factor β respectively, and $0 \leq \alpha + \beta \leq 1$.

4 Materials and Methods

In this section we describe the method for applying the proposed distance measure for whole genomes in order to build phylogenetic trees using the proposed

distance as the input of several distance based algorithms reported in literature [9]. The first step is to define the set of organisms in the study, and to establish a research hypothesis to test. The second step is to obtain the whole genomes or whole proteomes of these organisms from an appropriated database in the web. The third step is to obtain the BBHs for all pair of organisms using BLAST [1]. The fourth step is to make the genomic distance matrix using Definition 5 and finally, to run the algorithms that build the phylogenetic trees and to analyse the results with respect to the research hypothesis.

4.1 Data Description

In this study, the fungal phylogenetic trees are made with whole proteomes of 33 representative fungi from the following phyla: one organism of Zygomycota phylum, two of Chytridiomycota phylum, four of Basidiomycota phylum, and 26 of Ascomycota phylum. The proteomes used (see Table 1) are obtained from the Broad Laboratory [3], the Bordeaux Bioinformatics Center (CBiB) [2], and The European Bioinformatics Institute [6]. Each one is selected because it is a whole proteome.

Table 1 contains the list of the 33 fungal organisms for this study detailing its genus, species, variety, proteome size, and and the institute from where the data was obtained. Table 2 contains information important for reading and interpreting the obtained phylogenetic tree, such as the taxonomy in terms of the phylum, subphylum, class and order for all fungal proteomes in study.

4.2 Preprocessing of Data

A part of the data analysis is realized using BLAST (Basic Local Alignment Search Tool) [1] a free program from the National Center for Biotechnology Information (NCBI) [17]. This program performs different types of analysis. In this case the program is used to compare proteins that belongs to each pair of fungi. This program is used to obtain for each protein of the first fungus the one of the second fungus that best resembles it. This correspondence is denoted as a "best hit". The analysis is performed then in reversed order, and if the same protein that was a best hit is given back as a best hit too, it is said that a "bidirectional best hit" has occurred (see Definition 3) if the BLAST similarity scores in both directions are lower than 1.0 x 10^{-5}. The fundamental basis to the construction of the phylogenetic tree in our approach is to obtain all the BBHs existing between every pair of fungal proteomes.

The cut off point for determining which a best hit is and what is not, is determined by their expectation value. The expectation value (or E) in BLAST is a statistical significance threshold for reporting matches against database sequences. The typical value of E found in the literature for obtaining the BBHs is of 1.0 x 10^{-5} due the necessity of assessing the best resemblance between proteins of the species, not allowing alignments that appear very similar at first sight but, in closer examination they are not because they were well aligned by chance [4].

Table 1. Fungal organisms used in this analysis are listed.

No.	Identifier	Genus	Species	Variety	Proteins	Citation
1.	ASHBYA	*Ashbya*	*gossypii*	Q	4718	SP
2.	ASPERF	*Aspergillus*	*fumigatus*	Afu	9888	BROAD
3.	ASPERN	*Aspergillus*	*nidulans*	AN	10665	BROAD
4.	ASPERT	*Aspergillus*	*Terreus*	ATEG	10406	BROAD
5.	BATRAC	*Batrachochytrium*	*dendrobatidis*	BDEG	8818	BROAD
6.	BOTRYT	*Botrytis*	*cinerea*	BC1G	16389	BROAD
7.	CANDAL	*Candida*	*albican*	CAWG	6157	BROAD
8.	CANDGL	*Candida*	*glabrata*	CAGR	5215	GNL
9.	CANDGU	*Candida*	*guilliermondii*	PGUG	5920	BROAD
10.	CANDLU	*Candida*	*lusitaniae*	CLUG	5936	BROAD
11.	CANDTR	*Candida*	tropicalis	CTRG	6258	BROAD
12.	CHAETO	*Chaetomium*	globosum	CHGG	11124	BROAD
13.	COCCID	*Coccidiodes*	*immitis*	CIMG	10457	BROAD
14.	COPRIN	*Coprinus*	*cinereus*	CC1G	13544	BROAD
15.	CRYPTO	*Cryptococcus*	*neoformans*	CNAG	7302	BROAD
16.	DEBARY	*Debaryomyces*	*hansenii*	DEHA	6319	GNL
17.	FUSAGR	*Fusarium*	*graminearum*	FGSG	13321	BROAD
18.	FUSAOX	*Fusarium*	*oxysporum*	FOXG	17608	BROAD
19.	FUSAVE	*Fusarium*	*verticilloides*	FVEG	14195	BROAD
20.	HISTOP	*Histoplasma*	*capsulatum*	HGAC	9349	BROAD
21.	KLUYVE	*Kluyveromyces*	*lactis*	KLLA	5327	GNL
22.	LODDER	*Lodderomyces*	*elongisporus*	LELG	5796	BROAD
23.	MAGNAP	*Magnaphorte*	*grisea*	MGG	12832	BROAD
24.	NEUROS	*Neurospora*	*crassa*	NCU	9823	BROAD
25.	PUCCIN	*Puccinia*	*graminis*	PGTG	20567	BROAD
26.	RHIZOP	*Rhizopus*	*oryzae*	RO3G	17467	BROAD
27.	SACCHA	*Saccharomyces*	*cerevisiae*	SCRG	5388	BROAD
28.	SCHIZO	*Schizosaccharomyces*	*japonicus*	SJAG	5168	BROAD
29.	SCLERO	*Sclerotinia*	*sclerotiorum*	SS1G	14522	BROAD
30.	STAGON	*Stagonospora*	*nodorum*	SNU	16597	BROAD
31.	UNCINO	*Uncinocarpus*	*reesii*	UREG	7798	BROAD
32.	USTILA	*Ustilago*	*maydis*	UM	6522	BROAD
33.	YARROW	*Yarrowia*	*lipolytica*	YALI	6436	GNL

We take each pair combination P^i, P^j for every 33 whole proteomes in BLAST being the proteins of the fungus P^i the query and the proteins of the fungus P^j the database in order to obtain 1056 files of best hits. The next step is to take every pair of files in which we stored the best hits between P^i and P^j from $i \to j$ and from $j \to i$ to obtain the bidirectional best hits. After preprocessing the BBHs between all pairs of fungi, 528 independent files are obtained. This number corresponds to the number of necessary comparisons in order to obtain the BBH's number for each pair of fungi. When this magnitude is obtained for all fungi pair, it is possible to obtain the genomic distances matrix (see Definition 6) using the distance measure (Definition 5).

A computer program tests the accomplishment of restrictions made in the proof of triangle inequality, that is, the restrictions (14), (15) and (16) for all pair of organisms in our database, i.e., for any three fungi, the triangle inequality for the distance measure is fulfilled.

5 Results and Discussion

MEGA7 software [13] is used for running different phylogenies with its distance based algorithms. The genomic distances matrix with the distance values for

Table 2. The usual fungal taxomomy. The zygomycota and chytridiomycota phyla, have only one fungus respectively. The basidiomycota phylum has 4 fungi. The rest of the fungi belongs to ascomycota phylum, and are classified in three subphyla: pezizomycotina, saccharomycotina, and taphrinomycotina. This taxonomy corresponds to [11], and [26].

Identifier	Phylum	Subphylum	Class	Order
ASHBYA	Ascomycota	Saccharomycotina	Saccharomycetes	Saccharomycetales
ASPERF	Ascomycota	Pezizomycotina	Eurotiomycetes	Eurotiales
ASPERN	Ascomycota	Pezizomycotina	Eurotiomycetes	Eurotiales
ASPERT	Ascomycota	Pezizomycotina	Eurotiomycetes	Eurotiales
BATRAC	Chytridiomycota	Chytridiomycotina	Chytridiomycetes	Chytridiales
BOTRYT	Ascomycota	Pezizomycotina	Leotiomycetes	Helotiales
CANDAL	Ascomycota	Saccharomycotina	Saccharomycetes	Saccharomycetales
CANDGL	Ascomycota	Saccharomycotina	Saccharomycetes	Saccharomycetales
CANDGU	Ascomycota	Saccharomycotina	Saccharomycetes	Saccharomycetales
CANDLU	Ascomycota	Saccharomycotina	Saccharomycetes	Saccharomycetales
CANDTR	Ascomycota	Saccharomycotina	Saccharomycetes	Saccharomycetales
CHAETO	Ascomycota	Pezizomycotina	Sordariomycetes	Sordariales
COCCID	Ascomycota	Pezizomycotina	Eurotiomycetes	Onygales
COPRIN	Basidiomycota	Agaricomycotina	Hymenomycetes	Agarigales
CRYPTO	Basidiomycota	Agaricomycotina	Hymenomycetes	Tremellales
DEBARY	Ascomycota	Saccharomycotina	Saccharomycetes	Saccharomycetales
FUSAGR	Ascomycota	Pezizomycotina	Sordariomycetes	Hypocreales
FUSAOX	Ascomycota	Pezizomycotina	Sordariomycetes	Hypocreales
FUSAVE	Ascomycota	Pezizomycotina	Sordariomycetes	Hypocreales
HISTOP	Ascomycota	Pezizomycotina	Eurotiomycetes	Onygales
KLUYVE	Ascomycota	Saccharomycotina	Saccharomycetes	Saccharomycetales
LODDER	Ascomycota	Saccharomycotina	Saccharomycetes	Saccharomycetales
MAGNAP	Ascomycota	Pezizomycotina	Sordariomycetes	Magnaporthales
NEUROS	Ascomycota	Pezizomycotina	Sordariomycetes	Sordariales
PUCCIN	Basidiomycota	Pucciniomycotina	Pucciniomycetes	Pucciniales
RHIZOP	Zygomycota	Mucoromycotina	Zygomycetes	Mucorales
SACCHA	Ascomycota	Saccharomycotina	Saccharomycetes	Saccharomycetales
SCHIZO	Ascomycota	Taphrinomycotina	Schizo-saccharomycetes	Schizo-saccharomycetales
SCLERO	Ascomycota	Pezizomycotina	Leotiomycetes	Helotiales
STAGON	Ascomycota	Pezizomycotina	Dothideomycetes	Pleosporales
UNCINO	Ascomycota	Pezizomycotina	Eurotiomycetes	Onygales
USTILA	Basidiomycota	Ustiloginomycotina	Ustilaginomycetes	Ustilaginales
YARROW	Ascomycota	Saccharomycotina	Saccharomycetes	Saccharomycetales

each pair of proteomes of fungi is the input to MEGA 7 software. We use three distance based methods for building phylogenetic trees: the UPGMA, the Neighbor-Joining and the Minimum Evolution.

In Fig. 1 the phylogenetic tree using the UPGMA Method is presented [25]. The evolutionary history was inferred from the UPGMA method. The optimal tree with the sum of branch length = 7.06527874 is shown. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances are calculated from Definition 5.

In Fig. 2 the phylogenetic tree using the Neighbor-Joining method is presented. The evolutionary history is inferred using the Neighbor-Joining [24] method. The optimal tree with the sum of branch length = 7.09372415 is shown. The tree is drawn to scale, with branch lengths (next to the branches) in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances are calculated from Definition 5.

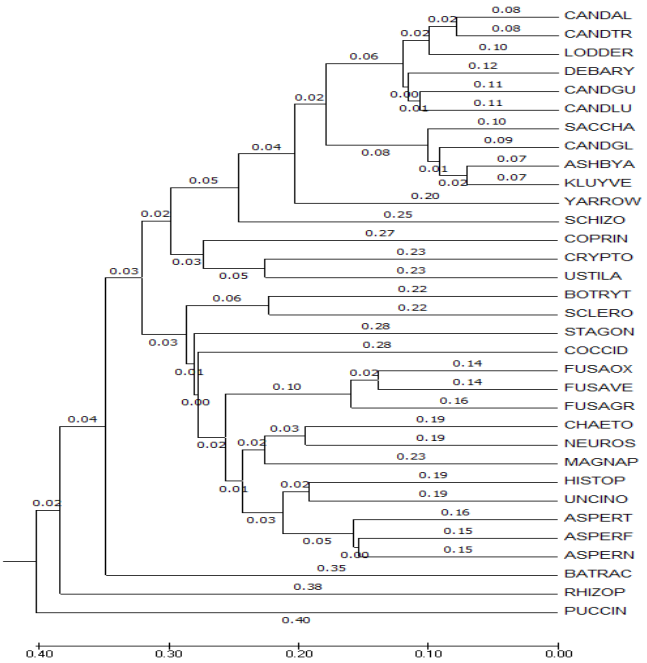


Fig. 1. Phylogenetic tree using UPGMA method.

In Fig. 3 the phylogenetic tree using the Minimum Evolution method is presented. The evolutionary history is inferred using the Minimum Evolution method [23]. The optimal tree with the sum of branch length = 7.09372415 is shown. The tree is drawn to scale, with branch lengths (next to the branches) in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances are calculated from Definition 5. The Minimum Evolution tree is searched using the Close-Neighbor-Interchange (CNI) algorithm [18] at a search level of 1. The Neighbor-joining algorithm [24] is used to generate the initial tree.

For all methods we can observe that the Saccharomycotina subphylum has been well identified respect to the main clades of subphylum, CTG and WDG in all phylogenetic trees. All organisms in our database that belong to Saccharomycotina subphylum are of the same order, that is, saccaromycetales. The members of CTG clade in our database are CANDLU, CANDGU, DEBARY, LODDER, CANDTR, and CANDAL. The members of WGD clade in our database are SACCHA, CANDGL, KLUYVE, and ASHBYA.

The Pezizomycotina subphylum has five different orders, Eurotiales, Helotiales, Sordariales, Onygales and Hypocreales. In all mentioned orders for all phylogenetic trees obtained, the topologies of the order's subtrees are the same as reported in [26].

