Finding patterns in a wildfire data warehouse by applying data mining techniques

María J. Somodevilla, Maribel Fortiz Flores, Ivo H. Pineda T., Concepción Pérez de Celis H.
FCC, Benemérita Universidad Autónoma de Puebla, México
{mariajsomodevilla, maribel.fortiz, ivopinedatorres,mcpelish}@gmail.com

(Paper received on June 30, 2013, accepted on August 15, 2013)

Abstract. This paper presents a decision support system on prevention of wildfires. Over the last 15 year in the state of Tlaxcala, it has been collected a large amount of data about wildfires and its origins, using this a data warehouse was developed. Applying descriptive and predictive data mining techniques, it is possible to determine wildfires causes, and other data related to.

Keywords: wildfires, Forest fires, data warehouse, data mining, classification methods, clustering, association.

1 Introduction

Names such as brush fire, bushfire, forest fire, desert fire, grass fire, hill fire, peat fire, vegetation fire, and veldfire may be used to describe phenomenon known as wildfires. Wildfires involve a major change in the ecological factors that govern the behavior of ecosystems and the importance, they have acquired in recent decades, is one of the most serious environmental problems which have to face environmental managers. Given the problems and concerns facing the managers of the environment, the operational use of new communication infrastructures with powerful and flexible information processing tools is sought.

This paper proposes a wildfires data warehouse for being exploited by an OLAP queries and data mining techniques. Data are provided by the National Forest Commission of the State of Tlaxcala (CONAFOR), to which it is applied a process Knowledge Discovery in Databases (KDD) in order to extract valid, useful and novel information. This research focuses on wildfires that have been generated in the state of Tlaxcala in the last 15 years. The work aims for a description of the conditions and characteristics that represent the major causes for wildfires, considering affected surfaces and vegetation, areas with the greatest impact, it is considered important to determine seasons and/or months with the highest incidence among others. The Data mining techniques implementations used come from WEKA.

The paper's structure is as follows. Section 2 describes current research as well state of the art related with wildfires, in Section 3 it is explained the architecture used for the development of this research work. Section 4 explains data mining techniques considered for analysis of wildfires or forest fires. Section 5 describes final results of several experiments that confirm our hypothesis. Finally, Section 6 which presents the conclusions drawn from experiments related with the research.

Special Issue in Information Technology
Research in Computing Science 64, 2013, pp. 77-86

2 Related work

Forest fires are one of the significant causes of deforestation and degradation of ecosystems, 90% of forest fires worldwide are manmade. Wildfires negatively affect environment through deforestation, erosion, loss of biodiversity and the emission of CO2, which affect landscape and wildlife habitat [1].
Fire prevention is crucial to avoid forest fires and/or to minimize its consequences once declared. Under this context paper[2] explains how to deal with a KDD process, including image analysis process based on image partitions in order to obtain two classes let's say presence or absence of wildfires, the experiments in that paper, use data from the Secretary of Environment of State of Puebla. In reference [3] found that in Chile, it was developed a system for the analysis of social-demographic variables affecting wildfire through data mining techniques. For instance in work [4], it addressed the problem of forest fires in Spain focusing on obtaining additional information in making decisions aimed at prevention. Based on such experiences, for this project it were considered two phases, called data preparation in which data are selected, cleaned and transformed for next phase called analysis , where different procedures are applied to extract information or a summary information.

3 Design of the decision support system

Data warehouses and OLAP operations are of great importance for large databases analysis. A data warehouse includes mainly historical data, for example facts about the context in which the organization operates [5].
In this research a system is used as a support for decision-making on forest fires generated in the state of Tlaxcala. Due to a lack of method to collect information, most of the information was provided by CONAFOR. The first step of the process consist of mapping an ER model to a multidimensional model, which was implemented as a SQL Server data warehouse, which is exploited by data mining techniques. These techniques allow finding patterns that explain and describe forest fires, particularly using classification and prediction methods.

4 Data mining techniques applied to the data warehouse

Data mining techniques are applied to the data set, having previously obtained in order to analyze the characteristics of forest fires and its causes that generate them, to thereby find patterns that explain and describe their behavior. These patterns will be used to identify areas of risk and creating contingency plans to be used in case a fire occurs.

