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.

pp. 125–134; rec. 2017-03-14; acc. 2017-05-06

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):

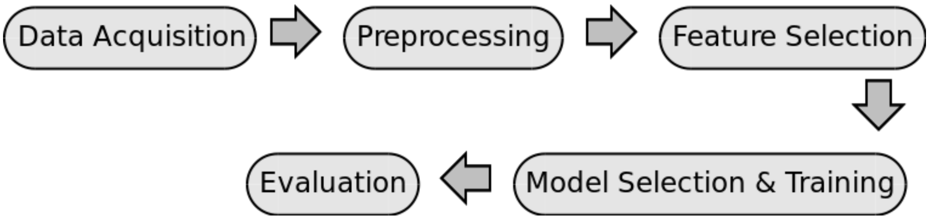

Fig. 1. Proposed methodology based on [12].

- 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

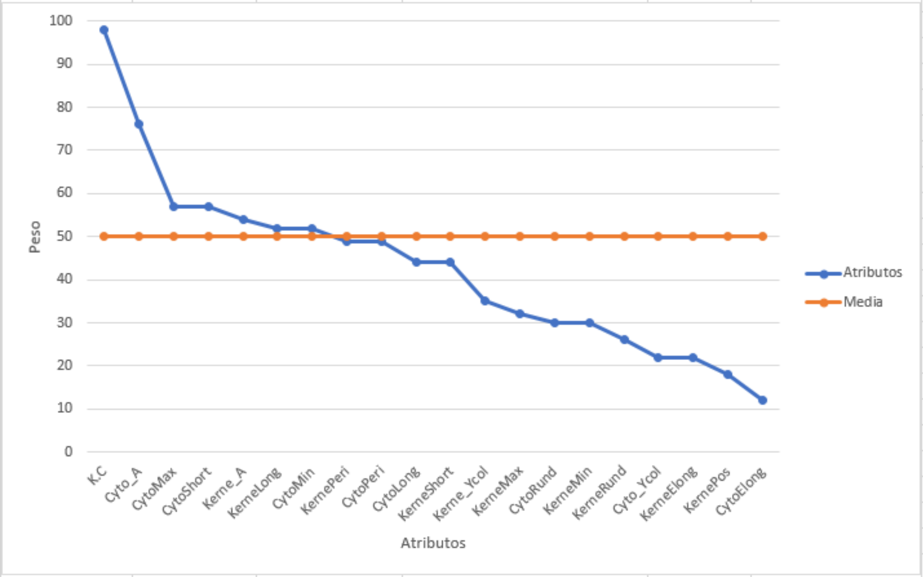

Fig. 2. Attributes with respect to the mean of the data.

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)