In our database four fungi belong to the Basidiomycota phylum, they have

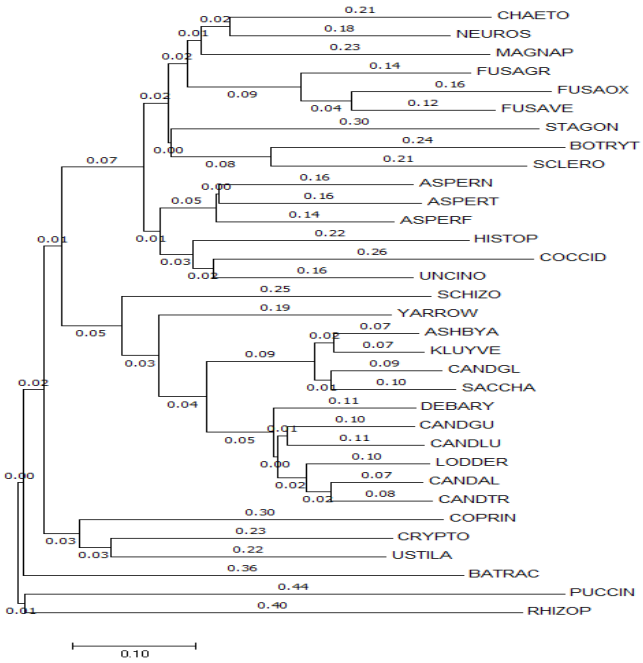


Fig. 2. Phylogenetic tree using Neighbor-Joining method.

been classified well in the Neighbor-Joining and Minimum Evolution methods. Only in the UPGMA method three fungi belonging to the Basidiomycota phylum (USTILA, CRIPTO and COPRIN see Table 2) have been classified wrong. On the other hand, the RHIZOP fungus belonging to the Zygomyccota phylum obtains an adequate position in the topology of the phylogenetic tree. The same situation occurs in the case of BATRAC fungus, which is classified as belonging to the Chytridiomycota phylum. In the case of the Taphrinomycotina subphylum exits the discussion about two different places where to assign it [15]. We obtain that the SCHIZO fungus, single representative of Taphrinomycotina subphylum in our study, is branching as a sister group to Saccharomycotina.

6 Conclusions and Future Work

The dissimilarity measure d between whole genomes has been demonstrated as a distance measure and it satisfies the four properties of a distance, i.e., the measure is a positive number or equal to zero, it is reflexive, symmetric and fulfils the triangle inequality if the suppositions (14), (15) and (16) are supported. Specially, the distance measure d is less than or equal to 1.

The resulting phylogenetic trees are in agreement with the most part of topologies and groups reported in the literature, for example, the fungal phylogenies

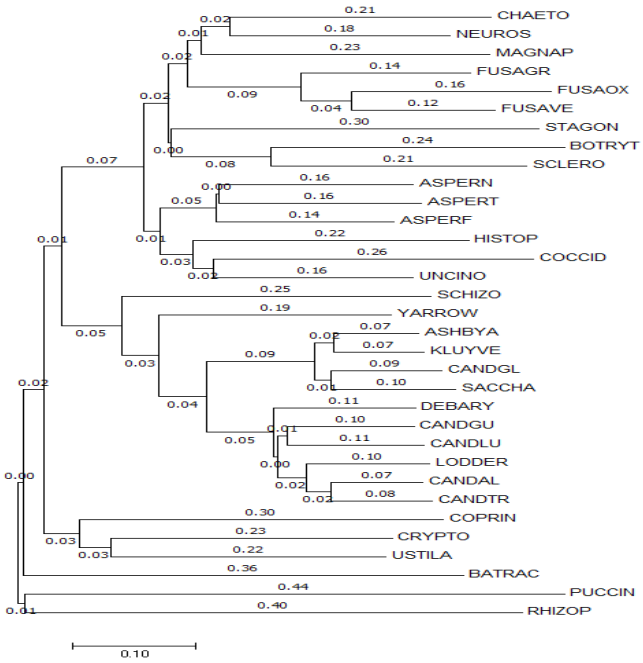


Fig. 3. Phylogenetic tree using Minimum Evolution method.

obtained in [11], and [26]. As seen in Table 2 where appears the taxonomy in terms of the phylum, subphylum, class and order for all fungal proteomes in study, our resulting phylogenetic trees classify correctly the Pezizomycotina subphylum until the order level. In the case of Saccharomycotina subphylum two main clades are identified. The Neighbor-Joining and Minimum Evolution methods obtain very similar phylogenies and the best results. They are also very similar to phylogenies reported in [11], and [26].

The future work will be to test if incorporating BBH - based phylogenetic tree structural information contributes to study the different functional groups of proteins for a set of organisms in study.

Acknowledgments. We would like to acknowledge the support for this project (PIINF15-1) given by the Autonomous University of Aguascalientes, Aguascalientes, Mexico.

References

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J.: Basic local alignment search tool. Journal of Molecular Biology 215(3):403–410 (1990)
2. Bordeaux Bioinformatics Center (CBiB), http://proteome.cgfb.u-bordeaux.fr

3. Broad Laboratory, http://archive.broadinstitute.org/ftp/pub/annotation/fungi
4. Brenner, S.E., Chothia, C., Hubbard, T. J.: Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc. Natl. Acad. Sci. USA, 95(11):6073–6078, (1998)
5. Deza, M. M., Deza,E.: Encyclopedia of Distances. Springer-Verlag (2013)
6. European Bioinformatics Institute, ftp://ftp.ebi.ac.uk/pub/databases/integr8/last_release/fasta/proteomes
7. Farris, J. S.: Methods for computing Wagner trees. Systematic Zoology 19:83–92 (1970)
8. Felsenstein, J.: Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach. Journal of Molecular Evolution, 17(6):368–376 (1981)
9. Felsenstein, J.: Inferring Phylogenies. Sinauer Associates, Inc. (2004)
10. Fitch, W.M.: Toward defining the course of evolution: minimum change for a specified tree topology. Systematic Zoology 20(4):406–416 (1971)
11. Fitzpatrick, D. A., Logue, M.E., Stajich, J. E., Butler, G.: A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evolutionary Biology 6(99):1–15 (2006)
12. Gertz, E. M.: BLAST Scoring Parameters. Available at:ftp://ftp.ncbi.nlm.nih.gov/blast/documents/developer/scoring.pdf (2005)
13. Kumar, S., Stecher, G., Tamura, K.: MEGA7 Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Molecular Biology and Evolution 33:1870–1874 (2016)
14. Lackie, J. M.: The dictionary of cell and Molecular Biology. Academic Press Elsevier (2007)
15. Liu, Y., Leigh, J. W., Brinkmann, H., Cushion, M. T.: Phylogenomic Analyses Support the Monophyly of Taphrinomycotina, including Schizosaccharomyces Fission Yeast. Mol. Biol. Evol. 26(1):27–34 (2009)
16. Mushegian, A. R.: Foundations of Comparative Genomics. Elsevier Academic Press (2007)
17. National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov
18. Nei, M., Kumar, S.: Molecular Evolution and Phylogenetics. Oxford University Press, New York (2000)
19. Overbeek, R., Fonstein, M., D'Souza, M., Push, G. D., Maltsev, N.: The use of gene clusters to infer functional coupling. Proc. Natl. Acad. Sci. 96(6):2896–2901 (1999)
20. Qi, J., Luo, H., Hao, B.: CVTree: a phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Res. 32:W451–W47 (2004)
21. Qi, J., Wang, B., Hao, B.: Whole Proteome Prokaryote Phylogeny without Sequence Alignment: A K-String Composition Approach. J. Mol. Evol. 58:1–11 (2004)
22. Rosen, K. H.: Discrete Mathematics and Its Applications. McGraw-Hill (1999)
23. Rzhetsky, A., Nei, M.: A simple method for estimating and testing minimum evolution trees. Molecular Biology and Evolution 9:945–967 (1992)
24. Saitou, N., Nei, M.: The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4:406–425 (1987)
25. Sneath, P.H.A., Sokal, R.R.: Numerical Taxonomy. Freeman, San Francisco (1973)
26. Wang, H., Xu, Z., Gao, L., Hao, B.: A fungal phylogeny based on 82 complete genomes using the composition vector method. BMC Evolutionary Biology 9(195):1–13 (2009)
27. Wolf, Y. I., Koonin, E. V.: A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. Genome Biol Evol. 4(12):1286–1294 (2012)

ISSN 1870-4069

Modeling Students' Dropout in Mexican Universities

Noel Enrique Rodríguez-Maya[1], Carlos Lara-Álvarez[2], Oscar May-Tzuc[3], Brian Alison Suárez-Carranza[1]

[1] Departamento de Sistemas y Computación, Instituto Tecnológico de Zitácuaro Michoacán, Mexico

[2] CONACYT Research Fellow - Centro de Investigación en Matemáticas (CIMAT), Zacatecas, Mexico

[3] Facultad de Ingeniería, Universidad Autónoma de Yucatán, Mérida, Yucatán, Mexico

nrodriguez@itzitacuaro.edu.mx, carlos.lara@cimat.mx, maytzuc@gmail.com, alison@hotmail.com

Abstract. Nowadays, the student dropout rate in Mexican higher education institutions have increased, affecting diverse life aspects such as economic, social, academic, and among others. The students dropout prediction is a challenge because of numerous interacting factors; an accurate prediction is useful for universities to implement strategies that reduce student failure –e.g., to implement tutorial action plans. The present research work proposes a model to predict dropout based on the student self-reported information and scores on the university entrance exam. A case of study to evaluate the proposed model was performed, results show a precision of about 86%. This model can serve to support decisions on strategies to reduce students' failure.

Keywords: students' dropout, predictive models, machine learning, data mining.

1 Introduction

The scholar dropout is a problem present in the majority of institutions of higher education in Mexico, affecting their terminal efficiency and academic performance. In addition, the scholar dropout has been increasing in the last years with repercussions in social, economic, and academic aspects. According to the National Association of Universities and Institutions of Higher Education (ANUIES), in Mexico, 25% of first-year students leave the university without completing their courses, that percentage increases to around 50% as they advance in their studies [1]. Students leave their studies for a variety of reasons

that may be personal, academic, physical, economic, and emotional in nature [2,3,4,5].

Although the problem is not new, it acquires special relevance in the so-called 'knowledge society', in which the knowledge is particularly important for the economic development of the nations, and people seek to improve their educational levels in order to be able to compete for better jobs and lifestyle [6].

Institutions of higher education in Mexico select new students based on entrance exams such as the EXANI II (National New Applications Exam for Universities) designed by the National Center of Evaluation for the Superior Education (CENEVAL, for its acronym in Spanish). This exam brings information about previous knowledge for the new possible students in predictive areas that represents the possible academic performance in the university.

In the last years, the amount of data has increased considerably. Data Mining (DM) has proven to be a powerful tool to identify interesting and hidden patterns through the search for existing relationships within the available data. Different objectives can be achieved with DM, among which prediction and description are the most important. Prediction is responsible for the creation of models that can approximate a set of given observations. In general, there are two predictive approaches: classification and regression. In both cases there are a set of predictive attributes and an objective attribute. The main difference is that in classification, the objective is discrete, while in the regression the objective is to learn a continuous function [7].

This paper presents a classification model to predict students' dropout in universities. The model is intended to accurately and promptly identify those students at risk of dropout, so that universities can generate plans and strategies for retention. The model uses a DM methodology to learn models based on decision trees. A case study was used with information from the 2010-2015 and 2011-2016 generational cohorts of the Technological Institute of Zitacuaro. The data collected include: qualifications (performance) of students, and results of the EXANI II, which contemplates different personal, economic, social, cultural and academic factors.

The model was generated with the integration of a dataset, with the most important attributes of the students (predictors) as well as their corresponding level of academic performance (objective). The task consisted of creating a classification model in which each tuple belongs to a class or category. Recently, the majority of works have proposed the use of different local academic attributes to predict the scholar dropout [18,19,20,21,22]. This contribution proposes the use of some attributes obtained from EXANI II to generate learning models that predict the drop out. On the other hand, the establishment of the academic performance is an important aspect to take into consideration. This work proposes to use the number of academic credits in an specific generational cohort as the performance of students. The dataset is composed by the sequences $\mathcal{M}$ and $\mathcal{P}$ represented by Eq. 1:

$$\mathcal{D} = \{(m_i, p_i) \mid i \in 1, \ldots, k\}, \tag{1}$$

where $m_i \in \mathcal{M} \subset \mathbb{R}^N$ are academic attributes from the EXANI II, $p_i \in \{false, true\}$ is the performance obtained by students in a specific generational cohort, and k is the number of instances (information of students).

The results yield a prediction model with accuracy greater than 86%. It is intended that this work can be easily generalized to be implemented in other universities. The generated models can be used as academic tool to prevent desertion and for the application of the necessary retention mechanisms since the selection of new students.

The rest of this paper is organized as follows: Section 2 presents the main related works; Section 3 presents the methodology and main used tools; Section 4 presents the Proposed Model; Section 5 shows the experiments and the main results; and the Section 6 presents the conclusions.

2 Related Work

The first aspect to be done in school dropout modeling is the discrimination of attributes that most affect this phenomenon. Rodríguez Echeverría et al. [8] performed an analysis on the characteristics of the admitted students to the Technological Institute of Sonora in 2002.

They shown the importance of the selection exam score and the performance obtained by the students in their previous degree. These parameters (qualification and performance) can serve as measures of academic behavior prediction in the university environment.

On the other hand, Artificial Neural Networks [14] have been proposed to predict scholar dropout at the University of Genova, Italy in the years 2008 and 2009. For the creation of the model, the authors used students' academic information from surveys, interviews, and students' characteristics before entering the university. The model was able to estimate those students with major scholar dropout incidence with an accuracy greater than 90%.

Scoot et al. [18] proposed the identification of the determinants of failure and subsequent dropout of students in the first year of a computer science course at Glasgow University. The information was collected through questionnaires, interviews, and data with academic students performance provided by the department of computer science. An analysis was made using different sets of data collected and DM techniques. It was concluded that the collected data were not sufficient to make a completely reliable prediction. However, interesting results were obtained that can serve as a point of reference to know the performance of the student and a possible identification of risk of scholar dropout.

