© International Journal of Combinatorial Optimization Problems and Informatics, Vol. 1, No. 1, May-Aug 2010, pp. 61-67. ISSN: 2007-1558.

Spatial Data Mining of a Population-Based Data Warehouse of Cancer in Mexico

Joaquín Pérez-Ortega[1], Fátima Miranda-Henriques[2], Gerardo Reyes-Salgado[1], René Santaolaya-Salgado[1], Rodolfo A. Pazos-Rangel[3], Adriana Mexicano-Santoyo[1]
[1]Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), México
[2]Secretaría de Saúde do Estado de Pernambuco, Brasil
[3]Instituto Tecnológico de Ciudad Madero, México
{jperez, greyes, rene, amexicano}@cenidet.edu.mx,fhenriques@saude.pe.gov.br, pazos@yahoo.com.mx

Abstract. In the health sciences area, data mining applications have had a fast growth due to its results concerning the generation of patterns of interest; however, its application to spatial population-based databases has been scant. This paper shows the results obtained by applying a spatial data mining system of our making to a real population-based data warehouse of cancer mortality in Mexico. The system consists of a pattern generator module, which uses a variant of a clustering algorithm proposed by us, and a spatial visualization module. Several interesting and potentially useful patterns of stomach cancer were found in the northwest of Mexico, which show promising results for extending the use of data mining in the area of epidemiology.

Keywords: Bioinformatics, Pattern Recognition, Spatial Data Mining, Learning Technologies.

1 Introduction

The purpose of data mining is to obtain unknown patterns from massive or large databases, which can potentially have a large value or interest for an organization Adriaans and Zantinge (1996)[1], Berry and Linoff (2004)[2], Flouris and Duffy (2006)[3], Larose (2006)[4], Kamath (2009)[5], Thangavel et al. (2006)[6], Pyle (1999)[7].

The diagnosis and treatment of cancer, on one hand, is expensive and absorbs an important part of the public health budgets of many countries Giannopoubu (2008)[8], including Mexico, and on the other hand, it has a large social impact since the patient's and his relative's lifestyles are affected due to the health care required by the patient. The knowledge of the standards of mortality and the space distribution of the illness among different regions and districts of the country, contributes to identify hypotheses to probable causal associations. Additionally, it is useful for directing prevention and control measures based on the regional differences identified and for reducing diagnosis and treatment costs. The standards identified for the present study can direct future epidemiological studies about stomach cancer mortality. The present method could be extended to other cancers and illnesses.

The occurrence of malignant stomach tumors is strongly related to the social economic standards of the populations Witten and Frank (2005)[9], Faggiano et al. (1997)[10], Bouchardy et al. (1993)[11]. This work shows the results obtained by applying a data mining system to a real database of stomach cancer mortality from Mexico. The system was developed ad hoc and consists of a pattern generator subsystem and a visualization subsystem.

1.1 Related Work to the Application of Data-Mining Techniques to Health-related Databases

In recent years the use of data mining applied to clinical cancer data has increased, some examples are the works in Kamath (2009)[5], Nomura (1996)[12], Labib and Malek (2005)[13], Mullins and Siadaty (2006)[14], Wheeler (2007)[15], Maheswaran et al. [16]. However, the application of data mining to cancer epidemiologic data has been very limited Berry and Linoff (2004) [2].

Received March 1, 2010 / Accepted March 29, 2010

In Berry and Linoff [2] a study is reported on the application of data mining to the analysis of epidemiologic data, where, specifically, the following techniques are mentioned: Classification and Regression Trees, Multivariate Adaptive Regression Splines, and Tree-Structured Classifiers. That paper presents interesting references on the application of data mining techniques. The paper concludes that the application of data-mining techniques to population databases has been limited and that its use may facilitate finding interesting patterns for this kind of data.

According to the specialized literature surveyed, no previous works have been reported where clustering techniques are applied to population-based data on cancer for obtaining mortality rate distributions.

2 Data Mining Methodology

2.1 Source of Data

In this research, real data from several official databases were used. The most important are described hereupon.

