Engineering Applications of Artificial Intelligence 22 (2009) 57–65

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

A search space reduction methodology for data mining in large databases

Angel Kuri-Morales [a,*], Fátima Rodríguez-Erazo [b]

[a] Department of Computer Science, Instituto Tecnologico Autonomo de Mexico, Rio Hondo No. 1, Col. Tizapan San Angel, C.P. 01000 México D.F., Mexico
[b] Posgrado en Ciencia e Ingeniería de la Computación, Universidad Nacional Autónoma de México, Ciudad Universitaria, Del. Coyoacán, C.P. 04510 México D.F., Mexico

ARTICLE INFO

Article history:
Received 20 April 2008
Accepted 26 April 2008
Available online 9 July 2008

Keywords:
Large databases
Sampling
Space reduction
Preprocessing
Clustering
Instance selection
Data mining

ABSTRACT

Given the present need for Customer Relationship and the increased growth of the size of databases, many new approaches to large database clustering and processing have been attempted. In this work, we propose a methodology based on the idea that statistically proven search space reduction is possible in practice. Two clustering models are generated: one corresponding to the full data set and another pertaining to the sampled data set. The resulting empirical distributions were mathematically tested to verify a tight non-linear significant approximation.

1. Introduction

Nowadays, commercial enterprises are importantly oriented to continuously improving customer–business relationship. With the increasing influence of CRM[1] Systems, such companies dedicate more time and effort to maintain better customer–business relationships. The effort implied in getting to better know the customer involves the accumulation of enormous databases where the largest possible quantity of data regarding the customer is stored.

Data warehouses offer a way to access detailed information about the customer's history, business facts and other aspects of the customer's behavior. The databases constitute the information backbone for any well-established company. However, from each step and every new attempted link of the company to its customers the need to store increasing volumes of data arises. Hence, databases and data warehouses are always growing up in terms of number of registers and tables which will allow the company to improve the general vision of the customer.

Data warehouses are difficult to characterize when trying to analyze the customers from company's standpoint. This problem is generally approached through the use of data mining techniques (Palpanas, 2000). To attempt direct clustering over a database of several terabytes with millions of registers results in a costly and not always fruitful effort. There have been many attempts to solve this problem: for instance, with the use of parallel computation, the optimization of clustering algorithms, via alternative distributed and grid computing and so on. But still the more efficient methods are unwieldy when attacking the clustering problem for databases as considered above.

In this article (the initial version of this article can be found in Kuri and Rodríguez, 2007), we present a methodology derived from the practical solution of an automated clustering process over large database from a real large sized (over 20 million customers) company. We emphasize the way we used statistical methods to reduce the search space of the problem as well as the treatment given to the customer's information stored in multiple tables of multiple databases. (An interesting reference in regard with the some issues discussed here may be found in Ahlemeyer-Stubbe, 2001.)

Because of confidentiality issues the name of the company and the actual final results of the customer characterization are withheld.

The paper is organized as follows. First, we give an overview of the analysis of large databases in Section 2; next we give a clustering, sampling, and feature selection overview. In Section 3, we briefly discuss the case study treated with the proposed methodology. Explanation of the methodology follows in Section 4. Finally, we conclude in Section 5.

2. Analysis of large databases

To extract the best information of a database, it is convenient to use a set of strategies or techniques which will allow us to analyze large volumes of data. These tools are generically known as data mining (DM) which targets on new, valuable, and

* Corresponding author. Tel.: +52 55 56284000x3649; fax: +52 55 5628 4065.
E-mail address: akuri@itam.mx (A. Kuri-Morales).

[1] Customer Relationship Management.

0952-1976/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2008.04.003

non-trivial information contained in large volumes of data. It includes techniques such as clustering (which corresponds to non-supervised learning) and statistical analysis (which includes, for instance, sampling and multivariate analysis).

2.1. Clustering in large databases

Clustering is a popular data mining task which consists of processing a large volume of data to obtain groups where the elements of each group exhibit quantifiably (under some measure) small differences between them and, contrariwise, large dissimilarities between elements of different groups. Given its importance as a very important data mining task, clustering has been the subject of multiple research efforts and has proven to be useful for many purposes (see Jain et al., 1999).

Many techniques and algorithms for clustering have been developed, improved and applied (see Berkhin, 2002; Kleinberg et al., 2004). Some of them try to ease the process on a large database as in Guha et al. (1998), Peter et al. (2003) and Raymong and Jiawei (1994). On the other hand, the so-called "Divide and Merge" (see Cheng et al., 2006) or "Snakes and Sandwiches" (see Jagadish et al., 1999) methods refer to clustering attending to the physical storage of the records comprising data warehouses. Another strategy to work with a large database is based upon the idea of working with statistical sampling optimization (see Palmer and Faloutsos, 2000).

2.2. Sampling and feature selection

Sampling is a statistical method to select a certain number of elements from a population to be included in a sample. There exist two sampling types: probabilistic and non-probabilistic. For each of these categories, there exists a variety of sub-methods. The probabilistic better known ones include: (a) Random sampling, (b) Systematic sampling, and (c) Stratified sampling. On the other hand, the non-probabilistic ones include methods such as convenience sampling, judgment sampling, and quota sampling. Some of the several ways to select the elements from a data set are discussed in Liu and Motoda (2002). This field of research, however, continues to be an open one (see Zhu and Wu, 2006; Brighton and Mellish, 2002).