Lam-On et al. [19] used the factors that would help to predict the students' performance by taking information such as pre-university characteristics, admission data and initial performance. Their proposed a method focuses on the transformation of data to a matrix where the relationships of the clustered sets were shown. It was hypothesized that multiple clusters of data may provide more useful information for classification. The results suggest that Principal Component Analysis leads to the most accurate prediction.

Márquez Vera et al. [20] collected information from three different sources: a survey designed to obtain personal and family data of the students, the results of the CENEVAL (EXANI I) exam and the scores obtained in the different subjects. The resulting information was 77 attributes of 670 students where 610 passed and 60 failed the course. At the end, a reduction was achieved to only 15 of the best attributes without losing the performance of the classification which was cost sensitive. The results showed an accurate classification in the minority class that is the one that matters in the research.

At the Eindhoven University of Technology in the Electrical Engineering career the scholar dropout rate was 40%, the students generally decide to dropout before the end of January, the goal of [21] was to build a model that predict the students' success, using as input data: information collected until the month of December of the 648 students of the institutional database, including pre-university and university data. Latter, different DM techniques were applied, such as Decision Trees, Bayesian Classifiers and Association Rules. The results indicate that an 80% accuracy was obtained with Decision Trees those proved to be better than other models applied.

Pal [22] applied an algorithm based on machine learning to analyze and extract information from the engineering students data of the Institute of Engineering and Technology of VBS Purvanchal University, Jaunpur, India. A predictive model was established with 1650 records, to identify the most likely students to dropout their studies. The collected data set was obtained through the enrollment form and includes demographic data, past performance data and personal data. Data were selected and transformed leaving only the fields required for DM. Before, four different classification techniques were applied which belong to the Decision Trees. Several experiments were carried out observing the resulting trees and the attributes with more relevance in the results. A precision percentage of more than 85% was obtained which is quite effective to reduce the school dropout rate.

In a previous work [25] we proposed experimentally a predictive model to determine the possible cases of scholar dropout. The model also is based on information collected from the selection exam and the academic performance of students. The main difference with this proposal is that in this research it is proposed a formal model for the creation of predictive models of students' dropout.

Martínez [13] developed a model based on descriptive techniques as Principal Component Analysis and Linear Discriminant Analysis, to detect possible cases of students that require tutoring to reduce the dropout at Technological Institute of Morelia in 2011. Using this tool, four profiles of students were formed: applicants, training, follow-up and closing, as well as three profiles of tutors: training, follow-up and closure. The results show that this tool is useful to generate precise profiles of students and tutors.

In [16] a model based on Chi-square Automatic Interaction Detector (CHAID) was constructed by means of which the most important predictor variables are determined in an Indonesian University. The results showed an accuracy greater

than 80%, determining the accuracy of models, the depth of the Decision Trees.

Many studies have used the DM approach to develop models in scholar dropout prediction [5,9,10,11,12]. Those models are based on Decision Trees and Artificial Neural Networks techniques using personal, economic and academic factors. The results reached precisions greater than 70%. Márquez [3] realized a DM process using information from the first period of students of a high school in the city of Zacatecas in 2012. The WEKA software was used to make the classification process: class balancing, attribute selection and the generation of models based on decision trees and rules of induction. The result was an early detection methodology for possible dropouts.

Cedano et al. [15] proposed a model to predict scholar dropout in an University at La Paz, Baja California Sur, Mexico based on the methodology Cross Industry Standard Process for Data Mining (CRISP-DM). They compared different techniques for model generation such as: Decision Trees, Artificial Neural Networks and Cluster k-means. The results showed precision values of 68% for the Decision Trees and k-means clusters, and 64% for the Artificial Neural Networks.

3 Methodology and Tools

This research was guided through the CRISP-DM methodology, that is conformed of six stages [24]:

1. *Business Understanding.* The aim of this stage is the comprehension of the project objectives from the point of view of the business perspective using a data mining scenario. The students dropout phenomenon was established as an institutional problem, therefore, it was considering all the possible variables (information from the selection exam and the academic performance).
2. *Understanding of data.* In this phase a data collection was performed, this by the understanding and identification of the main elements of the data that can be related with the problem; e.g., the establishment of the student performance.
3. *Data preparation.* Once the main data elements were determined, it begin with the construction of the dataset considering the raw data: this phase comprises a data cleaning and transformation, the selection of attributes, among other data activities.
4. *Modeling process.* In this phase various classifiers were selected according to the performance reported in literature and the selection of the proper software to make the experiments.
5. *Data Evaluation.* Once the most accurate model was selected, a data evaluation was performed; the model is examined in a depth way to determinate possible bias and to make assumptions about the model performance.
6. *Deployment.* Finally, the created model is deployed in a easy access platform to the final user; e.g., in a web platform.

The main tools used in this work are described below:

- *Data mining*: is a set of tools and techniques that allow the exploration of large information in order to find patterns that can explain the behavior of the data. The DM allows an in-depth analysis of data for the purpose to create models that allow, for example, to predict phenomena in the educational environment [4,5].
- *Weka Software*: the Weka platform is a collection of the state-of-the-art machine learning algorithms and data preprocessing tools [7]. It provides extensive support for the whole process of experimental DM that includes: preparing the input data, evaluating learning schemes statistically, and visualizing the input data and the outcome of learning. This diverse and comprehensive toolkit is accessed through a common interface. Therefore, its users can compare different methods and identify the most appropriate for the problem. Weka was developed at the University of Waikato in New Zealand. The system is written in Java and distributed under the terms of the GNU General Public License. It runs on almost any platform and has been tested under Linux, Windows, and Macintosh operating systems-even in a personal digital assistant.
- *Resample technique*: it is applied to a set of data to balance the number of patterns of each class defined in them [17]. Resample produces a random subsample of a dataset using either sampling with or without replacement. An example to configure it and that is of vital importance is to specify "NoReplacement" so that at the time of execution it does not return to select patterns that already have been selected. In Weka, the Resample technique adds instances to a class, this is done simply by adding instances of the class that has only a few instances multiple times in the result dataset. Thus, the resulting dataset is heavily biased in terms of a class for which only a few samples are available.
- *Random Forest*: this technique denotes an improved method for classification, by means of the creation of a large number of classifiers. Each one is constructed using a deterministic algorithm, the generated trees are different for two reasons; first, for each node, the best division is choosen from a random subset of predictors rather than the whole set of them. Secondly, each tree is constructed using a sample of the observations, approximately one-third of these are used to estimate the accuracy of prediction. Unlike other trees, in these there is no possibility of pruning or trimming [23].

4 Proposed Model

A supervised learning strategy is used to predict the student dropout. That is, from a set of training examples

$$\mathcal{D} = \{(m_i, p_i) \mid i \in 1, \ldots, k\}, \tag{2}$$

where $m_i \in \mathcal{M} \subset \mathbb{R}^N$ are academic attributes from the EXANI II; $p_i \in \{false, true\}$ is the performance obtained by students in a specific generational cohort, and k is the number of instances (information of students), we want to

learn a function F that maps a set of academic predictors $\mathcal{M}$ to indicators of student desertion I:

$$F : \mathcal{M} \rightarrow I, \tag{3}$$

where $I \in \{true, false\}$.

4.1 Predictive Variables

Every year CENEVAL defines the most representative variables to determinate the most appropriate students selection; i.e., some attributes can change. The EXANI II applications 2010 and 2011 contain some identical set of attributes, The first part of the model is the unification of predictive attributes taken from the EXANI II. The attributes for EXANI II 2010 and 2011 are[1]:

$EXANI{-}II\_2010$=$\{tipo\_exa$, opc_apli, ano_ver, $tipo\_reg$, $tipo\_resp$, cve_bpm, $apli$, $fecha\_apli$, cve_inst, $identifica$, $desc\_ident$, $lpos\_img$, $folio$, $matricula$, ape_pat, ape_mat, $nombre$, dia_nac, mes_nac, ano_nac, $sexo$, li_mad, li_pad, edo_proc, nom_proc, ciu_proc, cve_proc, reg_proc, mod_bac, $prom\_sec$, $prom\_bac$, exa_extr, dej_est, rec_beca, fal_dia, $porc\_ptar$, fre_tar, can_tar, cal_acl, cal_pun, cal_etar, $trab\_inv$, exa_dept, con_mrl, act_depo, act_cul, act_sal, hrs_trab, est_alca, no_lic, si_lic, si_pos, acc_con, acc_dud, acc_est, int_eje, fre_eje, ses_eje, act_pot, act_paa, act_ipf, fam_exa, pre_exa1, pre_exa2, apo_ami, apo_pro, apo_cur, apo_otr, uti_gui, $enlace$, $vive\_mad$, $vive\_pad$, $vive\_sup$, sup_cali, $trab\_mad$, $trab\_pad$, $esco\_mad$, $esco\_pad$, $cuan\_lib$, $cuan\_peli$, exp_pad, ser_per, ser_rev, ser_inte, ser_cabl, ser_luga, $cine$, $museo$, $espec$, con_lib, ex_ag, ma_pi, ser_tele, ser_lav, ser_ref, $bien\_mic$, ser_dvd, $bien\_pc$, ser_tv, ser_auto, pad_tarj, vac_rm, edo_rep, hab_cop, hab_elim, hab_vir, hab_ptex, hab_pres, hab_fbas, hab_int, hab_core, hab_baj, hab_adj, li_inter, li_taca, li_noti, pos_sel, $icne$, $percen$, $porcecne$, $pcne$, $prlm$, $pmat$, prv, $pesp$, $ptic$, $irlm$, $imat$, irv, $iesp$, $itic\}$.

$EXANI{-}II\_2011$=$\{tipo\_exa$, opc_apli, ano_ver, $tipo\_reg$, $tipo\_resp$, cve_bpm, $apli$, $fecha\_apli$, cve_inst, $identifica$, $desc\_ident$, $lpos\_img$, $folio$, $matricula$, ape_pat, ape_mat, $nombre$, dia_nac, mes_nac, ano_nac, $sexo$, con_dce, con_impe, con_esc, con_ver, con_len, li_mad, li_pad, edo_proc, nom_proc, ciu_proc, cve_proc, reg_proc, ano_bac, mod_bac, $prom\_bac$, bec_bdac, bec_bne, bec_bhd, por_cce, por_obap, por_bib, ppr_acm, ppr_eje, ppr_tar, ppr_pun, ppr_dud, ppr_asi, fev_eno, fev_rude, fev_ties, fev_tefe, niv_idat, niv_erel, niv_ride, niv_rpun, cua_lib, niv_coh, niv_err, niv_resl, niv_ensa, niv_cart, niv_repo, niv_doco, niv_exp, niv_duda, niv_deba, hrs_trab, est_alca, fam_exa, pre_exa1, pre_exa2, $vive\_mad$, $vive\_pad$, $trab\_mad$, $trab\_pad$, $esco\_mad$, $esco\_pad$, $cuan\_lib$, $cuan\_peli$, exp_pad, ser_tele, ser_lav, ser_ref, ser_hor, ser_inte, ser_cabl, ser_dvd, $bien\_pc$, ser_tv, ser_auto, ser_bano, $cine$, $museo$, $espec$, vac_rm, edo_rep, li_inter, li_taca, li_noti, hab_ptex, hab_pres, hab_fbas, hab_baj, pos_sel, $icne$, $percen$, $porcecne$, $pcne$, $prlm$, $pmat$, prv, $pesp$, $ptic$, $irlm$, $imat$, irv, $iesp$, $itic$.

To homogenize the information it is necessary a process to intersect the common attributes for both applications (2010 and 2011). The intersection set

[1] For more information http://www.ceneval.edu.mx/exani-ii

is defined by Eq. 4:

$$EXANI-II = EXANI-II\_2010 \cap EXANI-II\_2011. \tag{4}$$

Once the $EXANI-II$ set is established, it is necessary a process to determine the most representative attributes; i.e., the selection of the most correlated attributes with the objective.

4.2 Objective Attribute

The objective value is determined from the cumulative academic efficiency of student s_i after ten semesters, expressed as the ratio:

$$c(s_i) = \frac{\text{number of credits earned by } s_i}{\text{credits required for graduation}}, \tag{5}$$

therefore, the performance of student s_i is categorized according to

$$p_i = p(s_i) = \begin{cases} \text{false} & \text{if } c(s_i) \geq 1, \\ \text{true} & \text{otherwise,} \end{cases} \tag{6}$$

where 'true' represents a possible dropout, and 'false' the completion of the academic program.

5 Experiments and Results

Experiments were performed in the Technological Institute of Zitacuaro (ITZ for its Spanish abbreviation). ITZ is a superior educational institute with a technological orientation, placed at the east of the Michoacan State in Mexico. Actually, the ITZ offers nine bachelors degrees, increasing the number of students yearly in a constant way. In the last period, more than 2000 students signed up in the different academic programs; hence, it requires a complex academic and administrative planning. In the generational cohorts 2010-2015 and 2011-2016 were reported 16.2% and 19% of desertion, respectively, which affects directly the students' success rates.

Dataset $\mathcal{D}$ is composed by sequences $\mathcal{M}$ defined by the set $EXANI-II$ (Eq. 4), and the objective values $\mathcal{P}$ that correspond to performance values defined by Eq. 6. Data was collected in a separated by commas file (CSV) containing personal and academic information, and EXANI II results of ITZ students. A total of 671 records, from the generational cohorts 2010-2015 and 2011-2016, were prepared and processed. The dataset was structured in a two dimensional matrix containing the EXANI II information and its corresponding success or fail attribute for each student. The attributes are[2]: *cine*, *pcne*, *mes_nac*, *ser_inte*, *li_mad*, *nom_proc*, *fam_exa*, *hab_baj*, *tipo_resp*, *icne*, *ano_nac*, *cuan_peli*, *matricula*,

[2] The complete description of attributes can be found in: http://ceneval.edu.mx/web/guest/exani-ii

Table 1. Correlation of the selected attributes with the student performance.

