Fuzzy Model for Municipalities Classification in the State of Hidalgo in Mexico

AMAURY CABALLERO, KANG YEN
Florida International University
MIRIAM ALVAREZ , GILBERTO PEREZ-LECHUGA
Universidad Estatal Autónoma del Estado de Hidalgo
USA - MEXICO

Abstract. The state of hidalgo, in Mexico, is looking for a model to classify 84 municipalities according to certain economic and social parameters. The model will permit the development of policies for improving the conditions of those municipalities that are lagging behind. A previous model was created using statistical methods, applying the factorial analysis. A model using databases looks more practical, but the classical SQL methodology using crisp qualifiers causes difficulties in some decision making processes especially when it is mandatory to move to the definition of practical indicators or categories and to evaluate them according to certain practical assumptions. Recently fuzzy logic has been embedded in SQL to improve its performance. In this paper, we use a Fuzzy model for solving the stated problem. The result is compared with that from previous study with statistics method.

Key Words: Fuzzy Logic, Data Bases, Classification, Modeling

1 Introduction

In many developing or underdeveloped countries one interesting and critical task is the allocation of limited resources. The priority and level of development for different states, provinces, counties, or municipalities has to be analyzed before efficient distribution of resources. The strategic plan needed in resources allocation in the state of Hidalgo in Mexico is one of the examples. The government is interested in classifying all its 84 municipalities based on some predefined indicators. The result can provide them with the information about the stage of the economic development among different municipalities. This classification can assist the politicians make policies to boost the development of those municipalities that are lagging behind. It is expected that the result can be generalized in the study of the similar problem for other states in Mexico.

The objective of this paper is to show an easy-to-use methodology, which is based on fuzzy logic databases, in municipality classification tasks. The same problem was studies in [4], where a combination of two exploratory techniques: (1) the principal components analysis and (2) a hierarchical classification technique has been used.

2 Methodologies

2.1 Statistical Methods

One of the exploratory methods in statistics is the factorial analysis. The principal components analysis (PCA) is one of the basic models of the factorial analysis when the objective is to predict the minimum number of necessary factors to justify the maximum portion of the variance represented in the original variables. By means of a mathematical procedure [1] a smaller group of uncorrelated variables can be generated and called “principal components” that allows to identify a structure or the ownership from each individual to a specific group.

It is possible to complete the factorial analysis with a classification carried out on the total space or a sub-space defined by the first few significant factors. The classes consider the actual dimension of the cloud of points. The classification algorithms, particularly those of agglomeration, are locally robust since the low parts of the clusters (the nodes corresponding to the smallest distances) are independent of some isolated points [2]. This complementary character between the factorial analysis and the complete classification might complete the knowledge of data structure and allows a better interpretation of the data [3].

The inference and confirmatory statistics, however, allow us to validate the hypotheses formulated a priori (or after an exploratory phase) and to extrapolate the results to a wider population. This kind of statistics makes use of explanatory methods dedicated to explain and predict a variable starting from one or several explanatory variables, following the decision rules. Among these methods we find multiple and logistical regressions methods, discriminant analysis and analysis of the variance.

In [4], this method was applied to obtain the indicators to be used in grouping the municipalities in the state of Hidalgo in Mexico. From this, the municipalities were divided into three groups: less, medium and well development. Figure 2 shows a map of the state of Hidalgo after the classification was carried out. In the current work of using fuzzy concept, the same indicators obtained in [4] are adopted, but the proposed method for solving the same problem is to develop a query from a fuzzy database. With fuzzy approach, the classification is simpler and has flexibility, which cannot be obtained with the statistical method.

2.2 Fuzzy SQL Method

Although methods exist to facilitate the manipulation of data in a database, the most popular one is the declarative language known as Structured English Query Language (SQL). SQL can easily capture the mechanical intent of a query, but it lacks the ability to capture the semantics of a query. We can group and slice up collections of data in a variety of ways, but each division of the record collection proceeds along crispy lines [7]. For example, let’s assume that the cost of a product is defined through the regions *Low, Moderate,* and *High*. The situation can be shown in Figure 1.