Mortality data for cancer. This data was extracted from the Núcleo de Acopio y Análisis de Información de Salud or NAAIS (Collection and Analysis Core on Health Information) [17] from the Instituto Nacional de Salud Pública or INSP (National Institute for Public Health). From this database, all the records on deaths from lung and stomach cancer for the year 2000 were selected, and out of 38 attributes of the table only two were included: death cause and district of the deceased.

Population and geographic data. The data on districts population were obtained from the Sistema Municipal de Bases de Datos or SIMBAD (Database District System) [18] from the Instituto Nacional de Estadística y Geografía or INEGI (National Institute of Statistics and Geography). The data on the geographical position of each district and maps of Mexico were also obtained from INEGI.

2.2 Data Preprocessing

For each district, the gross lung and stomach cancer mortality rate per 100,000 inhabitants for the year 2000 was calculated, using formula (1),

$$rate = \frac{deaths}{population} *100,000 \qquad \textbf{(1)}$$

where: *deaths* = number of cancer deaths in a district in the year 2000, *population* = district population for the year 2000.

The data on the geographical position of districts was transformed, specifically minutes and seconds, to their equivalent in fractions of degree.

As a result of the preprocessing, the data warehouse was populated for the application of several modeling techniques showed in Hernández et al. [19].

2.3 Data Modeling Techniques

For knowledge extraction, the data was modeled through clustering. To this end, a variant of the *K*-means algorithm devised by us was used, whose results were promising and satisfactory for the solution of the addressed problem.

Several improvements to the standard *K*-means algorithm [20] have been carried out, most of them related to the initial parameter values. In contrast, we propose an improvement using a new convergence condition that consists of stopping the execution when a local optimum is found or no more object exchanges among groups can be performed. For assessing the improvement attained, the modified algorithm (Early Stop *K*-means) was tested on six databases of the UCI repository [21], and the results were compared against SPSS [22], Weka [23] and the standard *K*-means algorithm. Early Stop *K*-means attained an average reduction in the iterations number of 40%, 49% and 32% with respect to standard *K*-means, SPSS and Weka, with a simultaneous average improvement in the solution quality of 1.67%, 3.30% and 19.83% with respect to those algorithms, reported in Pérez et al. [24].

As a result of applying the algorithm (Early Stop *K*-means) to the data warehouse, patterns were generated as groups of districts with similar parameters regarding geographical position and mortality rate.

The results of the clustering algorithm are lists of element groups including the centroid of each group. Since the interpretation of patterns expressed in list form was difficult for specialists, it was considered convenient to present groups information in tabular form (Tables 1, 2, 3, 4); however, this was not entirely satisfactory, since there exist over 2000 districts in Mexico and it is not easy to locate a district on a map, specially if it is relatively unknown. In order to solve this problem with interpretation, a visualization subsystem was developed, which is described in the following section.

2.4 Visualization Subsystem

The cartographic visualization subsystem permits selecting and drawing on a map of Mexico one or more of the patterns generated by the clustering algorithm. The subsystem shows on the map the group centroids as small circles and the group elements (districts) as black dots; while the membership of an element to a group is indicated by a line that joins it to the group centroid (Figs. 1, 2, 3 and 4).

The visual representation of groups permitted to enhance the knowledge obtained facilitating the interpretation and assessment of the results by the system users.

3 Experimental Results

A set of experiments were conducted using the data mining system on the cancer data warehouse, selecting districts with population greater than 100,000 for the year 2000, and setting the number of groups *k* equal to 5, 10, 15, 20, and 30. The best result was obtained for *k* equal to 20 according to the specialists.

Out of the 20 groups generated, the two groups with the largest average rates were selected for attracting the most interest. Figure 1 shows group 1 which corresponds to the Chiapas heights in the southeast of Mexico. This group served to validate the data mining method used in this prototype, since clinical investigations have reported a high mortality rate for gastric cancer. Such investigations have claimed that one of the factors that contribute to the development of this type of cancer in the region is a chronic infection caused by a bacteria called helicobacter pylori (HP) [25]. The district details of the group and the mortality rates are shown in Table 1 including the mean value and the standard deviation.