The use of sampling for data mining has received some criticism since there is always a possibility that such sampling may hamper a clustering algorithm's capability to find small clusters appearing in the original data (see Palmer and Faloutsos, 2000). However, small clusters are not always significant; such is the case of customer clusters. Since the main objective of the company is to find significant and, therefore, large customer clusters, a small cluster that may not be included in a sample is not significant for CRM.

Apart from the sampling theory needed to properly reduce the search space, we need to perform feature selection to achieve desirable smaller dimensionality. In this regard, we point out that feature selection has been the main object of many researches (see, e.g., Vu et al., 2006; Zhang et al., 2007), and these had resulted in a large number of methods and algorithms (see Fodor, 2002). One such method is "multivariate analysis". This is a scheme (as treated here) which allows us to synthesize a functional relation between a dependent and two or more independent variables. There are many techniques to perform a multivariate analysis. For instance, multivariate regression analysis, principal component analysis, variance and covariance analysis, canonical correlation analysis, etc. (see Hair et al., 1999). Here, we focus on the explicit determination of a functional which maximizes the resulting correlation coefficient while minimizing its standard error. Clearly, this approach requires a sufficiently large number of models to consider, as will be discussed in the sequel. The main advantage of this method vs. other alternatives lies in two facts: (a) no assumption is made on the distribution of the variables in the sample (which makes it of a more general nature) and (b) the computational cost is kept at a minimum. This last issue becomes fundamental when addressing very large databases, as is the case.

3. Case study

A data-mining project was conducted for a very large multinational Latin American company (one of the largest in Latin America), hereinafter referred to as the "Company". The Company has several databases with information about its different customers, including data about services contracted, services' billing (registered over a period of several years) and other pertinent characterization data. The Company offers a large variety of services to millions of users in several countries. Its databases are stored on IBM® Universal Database version 7.0. In our study, we applied a specific data-mining tool (which we will refer to as "the miner") which works directly on the database. We also developed a set of auxiliary programs intended to help in data pre-processing.

The actual customer information that was necessary for the clustering process was extracted from multiple databases in the Company. Prior to the data mining process, the Company's experts conducted an analysis of the different existent databases and selected the more important variables and associated data related to the project's purpose: to identify those customers amenable to become *ad hoc* clients for new products under development and others to be developed specifically from the results of the study. Due to the variety of platforms and databases, such process of selection and collection of relevant information took several months and several hundred man-hours.

The resulting database displayed a table structure that contains information about the characteristics of the customers, products or services contracted for the customer and monthly billing data over a 1-year period.

To test the working methodology the project team worked with a subset of 400,000 customers' registers, consisting of a total of 415 variables divided in nine data tables. Table 1 displays the characteristics of the data sources treated in this study.

As stated above, the main objective of the data-mining project was to characterize the customers of the Company allowing in the near future—in accordance to customer characteristics—to offer new services and/or increase sales to existent or new customers.

4. Methodology

In order to apply a methodology, whereupon the search space is efficiently and effectively reduced, it is necessary to comply

Table 1
Data sources

Table	Columns	Rows	Description
TFB	25	400,000	Customer billing
TINT	121	400,000	Internet services
TPK	49	400,000	Data package services
TGRL	11	400,000	Customer's general data
TAC	2	73	Supply areas
TCC	2	4	Customer's credit rank code
TPA	3	183	Customer's permanence
TLPC	121	400,000	Local services
TSD	85	400,000	Digital services

with several steps leading to the adequate representation and/or behavior of the data regardless of its primary origin.

- Data preprocessing
- Search space reduction
- Clustering

These steps are discussed as follows.

4.1. Data preprocessing

This step included data cleaning by exhaustively searching for incomplete, inconsistent or missing data (see Delmater and Hancock, 2001 or Perner, 2002). Additionally, we also had to transform non-numeric to numeric data. Resulting from this process, unrecoverable registers were deleted. The number of such deleted records, however, was not significant.

From the original multiple-tables structure, we defined a single-table view structure for which a process of denormalization was performed. This followed from an analysis of the key-structure. In this view, tables with the same key were merged and tables with different keys were included in the referenced tables as additional columns. The transformation resulted in a view with a structure with 415 attributes.

4.2. Search space reduction

To reduce the search space, we worked with the original data to obtain a sample which is not merely a subspace but, rather, one that properly represents the original (full) set of data. We reduce the set both in the horizontal (reducing the number of tuples) and in the vertical (reducing the number of attributes) to obtain the "minable view". Simultaneous reduction—horizontal and vertical—yields the smallest representation of the original data set. Vertical reduction is possible from traditional statistical methods, while horizontal reduction, basically, consists of finding the best possible sample. In the following subsections, we discuss how both reductions were performed.