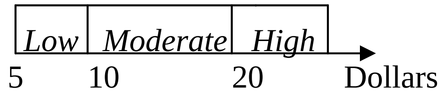

Figure 1. Regions defining the cost

The statement could be written as:

FROM *PRODUCT*
WHERE *COST < 10*

In this case, the selected rows have a COST lower than 10. There are at least two problems exposed:

1) Not everyone will agree with the rigid boundary imposed to each region.
2) For points near to the boundary, the query intent may not be captured.

Fuzzy logic will find solutions for both problems. In a fuzzy system, each of the regions that illustrate a particular situation is covered by more than one membership function, which may be triangular, trapezoidal or unitary Bell functions.

The main tasks considered in defining a fuzzy model are [8]:

1) Choose an appropriate family of membership functions.
2) Interview human experts to determine parameters of the membership functions.
3) Refine the parameters of the membership functions using regression or optimization techniques.

Recently, efforts have been made in the development of a methodology that reaches beyond the limitations of SQL. The use of fuzzy logic is one of them. Cox presents a method of composition for sorting the different alternatives using the Compatibility Index (*CI*), which can be expressed as the average of the degrees of membership μ for each alternative [9].

$$CI = \frac{1}{N}\sum_{i=1}^{N}\mu_i(x) \qquad (1)$$

Where $\mu_i(x)$ is the degree of membership for the parameter *i* of each alternative and *N* is the number of parameters considered in the process. Logically, one has to ensure that the same parameters are used for all the alternatives, and the number *N* is the same in all computations.

If the attributes do not have equal contribution, then relative weights must be involved. For example, if a particular user with an alternative having three parameters **B**, **P**, and **F** feels that **P** should be 1.5 times as important as **B,** then the weights assigned would be 1.0 for **B** and 1.5 for **P**. If we consider that **P** is the most important factor, then we would next rank **P** and **F.** If the relative weight obtained as a result of this pair-wise comparison is 2.0, then the relative weights for all three criteria become:

B:		1.0
P:		1.5
F:	1.5 x 2.0 =	3.0
Total		5.5

If these are normalized the resulting weights are as follow:

B: 1.0/(1.0 + 1.5 + 3.0) = 0.182
P: 1.5/(1.0 + 1.5 + 3.0) = 0.273
F: 3.0/(1.0 + 1.5 + 3.0) = 0.545

Then the modified formula for the compatibility index calculation becomes:

$$CI = \sum_{i=1}^{N} w_i\mu_i(x) \qquad (2)$$

Where w_i is the weight of the attribute *i*.

To create a query from a database it is necessary to follow the steps [9]:

1) Define the utilized parameters (columns in the database).
2) Organize different regions for each parameter (Labels).
3) establish the range of variation of each parameter (universe of discourse) and the boundaries for each region (scope/domain)
4) Apply the selection algorithm (Crisp –Fuzzy).

The steps 1) to 3) may be used independently of the implemented selection algorithm. When crisp variables are used, SQL provides a straight forward solution, but with the previously mentioned restrictions.

The factor, which influences most on the performance in fuzzy logic application, is the definition of membership functions. Typically experts are needed for defining the labels, the universe of discourse, the scope, and the shape of membership functions. The membership functions can be defined using experts from the specific application. Another way of solving this is using surveys. Another issue to be considered is the α-cut threshold definition. Moving the α-cut threshold upwards, only highly compatible participants are selected [10]. Moving it down, a wider but less compatible set of rows is selected. In a first approximation the α-cut threshold can be selected as zero. The compatibility index obtained at the end of the process will represent by itself the more or less compatibility of the selected rows in the selection process.

3 Analysis of Municipalities in Hidalgo

The data used in this study were obtained from the 2000 Mexico National Census [5]. The information of each family in the state of Hidalgo was collected and used. The initial database is a matrix with 84 municipalities and 6 indicators. Fuzzy logic database software [6] has been used for the classification purposes.