The predictive models attempt to estimate future values of unknown variables. Descriptive models identify patterns that explain or summarize the data [5].

4.1 Predictive model

4.1.1 Classification

This task is chosen in order to predict the new instances of the class for which the class is unknown. The decision trees technique. was chosen, the C4.5[8] in particular, to know municipalities and month(s), in which they are affected by forest fires.

The attribute that is to be predicted is known as the dependent variable for our situation is *cause of fire*, since its value depends upon, or is decided by, the values of all the other attributes such as *type of fire*. The attributes help in predicting the value of the dependent variable or wildfires, are the independent variables in the dataset.

4.2 Descriptive model

4.2.1 Clustering

This task forms groups such as the objectives of the same group are very similar to each other and at the same time are different from the goals of other groups.

Given n vectors x_1:::;x_n in R^d, and an integer k, find k points $\mu_1 .... \mu_k$ in R^d which minimize the expression:

$$f_k\text{-means} = \sum \min \| x_i - \mu_j \|^2$$

We aim to find k cluster centers. The cost is the squared distance between all the points to their closest cluster center. k-means clustering and Lloyd's algorithm [9] are probably the most widely used clustering procedure. This is for three main reasons:

- The objective function is simple and natural.
- Lloyd's algorithm is simple, efficient and often results in the optimal solution.
- The results are easily interpretable and are often quite descriptive for real data sets

For these reasons *K-Means* was selected for clustering for making main groups of conditions under which a fire takes place.

4.2.2 Association

With this work we seek to establish partnerships between databases instances. Association Rules do not imply a cause-effect relation, for instance, it may or may not be a cause for which the data is associated. Therefore, *Predictive Apriori* method was chosen in order to know the relationship between municipalities, months and affected surface.

5 Experiments and results

5.1 *Clustering*

K-Means was applied to the dataset from table 1.

Table 1 Data under analysis.

Atributes
Cause
Municipality
Month
Adult_Trees
Young_Trees
NoTrees_Bushes
NoTrees_Grasslands

The results found which factors discriminate groups, those are months and municipalities. As a conclusion it is possible to say that wildfires usually were promoted in some municipalities during months of February, March and April, when climatic conditions are more favorable for their generation and spread than other months. Another important aspect, that is observed is the type of vegetation involved, mostly in pasture followed by bushes which by their nature are easy to ignite and spread a fire faster.

A description of the results obtained by applying K-Means to municipalities can be found in Table 2. Figure 2 shows results obtained. Dimensions of vegetation affected are given in hectares.

Table 2 Applying K-Means for the municipalities.

Cause	1	1	1	1
Municipality	Teolocholco	Huamantla	Zitlaltepec	Chiautempan
Month	4	2	4	3
Adult_Trees	0.0051	0	3.8333	0.0296
Young_Trees	0.0409	0.0398	2.3654	0.1091
NoTrees_Bushes	0.5678	0.5962	14.9038	0.9663
NoTrees_Grasslands	1.4565	1.6573	16.266	2.6261