Attribute	Correlation	Attribute	Correlation
prom_bac	0.2626275	*ptic*	0.0960172
percen	0.2028458	*ser_cabl*	0.0944656
porcecne	0.1573105	*itic*	0.0935362
prlm	0.1541754	*ser_tv*	0.0828841
pos_sel	0.1516624	*bien_pc*	0.0800759
prv	0.1497613	hab_fbas	0.0783265
irv	0.1330568	li_taca	0.0776661
irlm	0.1329894	hab_ptex	0.0757085
icne	0.1307960	$identifica$	0.0710744
pmat	0.1275899	$folio$	0.0703989
pesp	0.1194576	ser_dvd	0.0698548
imat	0.1165091	ser_lav	0.0657144
edo_rep	0.1089331	li_noti	0.0640612
li_pad	0.0992314	ser_auto	0.0585061
ser_inte	0.0987749	vac_rm	0.0552041
...	...	...	...

$esco\_pad$, prv, edo_proc, $tipo\_exa$, exp_pad, $folio$, $tipo\_reg$, $desc\_ident$, $prom\_bac$, $ptic$, li_noti, $esco\_mad$, irv, ser_lav, $pmat$, hrs_trab, $museo$, est_alca, $cuan\_lib$, $irlm$, $vive\_mad$, $espec$, hab_pres, pre_exa1, $percen$, $trab\_mad$, hab_fbas, cve_bpm, ape_pat, pos_sel, ser_tv, ser_cabl, reg_proc, $sexo$, edo_rep, dia_nac, $nombre$, vac_rm, $imat$, ciu_proc, $iesp$, mod_bac, opc_apli, ser_dvd, $porcecne$, cve_inst, ser_auto, ano_ver, pre_exa2, $trab\_pad$, $itic$, $apli$, li_inter, ser_ref, $pesp$, $bien\_pc$, li_taca, cve_proc, $vive\_pad$, $lpos\_img$, $prlm$, ser_tele, li_pad, hab_ptex, ape_mat, $identifica$, $fecha\_apli$, $aluctr$, $performance$.

The next step is to determine those attributes most correlated with the objective attribute; i.e., performance (Eq. 5). Table 1 shows the correlation between each attribute and the performance, and Table 2 shows the selected attributes.

Attributes with a correlation greater or equal than 0.08 were selected; then, the dataset is composed of the following set:

$$\begin{aligned} EXANI-II = \{&prom\_bac, percen, porcecne, prlm, pos\_sel, prv, irv, irlm, icne,\\ &pmat, pesp, imat, edo\_rep, li\_pad, ser\_inte, ptic, ser\_cabl, itic,\\ &ser\_tv, bien\_pc, performance\} \end{aligned}$$

These attributes are related to students' score in different subjects such as: mathematics, verbal and logical reasoning, and Spanish; but it also includes cultural and economic aspects such as: number of mexican states that the student has visited, if students' parents speak an indigenous dialect, if student has access to computer, internet, and cable, among others. It is clear that the most important variables to predict the university studies success are those related with the basic subjects, technologies and cultural aspects.

Table 2. Description of the most correlated attributes.

Attribute	Description	Attribute	Description
prom_bac	Mean of qualification of high school	*pesp*	Percent of qualification of spanish
percen	Percentile of the selection exam	*imat*	Qualification of mathematics (ceneval index)
porcecne	% > CNE of the selection exam	*edo_rep*	Visited states (México)
prlm	Percent of qualification of mathematical logical reasoning	*li_pad*	Father speaks some indigenous dialect
pos_sel	Reached position for the student in the exam	*ser_inte*	Internet availability (in home)
prv	Qualification in verbal reasoning in percent	*ptic*	Qualification in Technologies of information and communications in percent
irv	Qualification in verbal reasoning (CENEVAL index)	*ser_cabl*	Availability of pay cable service (in home)
irlm	Qualification of mathematics logical reasoning (CENEVAL index)	*itic*	Qualification in Technologies of information and communications (CENEVAL index)
icne	CENEVAL index qualification in the exam of selection	*ser_tv*	Number of televisions in home
pmat	Qualification in Mathematics (CENEVAL index)	*bien_pc*	Number of computers in home

Table 3. Performance of different classifiers.

Classifier	Correctly Classified	Precision	Recall
Naive Bayes	63.48%	0.628	0.635
Multilayer Perceptron	81.67%	0.816	0.817
J48	80.18%	0.801	0.802
Random Forest	**86.14**%	**0.862**	**0.861**
Random Tree	83.61%	0.836	0.836

As shown in Fig. 1, data is composed of 274 and 397 instances for the $true$ (dropout) and $false$ (success) classes, respectively. In this sense, data has more instances for the class $false$ than $true$; hence, it is necessary a balancing process to ensure the most distributed patterns into the dataset. In this case, the re-sample tool provided by WEKA was used (more details in Section 3). Table 3 shows the performance of different classifiers. The generated models were validated using cross-validation with ten folds, and the WEKA classifiers used its default parameters.

The models based on decision trees obtained the most accurate results, while the model based on Multilayer Perceptron obtained the third best result, and finally the model based on Naive Bayes obtained the worst performance. On the other hand, Table 4 presents the confusion matrix for the Random Forest model (the most accurate model).

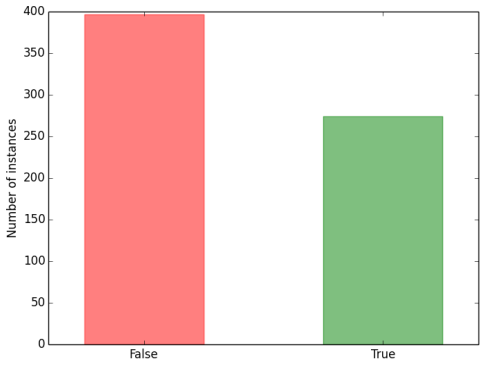


Fig. 1. Distribution of classes.

Table 4. Confusion matrix of model based on Random Forest.

true	**false**	
212	62	**true**
31	366	**false**

The model classifies the *true* instances with an accuracy of 77.4% while the *false* instances with an accuracy of 92.2%. It is clear that the model presents major problems when classify the possible students to dropout their studies.

6 Conclusions

This research presents a predictive model based on a classification task. The phase of knowledge extraction is based on a DM process, the Cross Industry Standard Process for Data Mining (CRISP-DM). The WEKA software was used to support the experiments. A model based on decision trees was the most accurate to predict student dropout in Mexican universities. The proposed model uses information obtained from the admission exam (EXANI II) and students' performance. The model was probed on a study case, particularly the Technological Institute of Zitacuaro in the generational cohorts 2010-2015 and 2011-2016. Results showed an accurate model with more than 86% of precision.

As a future work, the first step is the implementation of the model in a web based interface to support different mobile devices. Next, the prototipe will be implemented in a pilot proof-of-concept to determinate the success or possible dropout of students. The model will be completed with the academic information generated every period, and finally, the model will be implemented as a support to the tutorial works to guide the new ITZ students to increment the student terminal rates and the terminal efficiency (degree acquisition).

References

1. ANUIES: Desertion, lag and terminal efficiency in Higher Education Institutions. Methodological proposal for its study. In: Collection Library of Higher Education, Research Series. Mexico: National Association of Universities and Institutions of Higher Education (2002)
2. Magaña Hernández, M.: Causes of School Failure. In: XIII Congress of the Spanish Society of Adolescent Medicine, Spain (2002)
3. Márquez Vera, C.: Prediction of failure and dropout through Data mining techniques. University of Cordoba (2015)
4. Timarán Pereira, R., Jiménez Toledo, J.: Pattern Detection Student Dropout in Undergraduate Programs of Institutions of Higher Education with CRISP-DM. In: Iberoamerican Congress of Science, Technology, Innovation and Education, Buenos Aires, Argentina (2014)
5. Amaya, Y., Barrientos, E., Heredia, D.: Student Dropout Predictive Model Using Data Mining Techniques. IEEE Latin America Transactions, Vol. 13, No. 9 (2015)
6. González González, M. T.: Absenteeism and Dropout: A Single Situation of Educational Exclusion. Electronic Magazine Iberoamerican on Quality, Efficiency and Change in Education, 4: 1–15 (2006)
7. Witten, I. H., Frank, E., Hall, M. A.: Data Mining: Practical Machine Learning Tools and Techniques. Third Edition, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2011)
8. Rodríguez Echeverría, M., Nereida Aceves López J.: Performance of university students and their relation to the average and Preparatory Course: ITSON case. In: VI International Congress of Organizational Analysis (2008)
9. Martinho, V. R. C., Nunes, C., Minussi, C. R.: A New Method for Prediction of School Dropout Risk Group Using Neural Network Fuzzy- ARTMAP. In: International Conference on Artificial Intelligence 2013 - ICAI'13, 2013, Las Vegas, USA, v. 1., pp. 359–365 (2013)
10. Tejas, S., Swarna, A. and Mathey, R.: Classification with Wekatool for Predicting Student Failure. International Journal of Engineering, Science and Computing, pp. 874–877 (2014)
11. Harilatha, U., Sudhakaryadav, N.: Predicting Educational Performance of a Student Failure and Dropout by using Data mining Techniques. (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5(6) (2014)
12. Mendiola, J. L. A., Rosas, R. M. V., Velázquez, J. A. A., Eleuterio, R. A., Marcial-Romero, J. R.: Analysis of school dropout with mining data. Research in Computing Science, 93: 71–82 (2015)
13. Martínez Avila, G. I.: Intelligent Tutoring Model (MIT) in the Technological Institute of Morelia. Autonomous University of Queretaro, Faculty of Information Technology (2015)
14. Siri, D.: Predicting Students' Dropout at University Using Artificial Neural Networks. Italian Journal of Sociology of Education, 7(2), 225– 247 (2015)
15. Cedano, J. A. H., Castro, J. A.: Model of data mining for Identification of patterns that influence the academic. Thesis, La Paz, Baja California Sur, Mexico (2015)
16. Novita, R., Sabariah, M. K., Effendy, V.: Identifying factors that influence student failure rate using Exhaustive CHAID (Chi-square automatic interaction detection). In: Information and Communication Technology (ICoICT), 2015 3rd International Conference on, Nusa Dua, pp. 482–487 (2015)

17. Reyes, J. I., Artagaveytia, F.: Introduction to pattern recognition 2015: Classification of forests by cartographic information. Spain (2002)
18. Scoot, J., Graal, M.: Student Failure in First Year Modules in the Biosciences: An Interview Based Investigation. Bioscience Education (2007)
19. Lam-On, N., Boongoen, T.: Using cluster ensemble to improve classification of student dropout in Thai university. In: Soft Computing and Intelligent Systems (SCIS), 2014 Joint 7th International Conference on and Advanced Intelligent Systems (ISIS), 15th International Symposium on, Kitakyushu, pp. 452–457 (2014)
20. Marquez Vera, C., Romero, C., Ventura, S.: Predicting School Failure Using Data Mining. Zacatecas (2010)
21. Dekker, G., Pechenizkiy, M., VleeShouwers, J.: Predicting Students Dropout: A case study. Zacatecas (2010)
22. Pal, S.: Mining Educational Data to Reduce Dropout Rates of Engineering Students. I.J. Information Engineering and Electronic Business (2012)
23. Pizzuti, C., Ritchie, M. D., Giacobini, M.: Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics. In: 8th European Conference, EvoBIO, Instambul, Turkey (2010)
24. Shearer, C.: The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, Vol. 5, No. 4 (2000)
25. Rodríguez-Maya, N.E., Jiménez-Alfaro, A.J., Reyes-Hernández, L.Á., Suárez-Carranza,B.A., Ruiz-Garduño, J.K. Minería de Datos: Modelo Predictivo de Deserción Escolar. En: congreso IEEE Mexican Humanitarian Technology Conference (MHTC) (2017)

ISSN 1870-4069

Minería de datos aplicada para la identificación de factores de riesgo en alumnos

A. Reyes-Nava[1,3], Allan Flores-Fuentes[1], R. Alejo[2], E. Rendón-Lara[2]

[1] Universidad Autónoma del Estado de México[1],
Estado de México, México

[2] Instituto Tecnológico de Toluca, Estado de México,
México

[3] Tecnológico de Estudios Superiores de Jocotitlán,
Estado de México, México

adriananava0@gmail.com, allan_fflores@yahoo.com.mx, ralejoll@hotmail.com,
erendonl@toluca.tecnm.mx

Resumen. En este trabajo se presenta la automatización del Sistema Integral de Tutoría del Tecnológico de Estudios Superiores de Jocotitlán para el área de Ingeniería en Sistemas Computacionales, el objetivo es la implementación del sistema en una plataforma digital y realizar un procesamiento en línea para que sea más eficiente el proceso de tutorías. El sistema consiste en la aplicación de cuestionarios y entrevistas a alumnos, obtenidos del manual del tutor del Sistema Nacional de Institutos Tecnológicos. La información obtenida es procesada en el software Weka aplicando técnicas de minería de datos, siguiendo la metodología Proceso de Extracción del Conocimiento con el propósito de facilitar la interpretación de resultados y encontrar conocimiento útil, acerca de los alumnos. Se analizan y discuten los resultados parciales del procesamiento a través de minería de datos de la información de generaciones del 2008 al 2014 de la carrera de Ingeniería en Sistemas Computacionales, encontrando relaciones entre aspectos familiares, económicos y académicos que indican un probable riesgo de deserción en alumnos con comportamientos comunes, estas relaciones sirven para proporcionar información suficiente y apropiada en la toma de decisiones en el tratamiento de la deserción escolar.

Palabras clave: proceso de extracción del conocimiento, tutoría, toma de decisiones, minería de datos.

Data Mining Applied for the Identification of Risk Factors in Students

Abstract. In this paper presents the automation of the Tutoring System of the Technological Higher Education of Jocotitlán for the area of Engineering in

Computational Systems, the objective is the implementation of the system in a digital platform and perform online processing to be more efficient the process of tutorials. The system consists in the application of questionnaires and interviews to students, obtained from the tutor manual of the National System of Technological Institutes. The information obtained is processed in the Weka software by applying data mining techniques, following the Knowledge Extraction Process methodology in order to facilitate the interpretation of results and find useful knowledge about the students. We analyze and discuss the partial results of the processing through data mining of the information of generations from 2008 to 2014 of the career of Engineering in Computer Systems, finding relations between family, economic and academic aspects that indicate a probable risk of desertion in Students with common behaviors, these relationships serve to provide sufficient and appropriate information in decision-making in the treatment of dropout.