To apply the analysis to the proposed problem, the indicators in [4] have been used: percentage of population dedicated to agriculture (sect-prim); percentage of population dedicated to commerce and services (sect-terc); percentage of population over 15 years old only with primary education (pob-15-prim); percentage of population over 18 years old with college preparation (pob-18-uni); percentage of population, receiving between 1 and 2 minimum salaries (sal 1-2-min); percentage of population receiving more than 10 minimum salaries (sal>10-min). The membership function for the indicator "percentage of population dedicated to agriculture (sect-prim)" is presented in Figure 3. The membership functions for the other indicators have a similar form, and their domains are:

- Sect-terc:
  Low: from 2 through 8;
  Medium: from 4 through 14;
  High: from 8 through 25;
  Very high: from 14 through 30;
- Pob-15-prim:
  Low: from 10 through 18;
  Medium: from 13 through 23;
  High: from 18 through 27;
  Very high: from 23 through 30;
- pob-18-uni:
  Low: from 0 through 3;
  Medium: from 2 through 4;
  High: from 3 through 6;
  Very high: from 4 through 25;
- Sal>10-min:
  Low: from 0 through 2;
  Medium: from 1 through 3;
  High: from 2 through 6;
  Very high: from 3 through 6;
- Sal>10-min:
  Low: from 10 through 20;
  Medium: from 15 through 30;
  High: from 20 through 40;
  Very high: from 30 through 50.

The comparison was made by defining one "average" municipality, defined as per:

"pob-15-prim" is "LOW";
"pob-18-uni" is "VERY HIGH";
"sal-1-2min" is "LOW";
"sal>10min" is "VERY HIGH";
"sect-prim" is "MEDIUM"; and
"sect-terc" is "MEDIUM".

The membership functions are defined after consultation with specialists from the Universidad Autonoma Estatal del Estado de Hidalgo (State University of Hidalgo). According to the software [6], the *CI* for the first five municipalities is given in Table 1.

All the municipalities included on Table 1 coincide with that obtained in the "Medium Development" condition from Figure 2. This can be used as a confirmation that the used membership functions were properly defined.

The information for all the indicators used for the evaluation, for each of the given municipalities in Table 1, is given in Table 2.

As can be seen, the selection indicates which municipalities are more or less compatible under the selected conditions. When selecting these conditions, other factors may be taken into consideration. For example, if the analysis is made looking for rural municipalities, then the percentage of population dedicated to agriculture (sect-prim) should be taken as HIGH or VERY HIGH and, of course the municipalities more compatible with this new condition will be different ones.

4 Conclusions

The paper presents a methodology for comparing different municipalities, based on a group of previously defined indicators considered the priority and important for the evaluation, when the information is stored in a fuzzy logic database. The method results more simple, accurate and flexible than the previously used statistical methods.

The extension of the model previously studied, provides a refinement that allows the user of the basic model to better reflect their concerns in the ranking of the municipalities. The power of the fuzzy logic model is that it uses imprecise terms to arrive at ‘crisp’ values. Modifying these ‘crisp’ values by establishing weights, reflecting the importance of various attributes, is a logical next step.

The compatibility index (*CI*) gives a good measure in so far as a solution coincides with the previously imposed conditions. If the compatibility index is too low for any of the possible selections, it is necessary to revise the created fuzzy logic model: number and domain of the membership functions for each attribute. The results show that the fuzzy logic selection indicates which municipalities are more or less compatible with the selected conditions. If other factors are taken into consideration, the results will be different, depending on the input assumptions that can be changed, depending on the real situation under consideration.