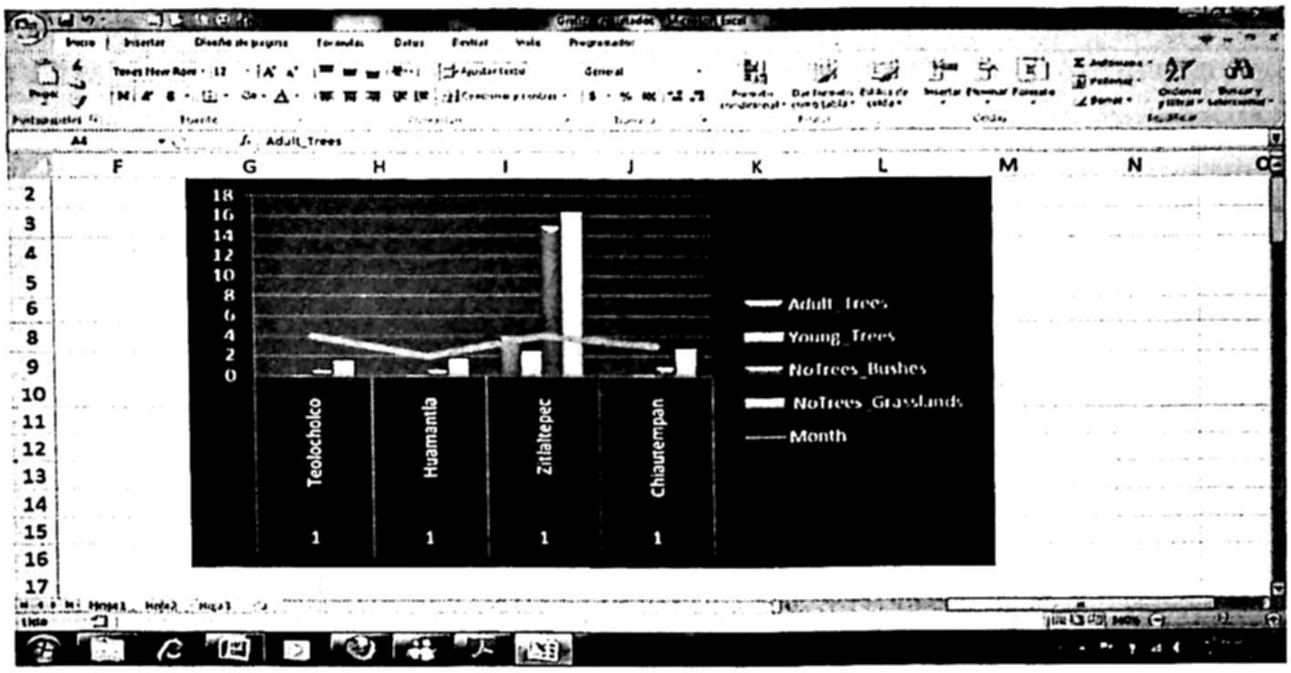

Figure 1 Results from *Simple KMeans* to the municipalities.

Following analysis of state-level data is necessary to examine the properties of municipalities Teolocholco, Huamantla Zitlaltepec and Chiautempan, which were obtained in the previous analysis.

To Teolocholco *K-Means* method is applied with three clusters, resulting that the cause is "1", which corresponds to farming and this affects all three properties. For San Luis' fires occur mainly in months of March and April, while in Acxotla only in March. In terms of vegetation, it is more affected grassland, while the adult trees have a negligible involvement. The data display already described can be seen in table 3.

Table 3 Results from *K-Means* for Teolocholco.

Cause	1	1	1
Site	San Luis	San Luis	Acxotla
Month	3	4	3
Adult_Trees	0	0	0.0909
Young_Trees	0.0323	0.0366	0.0455
NoTrees_Bushes	0.6694	0.5244	0.3977
NoTrees_Grasslands	1.5161	0.878	1.4943

In the municipality of Huamantla, Pillars was the site most affected with a wide variety of causes, among which we have cause 1, 2 and 6 for farming, forestry and walkers bonfires respectively, with a greater involvement in the months of March and April. The vegetation affected is grassland, while the adult trees have a zero involvement. Table 4 displays data already explained.

Table 4 Results from K-Means for Huamantla

Cause	2	6	1
Site	Los Pilares	Los Pilares	Los Pilares
Month	3	4	3
Adult_Trees	0	0	0
Young_Trees	0.1964	0.0656	0.5545
NoTrees_Bushes	0	0.225	0.3636
NoTrees_Grasslands	2.3036	1.0906	2.9273

In the municipality of Zitlaltepec, K-Means method is applied with two clusters, the result shows that the cause was 1 (agricultural activities), affecting the site of Javier Mina and St. Paul in the months of April and March respectively. As seen in Javier Mina, it exists a significant effect on grassland and scrub woodland but for adult involvement is nil. The affected vegetation in San Pablo is pastures. To corroborate the information described above can be seen in Table 5.

Table 5 Results from K-Means for Zitlaltepec.

Cause	1	1
Site	Javier Mina	San Pablo
Month	4	3
Adult_Trees	0	0
Young_Trees	0.1667	0.0847
NoTrees_Bushes	1.5714	0.1949
NoTrees_Grasslands	5.6176	2.0847

To analyze the properties of Chiautempan K-Means method was used with 3 clusters. The result was that the site of San Bartolo is affected by the case 1 (agricultural activities) in the month of April with a major involvement in pasture followed by bush. Farm Tlalcuapan was affected by cause 1 (agricultural activities) in March, mainly affected pastures. Finally, in the site Muñoztla fires are generated by the case 1 (agricultural activities) in the month of March, affecting mainly grassland. To get a better idea about the results and analysis described in Table 6.