Keywords: knowledge extraction process, tutoring, decision making and data mining.

1. Introducción

La Tutoría es un proceso a través del cual se busca acompañar, asesorar, aconsejar y apoyar al estudiante para lograr en él una educación de calidad, integral y acorde a las necesidades de su entorno académico, laboral o personal [1]. En el año de 1998 la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO) señaló la necesidad de modificar el proceso de enseñanza aprendizaje para dar respuesta a las necesidades del siglo XXI. Por su parte, en 2001 la Asociación Nacional de Universidades e Instituciones de Educación Superior (ANUIES) asume como eje rector para el mejoramiento de la educación la formación integrada, en la cual la tutoría tiene un papel distintivo. Asimismo, el plan de desarrollo Nacional 2013-2018 fija a la tutoría como medio para lograr disminuir la deserción escolar e incrementar los índices de eficiencia terminal. En ese mismo sentido, en los planes de desarrollo estatal e institucional del Tecnológico de Estudios Superiores de Jocotitlán se fija como meta el incremento de la eficiencia terminal.

El programa Institucional de Tutoría se implantó por primera vez en el Tecnológico de Estudios Superiores de Jocotitlán (TESJo) en el año 2008 y se operó por medio de los Jefes de División de cada carrera de acuerdo con los lineamientos establecidos por el Tecnológico Nacional de México (TECNM). En el 2010 se hizo un cambio, donde ya no eran los Jefes de División los encargados de la tutoría, sino los maestros. Para el 2012 se presentó el Manual del tutor, con la finalidad de que se convierta en una herramienta para el desarrollo de las actividades tutoriales.

A principios de 2014 el laboratorio de Minería de Datos y Reconocimiento de Patrones tomó como proyecto el estudio de la deserción escolar a través de técnicas de minería de datos. Para ello fue necesaria la construcción de una plataforma digital, en la cual, se pudiera desarrollar en línea el proceso de tutoría de acuerdo con lo establecido en el manual del tutor y adicionalmente se incorporaron algoritmos de minería de datos para el estudio de la deserción escolar. Actualmente, en el TESJo el

proceso de tutoría se rige con los lineamientos del Manual del Tutor del TECNM y es un proceso tradicional, en el sentido de que no se usan medios electrónicos o informáticos para las tareas de la tutoría. Sin embargo, en la carrera de Ingeniería en Sistemas Computacionales (ISC) se desarrolló un Sistema Integral de Tutoría del TESJo (SITTESJo), siendo la única carrera en llevar el proceso de tutoría de manera automatizada. Además se utilizan técnicas de minería de datos y reconocimiento de patrones para encontrar cuales son los factores que causan la deserción escolar.

2. Trabajo relacionado

2.2. Sistema de tutorías

El desarrollo de un sistema de tutorías en línea dentro del TESJo se realiza con la finalidad de hacer más eficiente el proceso de entrevista y recolección de información entre tutor y tutorado; de manera que al recabar la información de los alumnos pueda ser analizada para tener una perspectiva de quienes se encuentran en probable riesgo de deserción, fracaso estudiantil o inclusive saber quiénes son los alumnos destacados. Una vez que se obtienen los resultados de este análisis, la información es proporcionada a los tutores de cada grupo para emprender acciones tutoriales en su apoyo, con la finalidad de atacar a las problemáticas a las cuales se enfrenta la carrera de ISC como, la deserción escolar, inasistencias a clases, bajo rendimiento académico, entre otras.

Actualmente, en diferentes universidades se han desarrollado plataformas digitales para los procesos de tutoría. Por ejemplo, la Universidad Autónoma del Estado de México desarrolló el Sistema para la Tutoría Académica, en el cual se brinda el servicio de tutoría como apoyo al estudiante a través de un profesor del organismo académico que funge como tutor, se encarga de brindar el apoyo necesario y asesorar en los asuntos académicos como lección de unidades de aprendizaje, balance de créditos, canalización con asesores disciplinarios, entre otras actividades. Esto se lleva a cabo vía internet donde tutor y tutorado pueden interactuar para compartir información, analizar datos de diferentes tipos y formar un registro de las acciones emprendidas en su apoyo [2].

Por otro lado, [3] desarrolló un sistema de tutoría en línea para estudiantes de la carrera de ingeniería industrial de la Universidad del Valle de Atemajac, que surge por la necesidad de brindar una mejor educación a sus alumnos, tomando en cuenta los porcentajes de deserción y reprobación de las materias. Con la implementación del programa de tutoría se apoya a los alumnos desde el ingreso hasta el término de sus estudios haciendo uso del Programa General de Tutorías, asignando un tutor de apoyo desde el inicio de carrera y hasta el término de la misma.

También en [4], se hace una comparación entre los procesos de tutoría presencial y virtual en el Instituto Tecnológico de Sonora, donde los alumnos que están en la modalidad presencial, estadísticamente, tienen un promedio de calificaciones más bajo comparado con los que se encuentran en la modalidad virtual, lo cual hace notar que los alumnos prefieren un sistema de tutoría en línea, que un proceso tradicional.

No obstante, a pesar de que el problema de la tutoría se ha abordado de manera natural a través de herramientas digitales, la mayoría de los sistemas de tutoría son "tradicionales", es decir, plataformas en las cuales se interactúa con el alumno a través

de un portal de internet, pero que no van más allá. En este trabajo se presenta al SITTESJo como una alternativa a los sistemas convencionales de tutoría. El SITTESJo tiene como característica principal la inclusión de algoritmos de minería de datos que permiten el análisis de los datos y su posible aprovechamiento para mejorar la relación entre tutelados y tutores, y de esta forma cumplir con uno de los principales retos de la tutoría, reducir la deserción escolar e incrementar los índices de eficiencia terminal.

2.2. Aplicaciones de la minería de datos para el análisis de la deserción escolar

Actualmente en las instituciones existen diversos problemas que afectan el desarrollo de la educación sin importar el nivel educativo que se curse. De manera específica uno de los principales problemas a los que se enfrenta el sistema educativo nacional es el abandono de los estudiantes en algún punto durante el trascurso de sus estudios [5]. Pueden existir diferentes factores por los cuales los alumnos deciden abandonar sus estudios antes de concluir y pueden depender de diferentes circunstancias, por ejemplo, el ambiente social, personal y económico del alumno.

La deserción escolar es un problema que existe en las instituciones educativas desde hace mucho tiempo. En [6] se describe a la deserción como el abandono de las labores de un estudiante, dejando de cumplir con sus obligaciones como estudiante inscrito en la institución, así como no asistir a clases de manera regular, lo cual afecta los índices de eficiencia terminal.

Hasta el día de hoy se han realizado diferentes estudios, no solo en México, sino también en diferentes países, donde se presenta el problema de la deserción escolar. El objetivo principal de estas investigaciones es conocer las causas que lo originan e identificar cómo se puede resolver o tratar.

En este contexto [7, 8] presentan investigaciones realizadas en las universidades nacionales de Argentina donde se busca explicar la causa de la deserción, haciendo una relación entre el rendimiento académico, estilos de aprendizaje y competencia espacial, esta se refiere a la capacidad de generar, recordar y transformar información simbólica mentalmente. Aplican metodologías descriptivas y estadísticas cualitativas para encontrar los factores que afectan el rendimiento académico y pueden ser causa de abandono de los estudios. Las conclusiones en estas investigaciones, muestran que se necesitan promover conductas académicas en los alumnos para que obtengan un aprendizaje autorregulado de manera que se favorezcan sus capacidades intelectuales. Por otro lado, se pudo determinar que los factores individuales como el ambiente social y cultural del alumno pueden afectar el desempeño académico.

Por ejemplo en [9] se hizo un estudio tomando como fuente interna de información la base de datos de control de admisiones y registro académico de la institución y como fuente externa se seleccionó información del Departamento Administrativo Nacional de Estadística del Sistema para la Prevención de la Deserción en la Educación Superior. De la información obtenida de estas bases de datos, se realizó el análisis mediante clustering para obtener los patrones de comportamiento comunes que son causa del abandono de los estudios, encontrando así que las principales causas son las bajas calificaciones en las asignaturas de los primeros semestres de la carrera, el pago de la colegiatura y la zona de procedencia del estudiante. A partir de los resultados que se

obtuvieron en esta investigación se pueden formular políticas y estrategias para incrementar los índices de retención estudiantil.

En [10] se hizo un análisis con información obtenida de estudiantes desertores, éste se realizó mediante proyecciones capaces de ordenar datos nominales y numéricos de manera supervisada, con el fin de encontrar los atributos más relevantes para ser utilizados en la aplicación de algoritmos de minería de datos, obteniéndose así resultados claros que permitan la toma de decisiones en favor de la disminución de la deserción.

Con base en lo anterior se puede concluir, que el uso de tecnologías emergentes como la minería de datos para identificar a los alumnos desertores o con problemas académicos en las instituciones, ha sido tema de estudio para diferentes instituciones educativas.

3. Elementos que constituyen el proceso de tutoría del SITTESJo

El SITTESJo consiste de una serie de cuestionarios que el alumno debe contestar en forma manual, los cuales son establecidos en el manual del tutor del Sistema Nacional de Institutos Tecnológicos (SNIT) para el proceso de tutoría. De manera interna el TESJo asigna a algunos de ellos códigos específicos (FO-TESJo-039, FO-TESJo-040), mientras que a otros no los considera en el proceso formal de tutoría. No obstante, el SITTESJo incluye todos los cuestionarios y pruebas propuestas en el Manual del Tutor del SNIT, e inclusive incluye un módulo de calificaciones. A continuación, se describen brevemente los cuestionarios y pruebas incluidas en el SITTESJo.

1. FO-TESJo-039: Aquí se incluyen preguntas sobre aspectos familiares, socioeconómicos y académicos de los alumnos, a manera de conocer sus antecedentes; esta información se almacena en la base de datos del sistema y puede ser modificada en cada aplicación de los cuestionarios si es que existe algún cambio en la situación del alumno [11].
2. FO-TESJO-040: Este cuestionario incluye preguntas para conocer de manera específica si el alumno trabaja y en qué, quién y cómo solventa sus gastos, a quién mantiene informado de sus estudios y su estado de salud de manera general [11].
3. Test de autoestima, de asertividad, de estilos de aprendizaje y de habilidades de estudio, éstos son cuestionarios de preguntas cerradas para reconocer problemas referentes a organización de actividades, técnicas para sus estudios y su motivación a estudiar [11].
4. Calificaciones. De manera opcional se agregó el apartado de calificaciones donde los alumnos capturan la calificación obtenida por cada unidad de aprendizaje de los diferentes programas educativos que cursan.

4. Minería de datos y algunos clasificadores

Las bases de datos que existen actualmente contienen una gran cantidad de información que en muchas ocasiones no es aprovechada, porque no existe suficiente personal que se encargue de realizar el análisis de los datos. No obstante, en las últimas décadas se han popularizado nuevas formas de analizar estos datos y obtener información en forma de conocimiento; por medio del empleo de herramientas computacionales, es decir, a través del uso de la minería de datos. De acuerdo con [12] la minería de datos se puede definir como "el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos". La tarea fundamental de la minería de datos es, encontrar modelos entendibles y legibles para el usuario a partir de un conjunto de datos.

Otros autores como [13] definen de manera simple a la minería de datos como "el proceso de convertir datos en conocimiento". Esta definición es la que se puede comprender de manera más sencilla y enfoca el punto central de la minería de datos.

Para poder encontrar conocimiento a partir de un conjunto de datos es necesario tomar en cuenta el proceso de descubrimiento de conocimientos en bases de datos, en este caso se aplica KDD (Knowledge Discovery in Databases). Este es un proceso enfocado a identificar patrones válidos, potencialmente útiles y comprensibles a partir de un conjunto de datos, este proceso incluye no solo la obtención de patrones, sino también se realiza la interpretación de los resultados de manera que apoye a la toma de decisiones [14].

El proceso de descubrimiento de conocimiento en bases de datos consta de diferentes etapas, en la Figura 1 se muestra de manera esquemática cómo se lleva a cabo este proceso. Se parte desde un sistema de información de donde se obtienen los datos a ser analizados, los cuales pasaran por el proceso de preparación, aplicación de técnicas y algoritmos de minería de datos, obtención de patrones y evaluación o interpretación de los mismos para llegar a la obtención del conocimiento que ayuda a la toma de decisiones en el proceso de negocios.

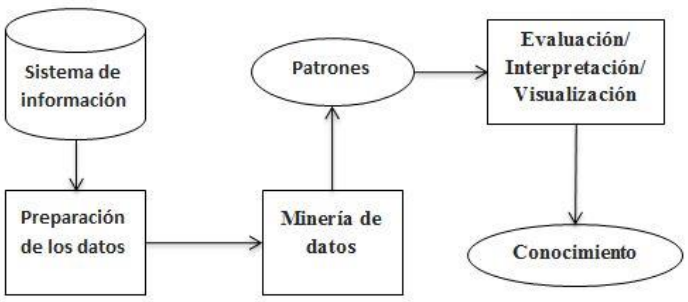


Fig. 1. Proceso KDD.

En el proceso de descubrimiento de conocimiento se identifica la etapa de minería de datos, donde se eligen los algoritmos a utilizar de acuerdo con la información con la que se cuenta y al conocimiento que se quiera obtener.

Un ejemplo de ello son los métodos bayesianos, los cuales se centran en el estudio descriptivo y predictivo, porque los algoritmos se enfocan en el descubrimiento de relaciones de independencia y relevancia entre las variables. Los métodos bayesianos se usan para realizar inferencias a partir de los datos y a partir de modelos probabilísticos usados para formular hipótesis [15].