4.2.1. Vertical reduction

To perform vertical reduction, multivariate analysis is required. There exist many methods to reduce the original number of variables. Here, we simply used Pearson's correlation coefficients. As stated, of primordial importance was an efficient treatment of the data. Therefore, we ruled out methods which involved the analysis of large databases and/or the processing of large numbers of attributes. Therefore, an exploration for correlated variables was performed over the original data, which is a relatively inexpensive task. We calculated a correlation matrix for the 415 variables and considered (after consulting with the company's experts) that those variables exhibiting a correlation factor equal or larger than 0.75 were redundant. Hence, from the original 415 variables only 129 remained as informationally interesting. In principle, out of a set of correlated variables only one is needed for clustering purposes. Which of these is to be retained is irrelevant; in fact, we wrote a program which simply performed a sequential binary search to select the (uncorrelated) variables to be retained.

4.2.2. Horizontal reduction

This step is based on the hypothesis that a sample will adequately represent the full set of data. The size of the sample was determined at the offset by the Company's experts; hence, 20% of the original data (after vertical reduction) was sampled. The elements of such sample were randomly (uniformly) selected. From the sample, we validated the representation adequacy of this subset. A central issue to our work was the way the sample is validated. The process consists of the following steps:

1. Select several n equally sized samples. In our case, $n = 5$.
2. Select sets of m variables to perform a goodness-of-fit test. We selected couples ($m = 2$) of variables (v_i, v_j) to prove that, within each sample, the behavior of the selected variables is statistically equivalent.
3. Perform a search for the best regressive function such that $v_i = f(v_j)$. To this effect, we programmatically analyzed, in every case, 34 models (listed in Table 2). From these, we selected the one which displayed the highest Pearson correlation factor.
4. Perform steps 2 and 3 as long as there are more variables to evaluate.

It is important to discuss what we mean by "as long as there are more variables to evaluate". In essence what determines the number of couples to evaluate (n) is the probability that the form of the best regressive function in (3) is the same for all n samples. We are assuming, as shown in the forthcoming graphs that, in effect, the best fit would correspond to like models. In order to establish a bound on n, we start by making the following definitions:

- *Model code*: Each of the models evaluated is encoded with an integer in the range 1–34.

Table 2
Evaluated regressive models

Family	Model	Equation
	Linear	$y = a+bx$
	Quadratic	$y = a+bx+cx^2$
	nth order polynomial	$y = a+bx+cx^2+dx^3+\ldots$
Exponential family	Exponential	$y = a\,e^{bx}$
	Modified exponential	$y = a\,e^{b/x}$
	Logarithm	$y = a+b\ln x$
	Reciprocal log	$y = 1/(a+b\ln x)$
	Vapor pressure model	$y = e^{a+b/x+c\,\ln x}$
Power law family	Power	$y = ax^b$
	Modified power	$y = ab^x$
	Shifted power	$y = a(x-b)^c$
	Geometric	$y = ax^{bx}$
	Modified geometric	$y = ax^{b/x}$
	Root	$y = ab^{1/x}$
	Hoerl model	$y = ab^x x^c$
	Modified Hoerl model	$y = ab^{1/x}x^c$
Yield-density models	Reciprocal	$y = 1/(ax+b)$
	Reciprocal quadratic	$y = 1/(a+bx+cx^2)$
	Bleasdale model	$y = (a+bx)^{-1/c}$
	Harris model	$y = 1/(a+bx^c)$
Growth models	Saturation-growth rate	$y = ax/(b+x)$
	Exponential association 2	$y = a(1-e^{-bx})$
	Exponential association 3	$y = a(b-e^{-cx})$
Sigmoidal models	Gompertz relation	$y = a\,e^{-e^{b-cx}}$
	Logistic model	$y = a/(1+b\,e^{-cx})$
	Richards model	$y = a/(1+e^{b-cx})^{1/d}$
	MMF model	$y = ab+cx^d/(b+x^d)$
	Weibul model	$y = a-b\,e^{-cx^d}$
Miscellaneous	Hiperbolic	$y = a+b/x$
	Sinusoidal	$y = a+b\cos(cx+d)$
	Heat capacity	$y = a+bx+c/x^2$
	Gaussian model	$y = a\,e^{-(x-b)^2/(2c^2)}$
	Rational function	$y = a+bx/(1+cx+dx^2)$

- *Fit*: It is the set of n model codes obtained from the regressive analysis of the couples of the n samples.
- *Maximality*: The largest number of similar models in a fit. For instance, the fit (2, 5, 6, 5, 5) has a maximality of 3.
- *Similar fits*: Fits with the same maximality.

The confidence level is expressed by (1):

$$c = \prod_{i=1}^{k} P_i, \tag{1}$$

where c is the confidence level, k is the number of couples to analyze, and P_i is the probability of fit i. We consider this probability independent from that of other fits.

We define the probability of a fit as the probability of similar fits. This value may be approximated as

$$P_i = \frac{\#\text{ Similar fits}}{\#\text{ Possible fits}}. \tag{2}$$

A small value of P_i implies that the fit has low probability and is, therefore, unlikely to have been resulted from chance alone. The number of similar fits and possible fits results directly from the theory of combinations with repetition[2] (CR).