Table 6. Results from K-Means to Chiautempan.

Cause	1	1	1
Site	San Bartolo	Tlalcuapan	Muñoztla
Month	4	3	3
Adult_Trees	0.0339	0	0

Young_Trees	0.0678	0.0287	0.1173
NoTrees_Bushes	1.3305	0.2644	1.0816
NoTrees_Grasslands	3.1102	1.1552	2.2194

Once analyzed and explained the results obtained with the method K-Means, it is concluded that the main cause of fires in the state of Tlaxcala is the one which corresponds to agricultural activities indicating the need to create fire prevention campaigns aimed to this sector.

5.2 Classification

Decision trees were used to determine affected municipalities any time of the year, by taking as inputs months, municipalities and affected area, specially to affected area a discrete filtering is applied considering in three intervals. The resulting model is robust so it can obtain valuable information.

In Figure 2, trees show the municipalities that are affected in a given month and together with the affected surface. It is observed that there are municipalities that do not have a high rate of fire (based on cluster analysis explained above), but the forest area affected is large, among these municipalities can be found Calpulalpan, Tlaxco and Altzayanca. Otherwise it was observed that some municipalities that have been identified with a high rate of fire, the affected area is small. One reason for this phenomenon could be its location, vegetation, climate, materials and fuels, and other factors such as influence the spread or fire can be extinguished easily.

Based on these results, it is proposed to the Government of the State of Tlaxcala, a campaign for adequate fire prevention at each location, and creates contingency plans for each municipality.

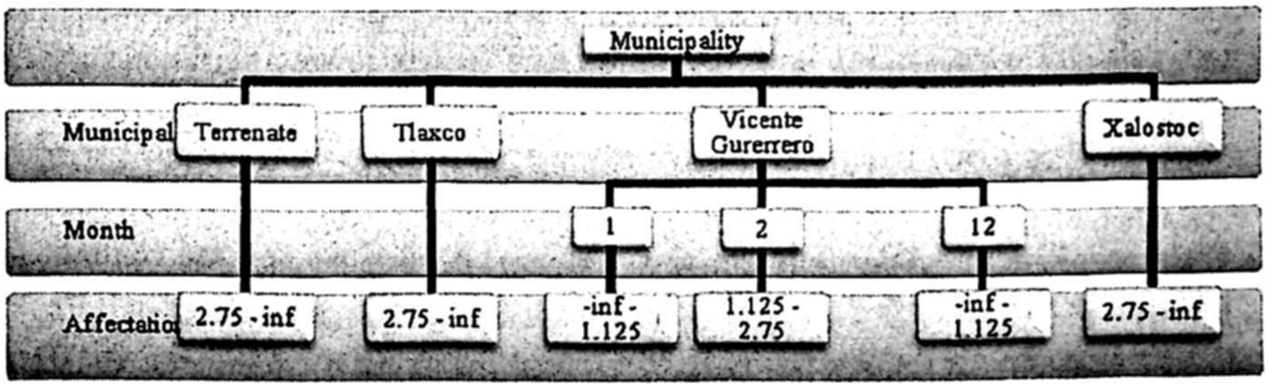

Figure 2 Decision Tree.