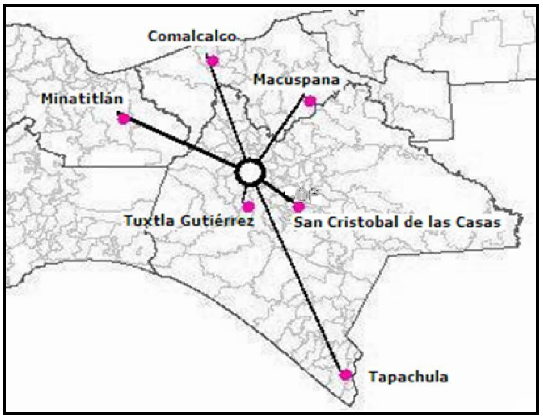

Fig. 1. Group1.

Table 1. Group1

District	Deaths	Population	Rate
Minatitlán	14	153001	9.15
Comalcalco	14	164637	8.50
Tapachula	21	271674	7.73
San Cristóbal	9	132421	6.80
Macuspana	9	133985	6.72
Tuxtla Gutiérrez	28	434143	6.45
		Average	7.56
		Standard deviation	0.99

Additionally, we report as a new finding another pattern of interest and potential usefulness in the northwest region: group 2 (Figure 2, Table 2), which has an average mortality rate larger than that of group 1.

According to the specialized literature, there are no studies reporting a concentration of high mortality rates for stomach cancer in this region. A possible explanation for this situation is that cancer statistics are usually analyzed statewise and group 2 spans two states.

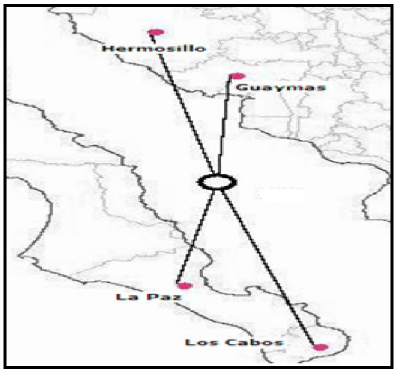

Fig. 2. Group2.

Table 2. Group 2

District	Deaths	Population	Rate
Guaymas	15	130329	11.52
Hermosillo	48	609829	7.87
La Paz	14	196907	7.11
Los Cabos	7	105469	6.64
		Average	8.28
		Standard deviation	1.92

As a result of the analysis of the patterns generated for lung cancer, two groups with the largest average rates were selected, since they attract the most interest. Figure 3 and Table 3 show group 1 in the northwest of Mexico.

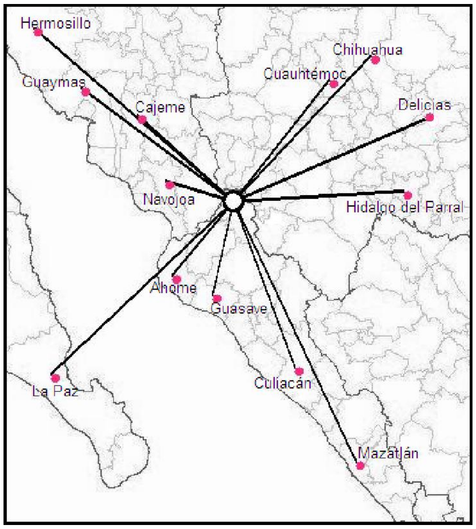

Fig. 3. Group 3.

Table 3. Group 3

District	Deaths	Population	Rate
Cajeme	67	356290	18.80
Hermosillo	104	609829	17.05
Hidalgo del Parral	16	100821	15.86
Culiacán	113	745537	15.15
Navojoa	21	140650	14.93
Ahome	52	359146	14.47
Guasave	39	277402	14.05
Delicias	16	116426	13.74
La Paz	27	196907	13.71
Mazatlán	51	380509	13.40
Guaymas	17	130329	13.04
Cuauhtémoc	14	124378	11.25
Chihuahua	75	671790	11.16
		Average	14.3546
		Standard deviation	2.03128

Another pattern of interest and potential usefulness was discovered in the north: group 2 for lung cancer (Figure 4, Table 4), which has a large average mortality rate.