4.2.3. Similar fits

To determine the number of similar fits of maximality t, we may use the next formula:

$$\#\text{ Similar fits} = \begin{cases} d & \text{if } t = n, \\ \binom{d}{n} & \text{if } t = 1, \\ d\left[\sum_{i=\lceil h/t \rceil}^{h} v\binom{d-1}{i}\right] & \text{otherwise,} \end{cases} \tag{3}$$

where d is the number of models to evaluate; h is the $n-t$; and v is the amount of combinations of i numbers less than or equal t which add up to h.

4.2.4. Possible fits

The total number of fits is given by the number of combinations with repetition for d models in n samples. Then

$$\#\text{ Possible fits} = \mathrm{CR}_d^n = \binom{d+n-1}{n} = \frac{(d+n-1)!}{(d-1)!n!}, \tag{4}$$

where d is the number of models to test and n is the samples to analyze.

Thus, given a desired confidence level, we may solve for n. For example, if $n = 5$ and $d = 34$ we see that $P_i = 34/501942 = 6.77369 \times 10^{-6}$.

Figs. 1–3 illustrate the fact that several functions resulting from paired variables yield similar regressive fits. The data displayed in graphs 1a and 1b are closely adjusted with an MMF model; those of graphs 2a and 2b are, analogously, adjusted by a 4th degree polynomial; finally, the data displayed in graphs 3a and 3b are tightly fit by a rational function. Interestingly, the correlation coefficient in all three couples is better than 0.93 indicating the very high quality of the fit. Hence, we rest assured that all samples display statistically significant equivalence. (We note that, because of space limitations, we are unable to show the entire set; however, very similar remarks do apply in all cases.) On the other hand, for different couples we obtain best fit with *different* models: MMF $[(ab+cx^d)/(b+x^d)]$ for couple 1; 4th degree polynomial $(a+bx+cx^2+dx^3+ex^4)$ for couple 2 and a rational function $[(a+bx)/(1+cx+dx^2)]$ for couple 3. This fact reinforces our expectation that different variables distribute differently even though the samples behave equivalently. A hypothetical possibility which is ruled out from this behavior is that all variables were similarly distributed. If this were the case, then ALL models would behave similarly and no significant conclusion could be derived from our observations.

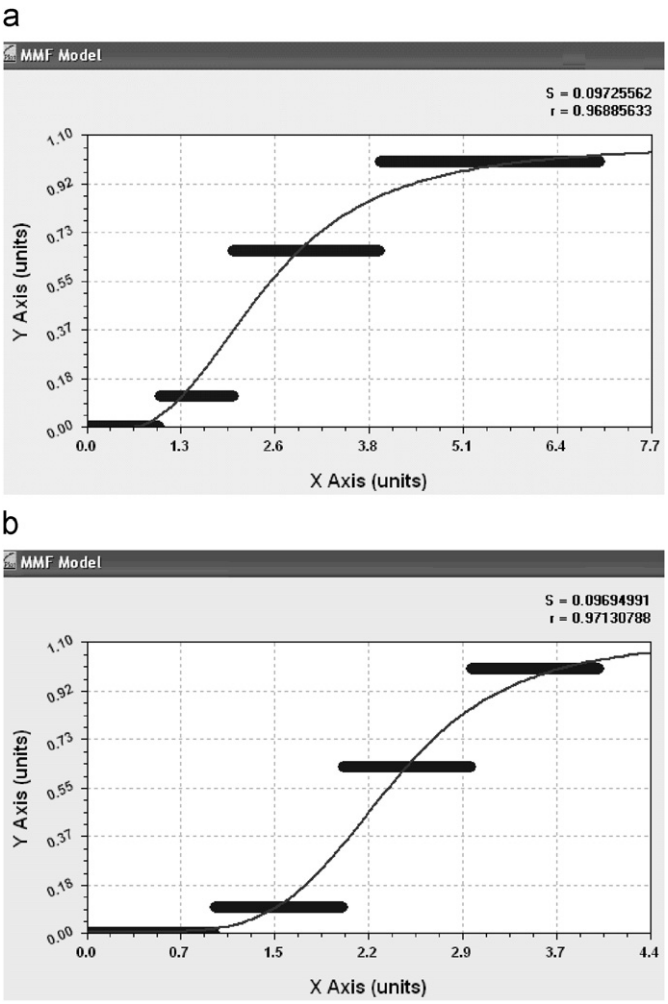

Fig. 1. Regressive fits: MMF model for (a) sample 1 and (b) sample 2.

It may be argued, upon first analysis, that the high correlation coefficients contradict the fact that our variables derive from the elimination of such correlation. Notice, however, that even if the variables with which we worked are not correlated (as discussed above) this non-correlation is *linear* (as pertaining to a Pearson coefficient) whereas the models considered here are basically highly non-linear, which resolves the apparent contradiction.

The probability of displaying results as shown by chance alone is $O(10^{-6})$. We must stress the fact that this analysis is only possible because we were able to numerically characterize each of the subsets in 34 different forms and, thus, to select the most appropriate ones. Furthermore, not only characterization was proven; we also showed that, in every case, the said characterization was similar when required and dissimilar in other cases.