Otro ejemplo son las redes neuronales artificiales, estas pueden ser usadas en aprendizaje supervisado o no supervisado. Una red neuronal artificial es en sí un método de aprendizaje cuya finalidad es la de simular procesos biológicos de información, las cuales parten de la capacidad del ser humano para procesar información [13].

4.4. Reglas de asociación

Las reglas de asociación son técnicas empleadas en el aprendizaje no supervisado para establecer posibles relaciones entre diferentes acciones aparentemente independientes entre sí, reconociendo como pueden ocurrir los eventos o acciones a partir de la aparición de otros. Así, este tipo de técnicas se usa cuando el análisis que se desea realizar es exploratorio [16].

Existen diversos algoritmos para encontrar reglas de asociación, el más utilizado es el algoritmo Apriori, el cual "se basa en la búsqueda de los conjuntos de datos con determinada cobertura" [13]. Este algoritmo es el empleado para este caso de estudio, a continuación se presenta el pseudocódigo para determinar las reglas.

> Paso 1. Determinar la confianza y cobertura.
> Paso 2. Se construyen los conjuntos formados por un solo item (conjunto) que superan la cobertura mínima.
> Paso 3. Extraer del conjunto de reglas las que tengan un nivel de confianza mínimo.

A partir de lo anterior, el algoritmo queda de la siguiente manera:

```
ALGORITMO Apriori (D: datos, MinC: cobertura mínima)
        i=0
        Rellena_Item(Ci)
        MIENTRAS Ci =! 0
                PARA_CADA X=elemento de Ci
                        SI  Cobertura(X)>=  MinC  ENTONCES
                                        Li=Li * X
                FIN_PARA
        Ci+1=Selecciona_Candidatos(Li)
        i=i+1
        FIN_MIENTRAS
        RETORNA C
FIN_ALGORITMO
```

4.2. Clustering

El clustering o agrupamiento es una técnica de minería de datos, la cual consiste en formar grupos, haciendo una división de los datos en grupos de objetos similares con base en un criterio de cercanía, aumentando la similitud de los elementos de un mismo grupo. Uno de los algoritmos de clustering o agrupamiento más común es el propuesto por MacQueen, llamado Simple K-Means, donde K determina el número de clusters o grupos que se desean obtener [17].

El algoritmo K-means inicia seleccionando o calculando los centroides iniciales, dependiendo del criterio de selección de centroides, posteriormente asigna los objetos a su centroide más cercano, para después recalcular los nuevos centroides esto lo realiza hasta que el algoritmo converge. El pseudocódigo se presenta a continuación:

Paso 1. Selecciona los K centroides iniciales.

Paso 2. Asigna los objetos X_i del conjunto de datos X, a su centroide más cercano.

Paso 3. Re-calcula los nuevos centros, regresa al paso 2, hasta que el algoritmo converge.

5. Metodología para el análisis de los datos almacenados por el SITTESJo

El análisis de los datos fue realizado a través de dos importantes herramientas de minería de datos: clustering y reglas de asociación; en particular, se utilizó el algoritmo de agrupamiento K-means y el algoritmo Apriori para las reglas de asociación. Para llevar a cabo el proceso de análisis de los datos se empleó la metodología KDD, la cual dentro de este trabajo consistió en:

1. Seleccionar los datos y análisis de sus propiedades. Una vez aplicados los cuestionarios a los alumnos y que la información está en la base de datos, se determina cuáles son los datos relevantes para hacer el análisis de la información disponible como antecedentes académicos, familiares y socio-económicos.

En este paso para reglas de asociación de utilizaron datos específicos de la base de datos, los cuales fueron: práctica de deporte, problemas económicos, si trabajan o no, promedio de preparatoria, interrupciones en sus estudios y el campo de acción de la carrera.

Para el caso de agrupamiento se emplearon los datos del porqué eligió la carrera, tratamientos médicos, dependencias económicas (si están casados o tienen hijos) y conocimiento de programas como la becas.

2. Pre-procesamiento de datos. En esta etapa se hace la limpieza de la base de datos, encontrando los datos inconsistentes, fuera de rango, faltantes o campos vacíos para ser integrados posteriormente y puedan ser utilizados en el análisis con

minería de datos, en esta etapa se convierte el archivo en formato .arff, el cual es el aceptado para el análisis con weka, este es un software diseñado para el análisis de bases de datos, aplicando diferentes algoritmos tanto de aprendizaje supervisado como no supervisado para obtener estadísticas y patrones de tendencia según sea lo que se estudie.

3. Aplicación de técnicas de minería de datos. Una vez que se obtienen los archivos en formato .arff, se procede a su uso en la herramienta weka, donde se carga el archivo y se procede a elegir la opción por la cual será procesada la información, para el caso de reglas de asociación, se elige la opción associate y posteriormente se selecciona el algoritmo apriori y se ajustan parámetros como el número de reglas, la salida, la cobertura mínima, entre otras, para posteriormente iniciar la ejecución y obtener los patrones.

En el caso de clustering, se sigue un procedimiento similar, seleccionando la opción cluster, después el algoritmo a emplear, en este caso k-means y se ajustan los parámetros, como el número de clusters en los que se va a agrupar la información y las iteraciones que tendrá.

4. Interpretación y evaluación. A partir de los resultados obtenidos de la fase anterior, se analizan los "patrones" obtenidos por weka y se evalúa si realmente son útiles, esta tarea es realizada por el analista.

6. Resultados

Esta sección presenta los principales resultados obtenidos del análisis de los datos del SITTESJo a través de técnicas de minería de datos. Se divide en dos subsecciones, una para la discusión de resultados obtenidos con reglas de asociación y otra para los resultados obtenidos del análisis por medio de herramientas de agrupamiento.

6.1. Reglas de asociación

Para el análisis mediante reglas de asociación se utilizó la información de la base de datos que corresponde a los apartados de antecedentes económicos, familiares y académicos, a partir de esta información se llegaron a las siguientes conclusiones.

1. Los alumnos que suelen practicar algún tipo de deporte como el futbol a su vez tienen problemas económicos y los que no practican ninguna actividad deportiva no presentan problemas económicos.
2. La mayor parte de los alumnos que deserta tienen problemas económicos, razón por la cual deben buscar una forma de obtener ingresos y deciden entrar a trabajar, y al pasar el tiempo deciden abandonar sus estudios para continuar obteniendo el ingreso económico.
3. Los alumnos que desertan son alumnos que en sus estudios previos han sido buenos con un promedio favorable, lo cual indica que la causa de su deserción no se debe directamente a situaciones académicas.

4. La mayoría de los alumnos que desertan nunca han interrumpido sus estudios en los niveles básico y medio superior.
5. La gran mayoría de los alumnos que ingresan a la carrera de ingeniería en sistemas computacionales tienen una idea errónea del enfoque de esta al momento en el que deciden inscribirse.

A partir de los resultados obtenidos con el uso de reglas de asociación se obtuvo un panorama general de la situación de los alumnos de la comunidad estudiantil de la carrera de ISC y las causas probables por las que deciden abandonar sus estudios. Por ejemplo, la contraposición de opiniones entre alumnos (hombres y mujeres) y la relación que existe entre estudiantes que practican deportes y su situación económica, da una referencia de como los alumnos pueden actuar y de lo que piensan. Estos factores pueden influir en que los alumnos tengan un bajo desempeño académico y pueda ser esto un detonante en decisiones como abandonar sus estudios. Sin embargo, estos resultados son preliminares y es necesario un estudio más extenso para poder obtener más información que ayude a la toma de decisiones por parte de la comunidad académica. Los resultados presentados corresponden a las generaciones del 2008 al 2013.

6.2. Clustering o agrupamiento

Una vez concluida la aplicación de reglas de asociación se utilizó la técnica de clustering k-medias, a un conjunto de datos más amplio, ya que se incluyeron nuevos datos obtenidos de la aplicación del cuestionario FO-TESJo-039 y los test de autoestima, asertividad y estilos de aprendizaje, además del registro de nuevas calificaciones de los alumnos. De la aplicación de este algoritmo se obtuvieron los siguientes resultados:

1. De un total de 831 alumnos, 269 de ellos tienen otra licenciatura en mente que les hubiera gustado estudiar, de estos 269, 15 prefieren no seguir estudiando y comenzar a trabajar para obtener un ingreso económico y el resto prefiere salir de la carrera para buscar el ingreso a la carrera de su agrado.
2. De un 60% de alumnos que toman medicamentos, sus padres tienen un nivel educativo superior al de secundaria por lo que cuentan con atención médica, pero se debe dar seguimiento para que no sea causa de deserción. En este caso la atención es porque del 100% de alumnos que toman medicamentos solo el 60% cuenta con atención médica mientras que el otro 40% no cuenta con la atención, lo cual, en algunos casos no es suficiente o bien sus padres no los apoyan al no saber del problema.
3. Los alumnos que tienen a alguien que depende económicamente de ellos, es decir, son casados o que trabajan son fuertes candidatos a desertar debido a su situación económica y familiar.
4. Todos los alumnos encuestados mencionan que conocen el campo de acción de la carrera y el 91% de ellos no ha estado inscrito en otra licenciatura, sin embargo únicamente el 4% de los alumnos no tiene motivos para no concluir la carrera, esto es que la carrera fue su primera opción y no tienen otra licenciatura en mente,

de manera que las causas de deserción se deben a causas sociales, económicas, personales o bien que no tienen el suficiente compromiso para culminar sus estudios.

5. A los alumnos que tienen otra carrera en mente se les dificulta la situación escolar ya que están en clases solo porque no tienen otra opción por el momento pero llegada su oportunidad ellos se decidirán por abandonar la carrera e inscribirse para tratar de entrar a lo que ellos realmente quieren estudiar.
6. La situación económica del estudiante propicia deserción cuando no se dan a conocer cuáles son las becas a las cuales el alumno puede acceder ya sean por parte de la institución o bien por algún organismo gubernamental, esta es tarea de los tutores o bien del personal de servicios escolares de la institución, pero en muchas ocasiones de deja de lado y no se presta la atención suficiente a la publicación de las convocatorias entre los alumnos.
7. De los alumnos enfermos se observa que el seguro institucional ayuda a cubrir las necesidades del caso, ya que de no tenerlo no podrían cubrir sus necesidades, lo cual lleva a que el alumno en ocasiones este estudiando por conservar el seguro y no porque realmente quiera hacerlo, lo cual se ve reflejado en las calificaciones de este al término de cada semestre.

Con los resultados obtenidos de este análisis, se puede observar que conforme se obtienen más datos de la aplicación de los cuestionarios, el análisis se hace más profundo y se obtiene información más precisa sobre las necesidades de los alumnos, lo cual representa un beneficio en la toma de decisiones que ayuden a la permanencia de los alumnos en la institución y puedan concluir la carrera de manera satisfactoria.

7. Conclusiones y trabajo futuro

En este trabajo se presenta el desarrollo del SITTESJo, enfocado a la automatización del proceso de tutoría para hacer este proceso más eficiente y eficaz, y que a la vez permita la adecuada toma de decisiones para atender problemas fundamentales como la deserción escolar.

El trabajo que se ha realizado hasta el momento sobre el desarrollo del SITTESJo se enfoca en la captura de información de los alumnos de ISC misma que es almacenada en la base de datos, para posteriormente ser analizada mediante técnicas de minería de datos y obtener patrones de comportamiento comunes entre los alumnos, con los cuales se pueden identificar problemas presentes en la comunidad estudiantil que afectan su desempeño y ocasionan problemas tanto para el alumno ya que puede no concluir sus estudios, como para la institución misma en el incumplimiento de la eficiencia terminal. La principal aportación del trabajo discutido en este artículo, es que además de sistematizar el proceso de tutoría, apoya el análisis de los datos por medio de herramientas de minería de datos y reconocimiento de patrones.

Con el análisis de la información que se realizó se obtuvieron resultados preliminares de los factores que pueden influir en la deserción y bajo desempeño de los estudiantes de la carrera de ISC del TESJo, entre estos se encuentran factores sociales, económicos y académicos.

Finalmente mediante este proyecto se el desarrollo un sistema integral de tutorías, que consiste en la página web y la aplicación de técnicas de minería de datos para obtener patrones. Sin embargo, como trabajo a futuro, es necesario probar con otras técnicas como un sistema experto basado en redes neuronales artificiales, para mejorar los resultados obtenidos hasta el momento.