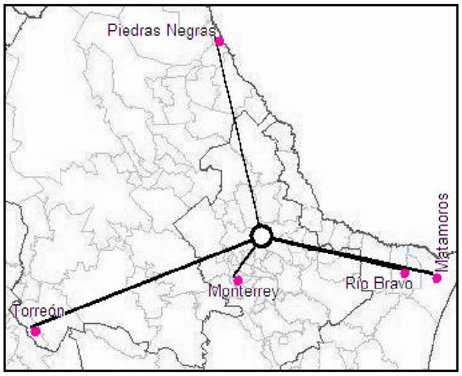

Fig. 4. Group 4.

Table 4. Group 4

District	Deaths	Population	Rate
Río Bravo	14	104229	13.43
Matamoros	54	4148141	12.91
Torreón	65	529512	12.27
Monterrey	113	1110997	11.97
Piedras Negras	15	128130	11.70
		Average	12.456
		Standard deviation	0.7065

4 Conclusions

One of contributions of this work is the development of a population-based data warehouse on cancer (stomach, lung, etc.) from official data sources, and specifically the integration of geographical data of districts with cancer statistical data.

A variant of the *K*-means clustering algorithm devised by us (Early Stop *K*-means) was designed and implemented. Our algorithm proved to be an adequate option for spatial clustering into regions.

A geographic visualization subsystem was implemented, which permitted to show the centroid and the districts of groups on a map, allowing to depict patterns as nation regions with similar mortality rates. This tool proved to be particularly useful for assessing and communicating the results because its visual expressiveness.

A set of experiments were conducted using the data mining system for different numbers of groups (k), and k = 20 yielded the best result.

As a result of the analysis of the patterns generated for stomach cancer, a well known pattern of districts with high mortality rate in the southeast of Mexico was determined (Figure 1, Table 1), which served for validating the method used in this prototype. Additionally, we report as a new finding another pattern of interest and potential usefulness in the northwest region: group 2 (Figure 2, Table 2), which has an average mortality rate larger than that of group 1.

As a result of the analysis of the patterns generated for lung cancer, a pattern of districts with high mortality rate in the northwest of Mexico was determined (Figure 3, Table 3) and another pattern of interest and potential usefulness in the north region: (Figure 4, Table 4), which has a large average mortality rate.

We consider that our data mining system can be improved by developing functions for adjusting mortality rates by age intervals and gender. Another improvement could consist of the integration of modules for the analysis of mortality rates for other diseases besides cancer.

Finally, we consider that the patterns generated by the data mining system, which are expressed as groups of districts with similar location and mortality rate parameters, can be useful as an aid tool for studies on cancer and for decision making concerning the allocation of resources for organizing specialized services for cancer prevention and treatment.

References

[1] Adriaans, P. and Zantinge, D.: Data Mining. Addison-Wesley (1996). GB Search.

[2] Berry, M. and Linoff, G.: Data Mining Techniques for Marketing, Sales, and Customer Relationships Management. Wiley Publishing, Inc. (2004). GB Search.

[3] Flouris, A. and Duffy, J.: Application of Artificial Intelligence Systems in the Analysis of Epidemiological Data. In European Journal of Epidemiology, Vol. 21, No. 3 (2006)167-170.

[4] Larose, T.D.: Data Mining Methods and Models. John Wiley & Sons, New Jersey (2006).GB Search.

[5] Kamath, C.: Scientific Data Mining, A Practical Perspective. SIAM (2009). GB Search.

[6] Thangavel, K., Jaganathan, P. and Esmy, P.: Subgroup Discovery in Cervical Cancer Analysis Using Data Mining. International Journal on Artificial Intelligence and machine learning, Vol. 6, No. 1 (2006) 29-36. View Item.