4.3. Clustering phase

Once the search space is reduced the clustering phase is reached. Before attempting the clustering proper, we impose certain a priori assumptions, as follows:

[2] A combination is an array of elements where the order is irrelevant. In a CR, there may be repeated elements.

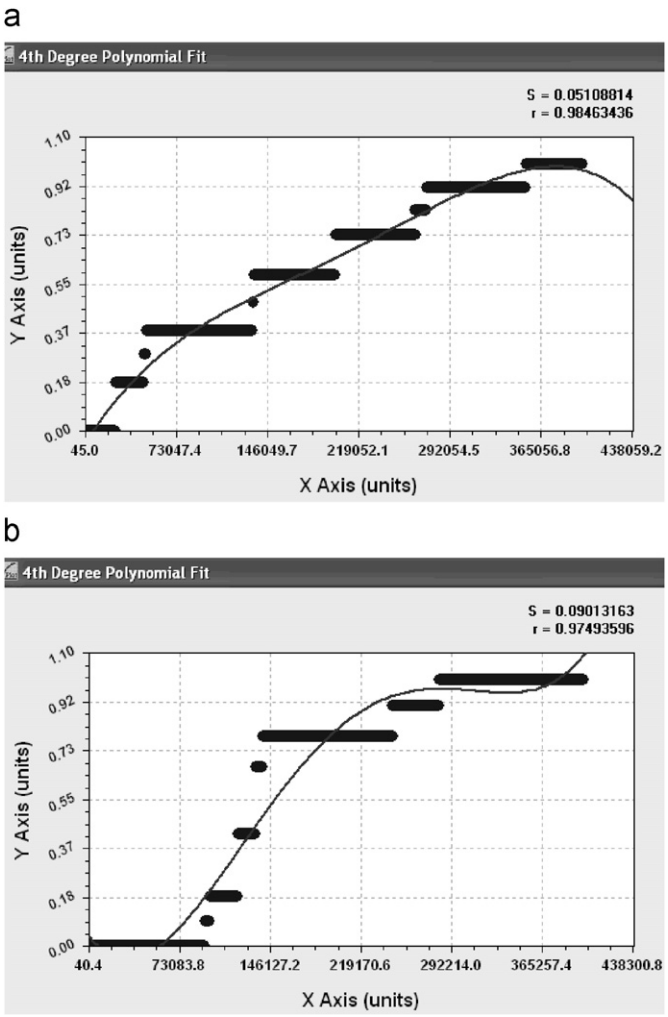

Fig. 2. Regressive fits: 4th degree polynomial model for (a) sample 1 and (b) sample 2.

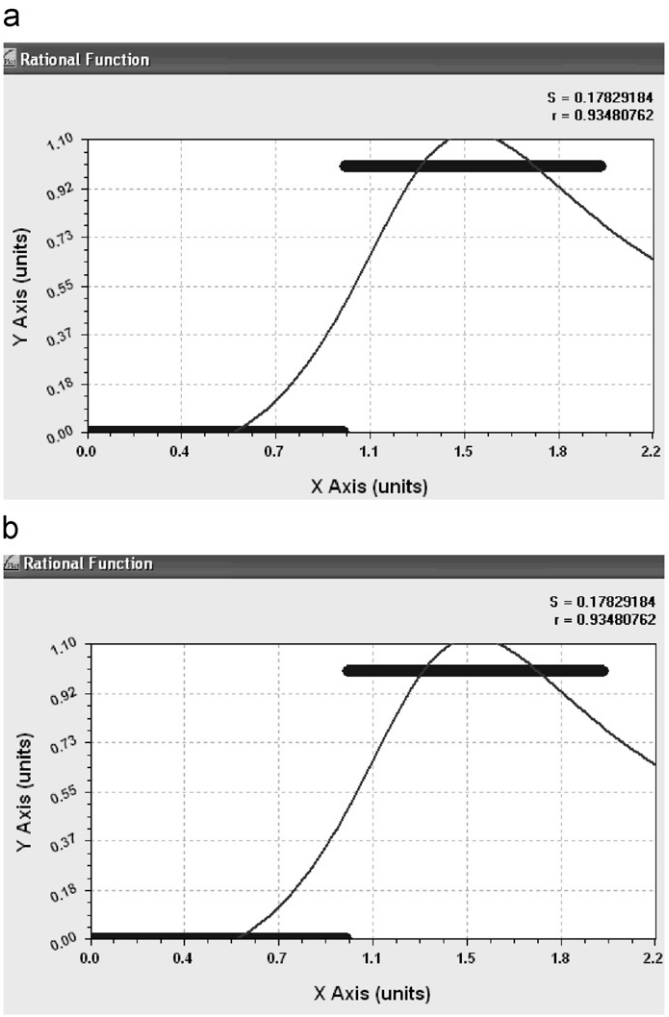

Fig. 3. Regressive fits: rational function model for (a) sample 1 and (b) sample 2.

- The number of clusters is to be determined automatically (without applying any aprioristic rules).
- The "best" number (N) of clusters is derived from information theoretical arguments.
- The theoretical N is to be validated empirically from the expert analysis of the characteristics of such clusters.