Referencias

1. García-Ibarra, C. A.: Manual del tutor del SNEST. México, SEP, pp. 145 (2012)
2. Sistema Inteligente para la Tutoría Académica. Disponible en: https:// www.sita.uaemex.mx/tutoria/index_ok3.html, México, SEP (2015)
3. Dávila-Avendaño, M.: Implementación de tutoría en línea para alumnos de la facultad de ingenierías de la universidad del valle de Atemajac en un entorno virtual de aprendizaje. En: X Congreso Nacional de Investigación Educativa, pp. 1–13 (2008)
4. García-López, R., Cuevas-Salazar, O., Vales-García, J., Cruz-Medina, I.: Impacto de la tutoría presencial y virtual en el desempeño académico de alumnos universitarios. Revista Iberoamericana De Educación, Vol. 58, No. 2, pp. 1–11 (2015)
5. Secretaría de Educación Pública: Programa sectorial de educación 2013-2018, México, SEP, pp. 117 (2013)
6. Chaín, R. R.: Deserción, rezago y eficiencia terminal en las IES: Propuestas metodológicas para su estudio. México, ANUIES, 1 Ed, pp. 244 (2001)
7. García de Fanelli, A.: Rendimiento académico y abandono universitario: Modelos, resultados y alcances de la producción académica en la Argentina. Revista Argentina de Educación Superior, Vol. 8, pp. 9–38 (2014)
8. Maris-Vázquez, S., Noriega-Biggio, M., Maris-García, S.; Relaciones entre rendimiento académico, competencia espacial, estilos de aprendizaje y deserción. Revista Electrónica de Investigación Educativa, Vol. 15, No. 1, pp. 29–44 (2013)
9. Tamirán-Pereira, R., Calderón-Romero, A., Jiménez-Toledo, J.: Descubrimiento de perfiles de deserción estudiantil con técnicas de minería de datos. Revista Vínculos, Vol. 10, No. 1, pp. 373–383 (2013)
10. Formia, S., Lanzarini, L., Hasperué, W.: Caracterización de la deserción universitaria en la UNRN utilizando Minería de Datos. Un caso de estudio. Revista Iberoamericana de Educación en Tecnología y Tecnología en Educación, Vol. 11, pp. 92–98 (2013)
11. Secretaría de Educación Pública: Formatos codificados TESJo. Disponible en: http://tesjo.edomex.gob.mx/formatos_codificados, México, SEP (2016)
12. Witten, I., Frank, E.: Data mining. San Francisco, Calif., Morgan Kaufmann, 4 Ed., pp. 664 (2016)
13. Hernández-Orallo, J., Ramírez-Quintana, M., Ferri-Ramírez, C.: Introducción a la minería de datos. Madrid, Pearson Prentice Hall, 1 Ed., pp. 680 (2004)
14. Valcárcel-Asencios, V.: Data Mining y el descubrimiento del conocimiento. Revista de la Facultad de Ingeniería Industrial, Vol. 7, No. 2, pp. 83–86 (2004)
15. Mesa-Páez, L. O., Rivera-Lozano, M., Romero-Dávila, J. A.: Descripción general de la Inferencia Bayesiana y sus aplicaciones en los procesos de gestión. La Simul. al Serv. la Acad. Fac. Adm. Univ. del Rosario, Vol. 2, pp. 1–28 (2011)
16. Pérez-López, C., Santín-González, D.: Minería de datos: Técnicas y herramientas. Madrid, España, Thomson, 2 Ed., pp. 808 (2007)

17. MacQueen, J. B.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Symposium on Math, Statistics, and Probability, Berkeley, CA, University of California Press, pp. 281–297 (1967)

ISSN 1870-4069

Estudio de la indumentaria indígena mexicana

Sandra Rodríguez-Mondragón, Oscar Herrera-Alcántara, Luis Jorge Soto-Walls, Manuel Martín Clavé-Almeida

Universidad Autónoma Metropolitana, Ciudad de México, México

srm@azc.uam.mx, oha@correo.azc.uam.mx, luissotowalls@gmail.com, mclavealmeida@gmail.com

Resumen. La indumentaria de origen indígena mexicano se caracteriza primordialmente por los materiales de los que está elaborada, el tipo de tejido, los íconos tejidos o bordados en ella y los colores aplicados en su composición. Definir su origen e identificación visual con base en los íconos requiere conocimientos etnográficos, técnicos e históricos. En este trabajo se presenta el análisis de imágenes (íconos) identificadas en indumentaria indígena del estado de Chiapas, México, en la región de los Altos de Chiapas, lo que corresponde a indumentaria de origen prehispánico Maya, y se describe el desarrollo de un programa que implementa una máquina de pila capaz de procesar cadenas de texto para reconstruir iconografía de indumentaria indígena desde su mínima expresión gráfica aplicando operadores. La reproducción de íconos, seleccionados por medio del análisis gráfico-visual y cadenas de texto, permiten identificar patrones y generar nuevos diseños que mantengan la identidad visual del textil indígena, además de contribuir a la preservación de la cultura mexicana.

Palabras clave: indumentaria indígena, identidad visual, identificación visual, máquina de pila, reconocimiento de patrones.

Study of Mexican Indigenous Clothing

Abstract. The clothing of Mexican indigenous origin is characterized primarily by the materials of which it is made, the type of fabric, the icons woven or embroidered therein and the colours applied in its composition. Defining their origin and visual identification based on the icons requires ethnographic, technical and historical knowledge. This work presents the analysis of images (icons) identified in indigenous clothing of the State of Chiapas, Mexico, in the region of the highlands, which corresponds to clothing of Maya prehispanic origin, and is presented the description of a software that implements a stack machine based model that processes strings of operators and icons, in order to generate the iconography of indigenous clothing from its minimal visual expressions of icons. The iconography expressed as strings allows to identify patterns, to develop new designs, to preserve the visual identity of the original indigenous clothing, and to contribute to preserve the Mexican culture.

Keywords: indigenous clothing, visual identity, visual identification, stack machine, pattern recognition.

1. Introducción

Generalmente los estudios sobre el diseño indígena se centran en aspectos antropológicos. En nuestro caso, el análisis está dirigido a los aspectos de lenguaje básico en el diseño que considera las cualidades simbólicas, icónicas, y el papel del diseñador ante sus elementos característicos. Las cualidades formales de la indumentaria son la clave de su comprensión y estudio, de la misma manera que la identificación de estos aspectos permite explorar las posibles propuestas contemporáneas. Así, por ejemplo, el ritual constituye el soporte principal de las prendas ceremoniales, en el diseño, los colores, la textura y la forma de los íconos, y muestran una vasta gama de posibilidades de diseño con identidad étnica. Diversos estudios arqueológicos, históricos y etnográficos, revelan el papel de la indumentaria indígena en la vida política, social y religiosa de Mesoamérica [1, 2].

Los nuevos diseños con identidad étnica mexicana, son generados principalmente por dos tipos de diseñadores: diseñadores occidentales (de habla hispana) y diseñadoras indígenas (en este caso de origen tzotzil o tzeltal) [3]. Actualmente, algunos diseñadores mexicanos se han dado a la tarea de aprovechar la belleza y armonía cromática de los textiles indígenas, algunas propuestas han innovado solo en la confección empleando el tejido indígena como material para sus diseños y otros han integrado de forma exitosa la confección con la forma del tejido. Otros, han optado por copiar el diseño y aplicarlo como estampado en otros materiales obteniendo resultados importantes en la moda internacional. Existe todo un discurso alrededor de esta temática que incluye definir hasta qué punto el diseñador contemporáneo puede hacer uso de estas tecnologías (materiales) y diseños sin considerar la parte simbólica del diseño indígena [4].

En nuestro trabajo se retoman los aspectos estéticos y formales revalorándolos desde la perspectiva del diseño desarrollado con tecnologías computacionales y de reconocimiento de patrones.

Es importante mencionar que con este trabajo no se pretende competir con el diseño indígena, sino retomar algunos de sus elementos, valor su armonía formal y preservar este conocimiento ancestral de los pueblos indígenas. Entonces, una motivación para esta investigación es pensar que podemos ser capaces de identificar los principios de diseño indígena con técnicas de inteligencia artificial, y reconstruir diseños antiguos, o generar nuevos automáticamente sin perder la esencia visual de la cultura mexicana.

Antes de continuar, consideramos importante precisar los siguientes conceptos sobre identidad visual e iconografía.

La identidad visual se refiere a identificar características de color, textura y forma de los íconos grabados en textiles indígenas.

La identificación visual se refiere a la acción de cómo dar datos característicos necesarios para reconocer “algo” por medio de la visión. En este caso específico, a través de la identidad visual.

La palabra ícono, de acuerdo con la RAE [5] tiene origen del francés icône, del ruso ikona, y del griego bizantino εἰκών, όνος eikṓn, ónos; y contempla las siguientes definiciones:

1. Representación religiosa de pincel o relieve, usada en las iglesias cristianas orientales.
2. Tabla pintada con técnica bizantina.
3. Signo que mantiene una relación de semejanza con el objeto representado.
4. En informática. Símbolo gráfico que aparece en la pantalla de una computadora u otro dispositivo electrónico y que representa un programa, un sistema operativo, etc.

De lo anterior, tomamos la definición de ícono como "el signo que mantiene una relación de semejanza con el objeto representado".

De igual forma, según la RAE, la palabra iconografía [6] tiene su origen en el latín medieval, iconographia, y del griego εἰκονογραφία eikonographía; y considera cinco definiciones:

1. Conjunto de imágenes, retratos o representaciones plásticas, especialmente de un mismo tema o con características comunes.
2. Representación o imagen de un personaje o de una realidad determinados.
3. Sistema de imágenes simbólicas.
4. Arte de la imagen o la representación plástica.
5. Estudio de las imágenes o representaciones plásticas en el arte.

De lo anterior, retomamos la última definición que alude al estudio de imágenes, que contienen información o códigos que se necesitan decodificar (ver [7]).

Por otro lado, tenemos que el estudio iconográfico en las artes visuales está constituido por dos aspectos: el formal y contextual; que son de suma importancia para el análisis de las obras de arte y sobre todo evitar una sobre-interpretación de la misma cayendo en errores comunes como la intromisión de elementos fuera de lugar. Así, según [8]:

"Una vez más y gracias a los estudios iconográficos, iconológicos, formales, sociales y psicológicos el arte puede estudiarse desde una perspectiva objetiva logrando aterrizar los conceptos y formas básicas de las mismas y sobre todo comprender el contexto sobre el cual fueron desarrollados, considerando esto una de las piezas fundamentales para poder reconocer el arte a través del tiempo."

Por lo que, la iconografía es vista como un "sistema de imágenes simbólicas", donde el análisis iconográfico es "el estudio de los íconos o figuras de carácter simbólico representadas", en este caso, en el textil indígena.

2. Metodología

La presente investigación es un estudio de caso. Los pasos seguidos en su desarrollo son los siguientes:

1. Realizar una investigación de campo para recopilar muestras fotográficas de textiles, pertenecientes al acervo algunas colecciones de indumentaria indígena mexicana y producción artesanal actual.
2. Generar un catálogo digital de íconografía del caso de estudio.
3. Realizar un análisis visual de iconografía del caso de estudio y determinar las cualidades formales que permitan realizar la identificación visual de la indumentaria indígena mexicana del caso de estudio.
4. Diseñar e implementar un programa de cómputo que realice la reconstrucción de gráficos textiles y de pauta para desarrollar la reconstrucción o propuestas de nuevos diseños.

2.1. Análisis visual

El análisis visual de íconos se realiza de forma manual, la descripción de este proceso se presenta en la Figura 1, y consiste en cuatro etapas:

1. Identificación de iconografía en indumentaria indígena
2. Digitalización de los íconos identificados
3. Análisis geométrico visual a partir de simetrías e isometrías
4. Desarrollo de un alfabeto gráfico de lenguaje formal de descripción visual

En el punto 3 anterior, isometría se refiere a aquella transformación en el plano que preserva la longitud, en donde se distinguen dos tipos:

a) Directa: preserva el sentido, que a su vez puede ser de rotación (alrededor de un punto llamado centro de rotación), o de traslación en una determinada dirección
b) Opuesta: invierte el sentido, y puede ser a su vez reflexión sobre una línea o reflexión con deslizamiento

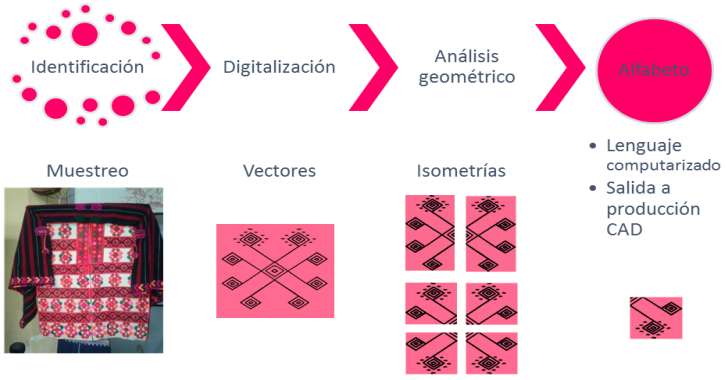


Fig. 1. Análisis formal de iconografía.

La Tabla 1 presenta dos ejemplos del desarrollo de este proceso, así como la forma en que se puede generar una cadena de descripción visual.

En el Ejemplo 1 se realiza el análisis de un ícono que se genera a partir de tres módulos o secciones primarias (A, B y C) y por medio de operadores de traslación, reflexión, yuxtaposición y **or** lógicos a nivel de pixeles, se genera una composición iconográfica.

Tabla 1. Análisis visual de íconos.

<table>
<tr><td colspan="3">Ejemplo 1. Íconos del grupo lingüístico tzeltal, origen municipio de Tenejapa, Chiapas, México.</td></tr>
<tr><td>A</td><td rowspan="4">= (trasladar verticalmente A)
or
(trasladar horizontalmente B)

= A*
or
(trasladar verticalmente C)</td><td>Repetir 3 veces</td></tr>
<tr><td>B</td><td rowspan="3">(trasladar horizontalmente

yuxtaponer verticalmente con</td></tr>
<tr><td>C</td></tr>
<tr><td>A*= Reflejar verticalmente A</td><td>: Repetir 3 veces</td></tr>
<tr><td colspan="3">Ejemplo 2. Ícono del grupo lingüístico tzotzil, origen municipio Aldama, Chiapas, México.</td></tr>
<tr><td>D</td><td>= D / D*
Yuxtaponer verticalmente</td><td>
yuxtaponer horizontalmente
con</td></tr>
<tr><td>D* =
Reflejar verticalmente D</td><td>reflejar horizontalmente</td><td></td></tr>
</table>

En el Ejemplo 2 solo se aplican operadores de reflexión, yuxtaposición y **or** lógicos a nivel de pixeles.

Nótese que en los Ejemplos 1 y 2 se usan reflexiones, traslaciones y yuxtaposiciones tanto en sentido horizontal como vertical.

A continuación, se describe el uso de una máquina de pila para procesar los íconos mediante la aplicación de operadores de manipulación de imágenes. La manipulación de las imágenes está dada por cadenas en notación posfija, mismas que pueden ser evaluadas con un algoritmo que hacer uso de una pila [12]. Cabe mencionar que las cadenas en notación posfija tienen su contraparte como cadenas en notación infija con paréntesis equilibrados, y dado que el conjunto de cadenas con paréntesis equilibrados no es un lenguaje regular [9], no se propuso un modelo de autómatas finitos. A manera de ejemplo considérese la cadena A op1 B op2 C donde A, B y C son iconos, y op1 y op2 son operadores. Si se requiere que se ejecute primero op1 entre A y B, se puede escribir con paréntesis como (A op1 B) op2 C, que en notación infija es A B C op1 op2, y cuya evaluación se realiza leyendo de izquierda a derecha, así que se introducen en la pila A y B, y al llegar al operador op1 se sacan de la pila A y B y se mete a la pila el resultado parcial A op1 B, como otro icono nuevo, luego se continua colocando en la pila a C, y al llegar a op2 se extraen los dos íconos de la pila para luego colocar el resultado en la cima de la pila como un ícono resultante (A op1 B) op2 C.