References

[1] Anderson, T.W., *An Introduction to Multivariate Statistical Analysis*, John Wiley, New York, 2nd edition,1984

[2] Sokal, R.R. & Sneath, P.H.A., *Principles of Numerical Taxonomy*, Freeman and Co., San Francisco, 1963

[3] Lebart, L., Morineau, A., & Poiron, M., *Statistique exploratoire multidimensinnelle,* Dunod, Paris, 1995

[4] Alvarez, M.M., *Factorial Methods and Discriminant Analysis for the classification of regions of poverty in Mexico*, WSEAS Trans. on Computers. Issue 5. Vol. 3. Nov. 2004. pp.1587–1591.

[5] INEGI, 2000. *Indicadores Sociodemograficos* (1930-1998), Mexico.

[6] *Fuzzy Query 1.0*. (1998), Sonalysts, Inc., 215 Parkway North, Waterford, CT. 06385

[7] Caballero, A., *Construction Project Management Using Fuzzy Logic*, Journal of Information. Vol 6, No. 4, October 2003. Japan. pp 463 – 474.

[8] Jang J-SW.R., et al., *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning Machine Intelligence*, Prentice Hall, 1996

[9] Cox, E.D., *Fuzzy Logic for Business and Industry*, Rockland: Charles River Media, 1995

[10] Caballero, A.A., & Dye, J.M., *Comparison of Construction Firms Based on Fuzzy Set*s, J. of Construction Education, Vol.III, No.3, 1999, pp.305-312

[11] Chiang, D. et al., *Fuzzy Information in extended Fuzzy Relational Databases*, Fuzzy Sets and Systems 92, 1997, pp.1-20

[12] Buckles B.P., & Petry, F.E., *A Fuzzy Representation of Data for Relational*

Databases, Fuzzy Sets and Systems, Vol.7, No.3, 1982, pp.213-226

[13] Galindo, J. et al., *Applying Fuzzy Databases and FSQL to the Management of Rural Accommodation*, Elsevier Science, 2003

[14] Marin, N. et al., *Complex Object Comparison in a Fuzzy Context*, Elsevier Science, 2003

Table 1. Compatibility Index

Municipality	*Compatibility Index*
4	0.468
59	0.460
7	0.456
38	0.454
57	0.453

Table 2. Information for the Municipalities Given in Table 1.

Municipality	*sect_prim*	*sect-terc*	*pob_15_prim*	*pob_18_uni*	*Sal_1-2min*	*sal>10min*
4	9.63	8.89	22.34	1.5	0.45	32.68
59	9.97	7.38	30.63	2.16	0.73	39.12
7	12.31	8.7	24.8	1.88	0.31	38.3
38	10.02	10.52	20.81	2.57	0.54	36.15
57	10.56	9.62	23.24	2.16	1.1	40.12

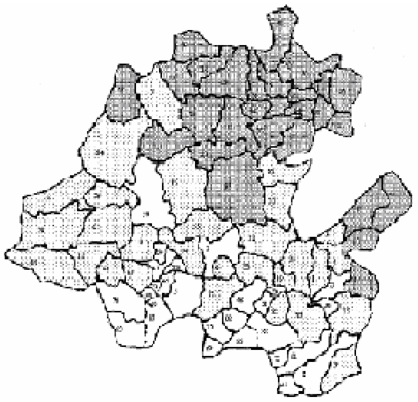

Figure 2. Map of the State of Hidalgo
-White: High Development
-Doted: Medium Development
-Black: Low Development

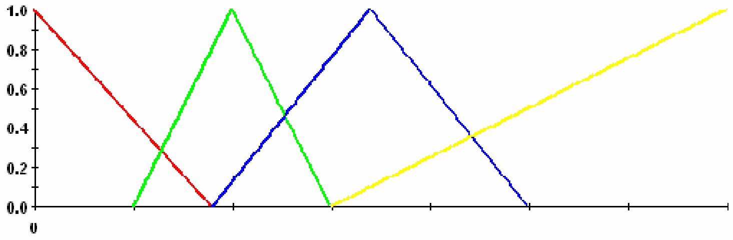

Figure 3. Membership functions for one indicator: Low, Medium, High, Very high.