In order to comply with our assumptions, we follow the next steps:

1. Consecutively obtain the clusters (via a Fuzzy C Means algorithm) assuming n clusters for $n = 2, 3, \ldots, k$, where "k" represents the largest acceptable number of clusters (in our case, $k = 20$).
2. Determine the "optimal" number of clusters according to "elbow" criterion (see Palmer and Faloutsos, 2000).
3. Clustering with a self-organizing map algorithm to find the optimal segmentation.

The minable view with the 129 variables was processed. The Fuzzy C Means (FCM) algorithm was used on the uncorrelated data and the elbow criterion was applied (see Bezdek, 1974). It is important to stress the fact that the use of fuzzy logic allows us to determine the content of information (the entropy) in every one of the N clusters into which the data set is divided. Other clustering algorithms based on crisp logic do not provide such alternative. Since the elements of a fuzzy cluster belong to all clusters, it is possible to establish an analogy between the membership degree of an element in the set and the probability of its appearance. In this sense, the "entropy" is calculated as the expected value of the membership for a given cluster. Therefore, we are able to calculate the partition's entropy PE. Intuitively, as the number of clusters is increased the value of PE increases since the structure within a cluster is disrupted. In the limit, where there is a cluster for every member in the set, PE is maximal. On the other hand, we are always able to calculate the partition coefficient: a measure of how compact a set is. In this case, such measure of compactness decreases with N. The elbow criterion stipulates that the "best" N corresponds to the point where the corresponding *tendencies* of PE to increase and PC to decrease *simultaneously* change. That is, when the curvature of the graph of tendencies changes we are faced with an optimal number of clusters. Table 3 displays part of the numeric data values of PC and PE. These coefficients were calculated with formulas (5) and (6):

$$\mathrm{PC} = \sum_{k=1}^{K}\sum_{i=1}^{c}\frac{(\mu_{ik})^2}{K}, \tag{5}$$

$$\mathrm{PE} = -\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{c}\mu_{ik}\,\ln(\mu_{ik}), \tag{6}$$

where μ_{ik} is the degree of membership of object k to cluster i.

Fig. 4 shows the graph for the numeric data of Table 3. In the graph, the "elbow" point is located between the clusters 6 and 7, indicating that there is a high probability that the optimal number of clusters is in that point, i.e., $N = 6$.

The last phase of our analysis implied the use of the miner and to find the theoretically determined best number of clusters, as shown in Fig. 5.

Table 3
Numeric data for the elbow criterion

	Clusters										
	2	3	4	5	6	7	8	9	10	11	12
PC	0.879	0.770	0.642	0.560	0.498	0.489	0.413	0.414	0.400	0.359	0.349
PE	0.204	0.436	0.639	0.812	0.982	1.036	1.220	1.224	1.272	1.403	1.433

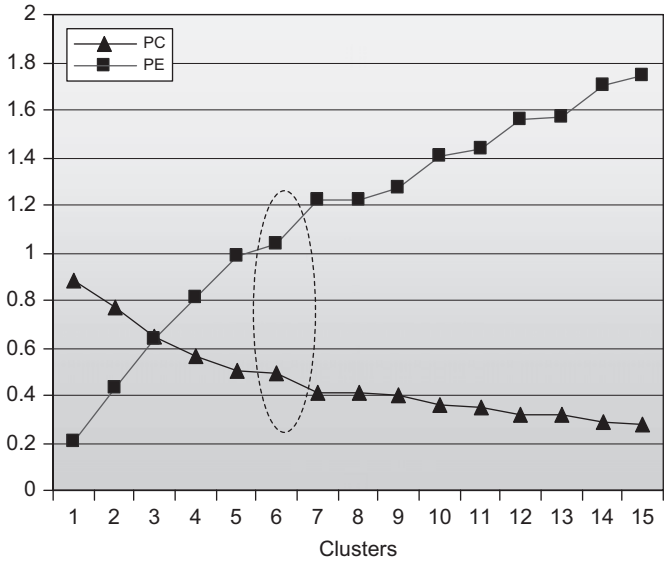

Fig. 4. Graph for the elbow criterion.

The graph shows, on the left, the percentage of elements grouped in each cluster. On the right, the neuron number which represents the cluster. Each cluster shows the more important variables for the results, ordered by Chi-squared characterization of the variable's behavior in the cluster and in the whole sample. The cluster information for the Company can be extracted from the graph and reports supplied by the tool. We should now prove that clustering resulting from the reduced search space reflects a correct clustering view of the population.

4.4. Validation of the reduced search space

To ease the understanding of the process in what follows, we will call the clustering model from the sample "Model 1"; likewise, we will call the clustering model derived from the complete data "Model 2".

We followed the next steps:

1. Reduce the original data set only vertically.
2. Execute a clustering process over the full set of data to obtain Model 2.
3. Label all the original data set and the sample data set with Models 1 and 2.
4. Compare the resulting distribution of elements labeled with both models.

The results are discussed in what as follows.

4.4.1. Comparison of Models 1 and 2

Table 4 shows the percentages for the two models. The names of the clusters were replaced by letters to avoid possible confusions with the neuron numbers shown in the miner's results. As Table 4 shows, the result clusters are very similar.