Por otro lado, dado un tamaño en pixeles de un ícono (altura H y amplitud W) y asumiendo valores binarios de pixel blanco o negro, el número posible de combinaciones diferentes para un ícono es 2^{H*W}, es decir, el número de iconos diferentes que podrían usarse como parte del alfabeto es finito.

2.2. Máquina de pila para procesar íconos

Una máquina de pila [9] M es una 7-tupla $M = (Q, \Sigma, \Gamma, \delta, q0, Z, F)$ donde:
Q es el conjunto finito de estados,
Σ Es el conjunto finito de símbolos de entrada,
Γ Es el conjunto finito de símbolos de la pila,
δ Es un subconjunto de $Q \times (\Sigma \cup \{ \varepsilon \}) \times \Gamma \times Q \times \Gamma$ que es la función de transición,
$q_0 \in Q$ es el estado inicial,
$\# \in \Gamma$ es el símbolo inicial de la pila,
$F \subseteq Q$ es el conjunto de estados de aceptación .

En este caso particular es relevante precisar que:

- Σ son los operadores de la segunda columna de la Tabla 2, en unión con los íconos identificados por letras mayúsculas (que a su vez indican la ruta en el sistema de archivos de imágenes en formato BMP)
- Γ son los íconos (imágenes) que se almacenan en la pila, y los resultados intermedios (imágenes) obtenidos al aplicar los operadores sobre los íconos. En la práctica, en lugar de almacenar cada ícono, se genera un archivo por computadora y se almacena la ruta completa en el sistema de archivos.
- # es el símbolo inicial de la pila, y en la práctica indicará que la pila está vacía, por lo que no se almacenará un valor específico.

– δ determina las reglas de transición, que en este caso está dada por los pasos del Algoritmo 1, en donde la idea básica es que al leerse un operando éste se inserta en la pila, y al leerse operador k-ario se toman k elementos de la pila, se ejecuta el operando con ellos, y se vuelve a colocar el resultado en la pila.

La ejecución de la máquina de pila se describe en el Algoritmo 1, en donde, al proporcionar cadenas válidas (ver Tabla 3), la máquina se ejecuta satisfactoriamente y deja en la cima de la pila el resultado que corresponde a la imagen de la indumentaria indígena reconstruida desde sus íconos. En el caso del ejemplo, el resultado de (A op1 B) op2 C quedará como único elemento en la cima de la pila, (porque el símbolo # no ocupa lugar en la pila), lo cual indica que, al sacar el resultado de la pila ésta quedará vacía.

Los resultados intermedios se dejan opcionalmente en un directorio temporal, a fin de facilitar la depuración del avance de la reconstrucción.

Cada uno de los operadores fue implementado en lenguaje Java [10] para imágenes en formato BMP, visto como matrices de pixeles con amplitud W y altura H.

Cada operador es ejecutado por un autómata de pila que aplica las siguientes reglas a los tokens separados por comas, por lo que el analizador léxico identifica cada token separado por comas, y puede ser del tipo operador o ícono.

Existen operadores unarios como la traslación que toma como entrada un ícono, y un porcentaje de traslación según la amplitud o altura del ícono y genera una imagen de mayor tamaño, pero desplazando los pixeles en negro, por lo que la altura del ícono se mantiene, pero la amplitud se extiende.

El alfabeto se compone de operadores e íconos, y los íconos se identifican como letras mayúsculas, que en notación gramatical serían:

Alfabeto → Operador | Ícono,
Operador → -, *, x, y, ~, /, +, %, or,
Ícono → A, B, C, D, E, F.

En la gramática y en el Algoritmo 1, las letras mayúsculas son rutas absolutas en el árbol de directorios para los archivos BMP de los íconos correspondientes. Por lo que la máquina de pila procesa en realidad imágenes de acuerdo a los diferentes operadores de la Tabla 2.

Algoritmo 1. Máquina de pila que procesa íconos para obtener imágenes de indumentaria indígena: algoritmo de evaluación de cadenas en notación postfija.

Entrada: Cadena de texto que incluye íconos y operadores en notación posfija
Salida: Imagen de indumentaria indígena

Identificar tokens separados por comas en la cadena a procesar
Por cada token hacer:
Si el token es operador unario, sacar elemento de la pila y aplicar el operador, luego insertar el resultado en la pila. En caso de que el operador requiere parámetros adicionales leerlos como tokens desde la cadena de entrada.

Si el token es operador binario, sacar dos imágenes de la pila y aplicar el operador, luego insertar el resultado en la pila

Si el token es un Ícono (letras A, B, C, …) meter a la pila la imagen correspondiente

La salida está en la cima de la pila.

A continuación, se describen los operadores necesarios identificados a partir de las imágenes de indumentaria indígena.

2.3. Descripción de operadores

La aplicación de los siguientes operadores a los íconos, ha permitido reconstruir una imagen original.

- **Reflexión horizontal.** Genera una imagen o ícono con el reflejo pixel a pixel sobre el eje Y.
- **Reflexión vertical.** Genera una imagen o ícono con el reflejo pixel a pixel sobre el eje X.
- **Traslación horizontal**. Toma como entradas un ícono y un porcentaje de traslación según la amplitud del ícono y genera una imagen de mayor tamaño en donde se desplazan los pixeles (en negro) por lo que la altura del ícono se mantiene, pero la amplitud se extiende.
- **Traslación vertical**. Toma como entradas un ícono y un porcentaje de traslación según la altura del ícono y genera una imagen de mayor tamaño en donde se desplazan los pixeles (en negro) por lo que la amplitud del ícono se mantiene, pero la altura se extiende.
- **Rotación**. Realiza la rotación en grados de un ícono, en sentido inverso a las manecillas del reloj.
- **Yuxtaposición vertical.** Realiza la y unión vertical de dos íconos. Las amplitudes de los íconos deben coincidir.
- **Yuxtaposición horizontal.** Realiza la y unión horizontal de dos íconos. Las alturas de los íconos deben coincidir.
- **Escalamiento**. Modifica el tamaño del ícono según un valor positivo de punto flotante.
- **Or**. Realiza la operación OR lógica pixel a pixel de dos íconos dados.

Es importante mencionar que, aunque existen otros operadores tales como el and, not y xor lógicos, aún no han sido considerados porque el análisis de las imágenes de la indumentaria indígena no ha requerido este tipo de operadores, sin embargo, no se descarta su futura inclusión a fin de enriquecer la cantidad de los diseños generados, incluidos operadores más avanzados que involucren proporciones áureas, uso de series de Fibonacci, y fractales.

Tabla 2. Operadores para íconos.

Operador	Símbolo	Tipo	Ejemplo	Descripción
Reflexión horizontal	**-**	Unario	A,-	Aplicar la simetría horizontal
Reflexión vertical	*****	Unario	A,*	Aplicar la simetría vertical
Traslación horizontal	**x**	Unario	A,10,x	Trasladar un 10% el ícono hacia la derecha
Traslación vertical	**y**	Unario	A,20,y	Trasladar un 20% el ícono hacia abajo
Rotación	**~**	Unario	A,30,~	Rotar el ícono 30 grados
Yuxtaposición vertical	**/**	Binario	A,B,/	Unir verticalmente los íconos dados en los archivos A y B en una sola imagen resultante
Yuxtaposición horizontal	**+**	Binario	A,B,+	Unir horizontalmente los íconos dados en los archivos A y B en una sola imagen resultante
Escalamiento	**%**	Unario	A,40,%	Amplificar en un 40% el ícono A
Or	**or**	Binario	A,B,or	Aplicar pixel a pixel un OR lógico a los íconos dados en los archivos A y B y poner el resultado en una sola imagen

3. Experimentos y resultados

El análisis de imágenes de indumentaria indígena permitió expresarlas como cadenas de texto, que son aceptadas por el autómata de pila M descrito en la Sección 2. La Tabla 3 reúne las cadenas de 16 imágenes estudiadas de los municipios de Aldama y Tenejapa, de Chiapas, México.

En la Tabla 3, se observa la repetición del patrón "**A,-,A,+"** en varias cadenas, lo cual permite identificar una identidad visual de simetría horizontal y vertical.

También se observa que el patrón **"A,-,A,+**,*,**A,-,A,+**,/" se repite en varias ocasiones, y si α**= A,-,A,+**, da lugar a "α*α /" en donde que de manifiesto una identidad visual basada en una simetría y yuxtaposición vertical.

El análisis gráfico realizado hasta el momento ha permitido determinar la identificación visual de textiles indígenas mexicanos de la región de los altos de Chiapas, tomando como caso de estudio huipiles de origen tzotzil y tzeltal. Conforme se trabaje con un mayor número de imágenes e íconos, se esperan identificar patrones más complejos, e inclusive dar pauta a una herramienta de clasificación automática basada en patrones de identidad visual.

El programa desarrollado en Java almacena la información de la Tabla 3, en una base de datos de SQLite [11], para facilitar el procesamiento de los iconos.

Tabla 3. Cadenas que generan imágenes de indumentaria indígena.

Cadena	**Nombre**
A,B,B,B,B,B,B,B,B,B,C,/,/,/,/,/,/,/,/,/,/,- ,A,B,B,B,B,B,B,B,B,B,C,/,/,/,/,/,/,/,/,/,/,+	Aldama1
A,-,A,+,*,**A,-,A,+,/**	Aldama2
A,-,A,+,*	Aldama3
A,-,A,+,*,**A,-,A,**+,/	Aldama4
A,-,A,+,*,**A,-,A,+**,/	Aldama5
A,A,A,/,/,A,A,A,/,/,*,A,A,A,/,/,+,+,A,A,A, /,/,A,A,A,/,/,*,A,A,A,/,/,+,+,-,+,*	Aldama6
A,-,A,+,*,**A,-,A,+**,/	Aldama7
A,-,A,+	Aldama8
A,A,A,+,+,A,A,A,+,+,*,/	Tenejapa1
A,*,A,/,-,A,*,A,/,+	Tenejapa2
A,-,A,+	Tenejapa3
A,*,A,/,-,A,*,A,/,+	Tenejapa4
A,-,A,+	Tenejapa5
A,*,A,/,-,A,*,A,/,+	Tenejapa6
A,*,A,/,-,A,*,A,/,+	Tenejapa7
A,-,A,+	Tenejapa8

4. Conclusiones y trabajo futuro

Los experimentos reflejan la factibilidad del análisis iconográfico de imágenes de indumentaria indígena, y la posibilidad de procesarlas con un modelo computacional de máquina de pila, que es una herramienta útil en la generación de nuevos lenguajes de programación, y que en este caso hemos aplicado a un lenguaje de manipulación icónica.

Actualmente el reconocimiento de características de imágenes lo realiza un experto, así que no se descarta la posibilidad de que en un futuro se apliquen técnicas de reconocimiento de patrones en imágenes para tal efecto.

El modelo de máquina de pila, facilita la identificación de patrones, así como la reconstrucción de imágenes ya existentes. Además, da pauta a la generación de nuevos diseños que mantengan la identidad visual en un intento por conservar nuestra cultura mexicana, en convivencia con nuevas tecnologías y futuras aplicaciones de técnicas de inteligencia artificial.

Como trabajo futuro se espera agregar un análisis de color y modulación que incluya el movimiento de íconos en una trayectoria armónica, además de agregar el procesamiento de cadenas en notación infija dado que actualmente sólo se procesan cadenas en notación posfija, así como la inclusión de operadores más avanzados que contemplen proporciones áureas, series, y fractales.

Referencias

1. Aguilera, M.: La faja ralámuli. Un entramado cosmológico. UNAM, Instituto de Investigaciones Antropológicas e Instituto de Investigaciones Estéticas, México (2011)
2. Johnson, K.: Saberes enlazados: La obra de Irmgard Weitlaner Johnson. CONACULTA, México (2015)
3. Freund, R.: Tzotziles y Tzeltales. Disponible en: http://www.cdi.gob.mx/print.php?id_seccion=357
4. OMPI: La propiedad intelectual y la artesanía tradicional. Disponible en: http://www.wipo.int/edocs/pubdocs/es/wipo_pub_tk_5.pdf
5. Diccionario de la lengua española: Real Academia Española RAE definición de ícono. Disponible en: http://dle.rae.es/?id=KsRzX3u
6. Diccionario de la lengua española: Real Academia Española RAE definición de iconografía. Disponible en: http://dle.rae.es/?id=KsZhO9s
7. Lenguaje iconográfico. Disponible en: http://pendientedemigracion.ucm.es/info/romanica/personales/docsJMLM/01_5Lenguaje%20iconografico.pdf
8. Análisis-iconográfico-iconológico. Disponible en: https://hermeneuticacui2012.files.wordpress.com/2012/03/analisis-iconografico-iconologico.pdf
9. Hopcroft, J. E., Motwani R., Ullman, J. D.: Introducción a la Teoría de Autómatas, Lenguajes y Computación. Edición en español, Pearson Educación, Madrid (2008)
10. Java. Disponible en: http://www.oracle.com/technetwork/es/java/javase/overview/index.html
11. SQLite. Disponible en: http://www.sqlite.org
12. Aho, A. V., Sethi, R., Ullman, J. D.: Compilers principles, techniques and tools. Addison-Wesley, CA, USA (1986)

Impreso en los Talleres Gráficos
de la Dirección de Publicaciones
del Instituto Politécnico Nacional
Tresguerras 27, Centro Histórico, México, D.F.
noviembre de 2017
Printing 500 / Edición 500 ejemplares