5.2 Association Rules

The association rules technique was used to analyze the relationship between municipality, month and area concerned. The association algorithm implemented in WEKA is *PredictiveApriori* algorithm. For simplicity, a filter attribute discretization in four intervals for affected area to explore the most significant relationships was applied. The algorithm is computed with the default parameters. For this case, among the most significant rules are obtained are:

```
1. Municipality=calpulalpan Beginning_Month=5 9 ==> Affected_Area='(2.75-inf)' 9 acc:(0.9873)
2. Municipality=xaltocan 7 ==> Affected_Area='(2.75-inf)' 7 acc:(0.9793)
3. Municipality=altzayanca 17 ==> Affected_Area='(2.75-inf)' 16 acc:(0.96005)
4. Municipality=terrenate Beginning_Month=5 5 ==> Affected_Area='(2.75-inf)' 5 acc:(0.95785)
5. Municipality=terrenate Beginning_Month=3 4 ==> Affected_Area='(2.75-inf)' 4 acc:(0.9337)
6. Municipality=juan cuamatzi Beginning_Month=6 4 ==> Affected_Area='(-inf-1.125]' 4 acc:(0.9337)
7. Municipality=panotla Beginning_Month=5 4 ==> Affected_Area='(2.75-inf)' 4 acc:(0.9337)
8. Municipality=xicohtencatl Beginning_Month=1 4 ==> Affected_Area='(2.75-inf)' 4 acc:(0.9337)
9. Municipality=huamantla Beginning_Month=6 4 ==> Affected_Area='(-inf-1.125]' 4 acc:(0.9337)
```

These rules provide information not so trivial: 98% of fires in Calpulalpan generated in the month of May exceed 2.75 hectares affected, and 97% generated in Xaltocan fire affected more than 2.75 hectares.

It is significant to note that the affected area in the municipalities listed with a high rate of fire is lower, as in the case of rule 9, which states that 93% of fires in the Municipality of Huamantla in the month of June show a lower affectation to 1,125 acres.

This analysis concludes that there are municipalities that do not have a high incidence of fires, but with a high degree of involvement, which must be considered for prevention campaigns, floor cleaning, training brigades and creating contingency plans.

6 Conclusions

Forest fires are considered a major problem in both forests and for society in general. It is clear the need for explicit information especially about the phenomenon in order to analyze possible occurrence patterns that could help make the surveillance and protection of the most affected areas. In this sense, this work was developed with support from WEKA and data analysis methods to assist the interpretation of information.

A data warehouse has been created, which contains information on forest fires that have occurred in the state of Tlaxcala, and then they are analyzed by data mining tools. This has allowed us to know in great detail the behavior of forest fires in Tlaxcala. Overall it can be concluded that:

The fire´s behavior in every month of the year was analyzed in order to discover what months have highest rate of fire. Clustering was applied concluding that March followed by Abril was the months with the highest rate of fire, where rainfall is scarce and dry land creating an ideal setting for the generation and spread of fire.

The fires causes were also analyzed by applying *K-Means method,* concluding that fires are generally initiated by agricultural activities followed by bonfires of strollers.

Using a decision tree, generated by the C4.5 algorithm, a relationship between the municipality and affected area was found. The affected area and in some cases the previous relationship is taken for months.

Finally, association rules helped to identified municipalities with a high degree of involvement in forest resources, concluding that these municipalities are among those with a higher incidence of fires.

7 References

[1]. Sistema estatal de protección civil de Chiapas, Ciencias de la tierra para la sociedad, tema: incendios forestales. On line jan-2013: http://www.proteccioncivil.chiapas.gob.mx/ciencia/ciencia&tierra/incendios.pdf

[2]. Detección de incendios forestales a través de imágenes digitales usando árboles de clasificación. System by Arturo Bustamante Blanco. On line jan-2013:< http://perseo.cs.buap.mx/bellatrix/tesis/TES1411.pdf>

[3]. Análisis de variables sociodemográficas que inciden en incendios forestales, a través de técnicas de Data Mining. System by Hans Carlos Lucero Salinas. On line jan-2013: <http://www.cifag.cl/_file/file_294_tesis_hans_lucero.pdf>

[4]. Aplicación de un Sistema de Información Geográfica al análisis de los datos de incendios forestales en España, System by Rosa Almudena Seco Granja On line jan-2013: <http://digital.csic.es/handle/10261/25971>

[5]. J. Hernández-Orallo, M. J. Ramírez-Quintana y C. Ferri. Introducción a la Minería de Datos. Prentice Hall / Addison-Wesley, 2004.

[6]. Jiawei Han,Micheline Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, Third Edition 2012.

[7]. William H. Inmon, Derek Strauss, Genia Neushloss. DW 2.0: The Architecture for the Next Generation of Data Warehousing. Morgan Kaufmann, 2008.

[8]. Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[9]. Stuart P. Lloyd. Least squares quantization in pcm. IEEE Transactions on Information Theory, 28:129-137, 1982.