4.4.2. Clustering from sampling (Model 1)

Having Model 1, we labeled the sample data and the full data sets. The resulting distribution of elements into the different six clusters was expressed in percentages for comparison effects. As Table 5 shows, the resulting distribution for the sample and for the full data are almost equal. This proves that the sample represents the full data set adequately.

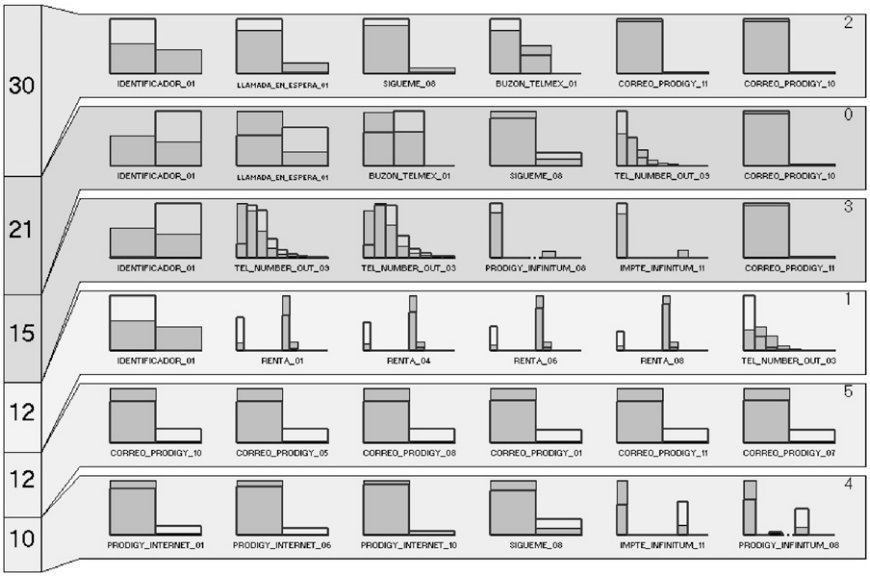

Fig. 5. Graph view of the clustering result supplied by the miner.

4.4.3. Cross-validation

Finally, we labeled the sample and the entire data with both algorithms. The results are shown in Table 6.

As Table 6 shows, the differences between the distributions of elements into the clusters are similar between the two clustering models. Analog clusters share the same cardinality with a difference of less than 3%.

Table 4
Clusters' comparison for Models 1 and 2

Clusters	Model 1 (%)	Model 2 (%)	Difference (%)
A	30	27	3
B	21	20	1
C	15	18	3
D	12	15	3
E	12	12	0
F	10	8	2

Table 5
Labeling from the Model 1 applied to the sampled and full data sets

Cluster	Sample (%)	Full data (%)	Difference (%)
A	30.06	30.24	0.18
B	21.01	20.91	0.10
C	15.45	15.37	0.08
D	12.27	12.25	0.02
E	11.54	11.55	0.01
F	9.67	9.68	0.01

4.5. Cost

The method allows, therefore, to replace a large sample with a smaller one. But this would be of no use if we cannot determine if the method is cost effective. To estimate such cost, we have to know:

- The cost of the clustering algorithm in terms of data access.
- The sample's size.
- The maximum number of probes to be performed over the samples.
- The number of models to be evaluated in the regression process.
- The number of variables for the regression process.

The methodology's cost is determined by equation

$$\text{Reduced cost} = y \times n \times q \times a + z \times \text{original cost}, \quad (7)$$

where y is the number of variables for the regression process, n is the total number of rows for the problem, q is the number of models to be evaluated in the regression process, a is the number of probes over the samples, z is the sample's size, and original cost is the clustering algorithm's cost without apply horizontal reduction. It is equivalent to the number of accesses that is needed by the algorithm for each data to be processed (rows × variables of the problem).

The first part on the right side of Eq. (7) corresponds to the validation cost; the second part to the clustering cost for the sample.

Fig. 6 shows how the cost reduction rate is affected by the number of probes performed over the sample.

Table 6
Comparison of full and sampled data clusters

Cluster	Labeling derived from sampled data				Labeling derived from complete data			
	Model 1	Model 2	Difference	% Population	Model 1	Model 2	Difference	% Population
A	23906	21503	2403	3	120961	108084	12,877	3
B	16712	16155	557	1	83655	81420	2235	1
C	12285	14327	2042	3	61471	72367	10,896	3
D	9760	11828	2068	3	49013	59313	10,300	3
E	9179	9580	401	1	46195	48356	2161	1
F	7687	6136	1551	2	38705	30460	8245	2