[7] Pyle, D.: Data Preparation for Data Mining. Morgan Kauffman Publishers, Inc. (1999). GB Search.

[8] Giannopoubu, E.: Data Mining in Medical and Biological Research. In-Teh (2008). GB Search.

[9] Witten, I.H. and Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. 2nd Edition. Morgan Kaufmann. San Francisco (2005). GB Search.

[10] Faggiano, F., Partanen, T., Kogevinas, M. and Boffetta, P.: Socioeconomic Differences in Cancer Incidence and Mortality. In: Kogevinas, M., Pearce, N., Susser, M., Boffetta, P. (eds.): Social Inequalities and Cancer, No. 138. International Agency for Research on Cancer (IARC), Lyon, France (1997) 65-176. GS Search.

[11] Bouchardy, C., Parkin, D.M. and Khlat, M.: Education and Mortality from Cancer in São Paulo, Brazil. Annals of Epidemiology, Vol. 3, No. 1. International Agency for Research on Cancer (IARC) Lyon, France (1993) 64-70. GS Search.

[12] Nomura, A.: Stomach Cancer. In: Schottenfeld, D., Fraumeni. J.F. Jr. (eds.): Cancer Epidemiology and Prevention. Oxford University Press, New York (1996).

[13] Labib, N.M. and Malek, M.N.: Data Mining for Cancer Management in Egypt Case Study: Childhood Acute Lymphoblastic Leukemia. Transactions on Engineering. Computing and Technology, Vol. 8 (2005) 309-314. GS Search.

[14] Mullins, I.M. and Siadaty, M.S.: Data Mining and Clinical Data Repositories: Insights from a 667,000 Patient Data Set. Computers in Biology and Medicine, Vol. 36, No. 12 (2006) 1351-1377.

[15] Wheeler, D.: A Comparison of Spatial Clustering and Cluster Detection Techniques for Childhood Leukemia Incidence in Ohio, 1996-2003. International Journal of Health Geographics, Vol. 6, No. 13 (2007). GS Search.

[16] Maheswaran, R., Strachan, D., Dodgeon, B. and Best, N.: A Population-based Case-control Study for Examining Early Life Influences on Geographical Variation in Adult Mortality in England and Wales Using Stomach Cancer and Stroke as Examples. International Journal of Epidemiology, Vol. 31 (2002) 375-382. GS Search.

[17] NAIIS, Instituto Nacional de Salud Pública, Núcleo de Acopio y Análisis de Información en Salud (2003). View Item.

[18] SIMBAD, Instituto Nacional de Estadística, Geografía e Informática [INEGI], Sistema Municipal de Base de Datos (SIMBAD) (2007). View Item.

[19] Hernández Orallo, J., Ramírez Quintana, M.J., Ferri Ramírez, C.: Introducción a la Minería de Datos, Pearson Educación, Madrid (2004).

[20] MacQueen J.B.: Some Methods for classification and Analysis of Multivariate Observations, fifteenth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, (1967) 281–297.

[21] Asuncion, A., Newman, D.: UCI Machine Learning Repository. Irvine, CA: University of California, Department of Information and Computer Science, 2007. View Item.

[22] SPSS, Ibérica, an IBM Company. View Item.

[23] Weka, University of Waikato, available at: View Item.

[24] Pérez, J., Pazos, R., Cruz, L., Reyes, G., Basave, R. and Fraire, H.: Improving the Efficiency and Efficacy of the K-means Clustering Algorithm Through a New Convergence Condition, Lecture Notes in Computer Science, Vol. 4707 (2007) 674-682. GS Search.

[25] Mohar, A., Ley, C., Guarner, J., Herrera-Goepfert, R., Sánchez L., Halperin D. and Parsonnet J.: Alta Frecuencia de Lesiones Precursoras de Cáncer Gástrico Asociadas a Helicobacter Pylori y Respuesta al Tratamiento, en Chiapas, México. Gaceta Médica de México, Vol. 138, No.5, (2000) 405-410.