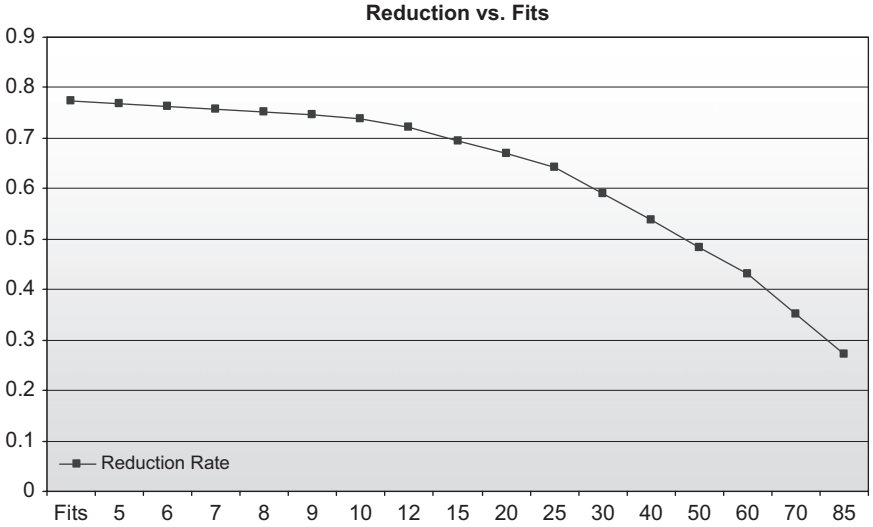

Fig. 6. Reduction rate vs. number of probes on the sample.

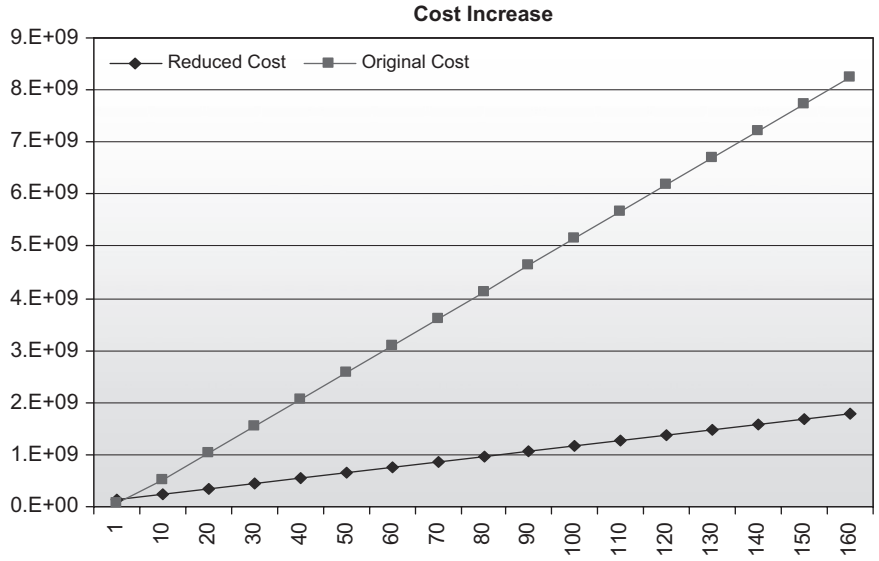

Fig. 7. Reduced cost vs. original cost.

Fig. 7 shows how, when the clustering algorithm's complexity increases, the difference between the original cost and the methodology cost increases. The graph was calculated for five samples, five probes on sample (size = 20%) and 34 models.

5. Conclusions

As we pointed out in the introduction, data mining may be an important strategic tool for commercial enterprises. But the management of large volumes of data (both physically and logically) may become a practical problem of large proportions and difficult to solve. Applying the methodology advanced herein it is possible to drastically reduce the size of the database to be processed. In this case, we were able to reduce the size in close to 93.78%. Originally we had to deal with 166 million elements (i.e., 400,000 registers with 415 attributes each); instead we used a sample with only 10.32 million such elements (80,000 records with 129 attributes). The reduced sample, however, performed in a way that made it statistically indistinguishable from the original data. Apart from the benefit resulting from having quicker access to strategic information the use of this methodology yields economic benefits derived from the ability to process a smaller sample (increased speed and capacity for data processing; decreased amount of primary and secondary storage, costs of software and hardware, among others). Considering that the company had important improvements with the application of the results of this investigation, we consider that continuing research is needed and justified, since much work remains to be done if we wish to set a bound on the characteristics of the data which will allow us to generalize the results reported here.

References

Ahlemeyer-Stubbe, A., 2001. Analyseorientierte Informationssysteme = Datawarehouse. In: Perner, P. (Ed.), Data Mining, Data Warehouse, Knowledge Management, Proceedings of the Industrial Conference on Data Mining ICDM 2001, IBal Report 2001, pp. 1–30.

Berkhin, P., 2002. Survey of Clustering Data Mining Techniques. Accrue Software Inc.

Bezdek, J.C., 1974. Cluster validity with fuzzy sets. Journal of Cybernetics 3, 58–72.

Brighton, H., Mellish, C., 2002. Advances in instance selection for instance-based learning algorithms. Data Mining and Knowledge Discovery 6, 153–172.

Cheng, D., Kannan, R., Vempala, S., Wang, G., 2006. A divide-and-merge methodology for clustering. ACM Transactions on Database Systems (TODS) 31 (4), 1499–1525.

Delmater, R., Hancock, M., 2001. Data Mining Explained: A Manager's Guide to Customer-Centric Business Intelligence. Digital Press (Chapter 6).

Fodor, I.K., 2002. A Survey of Dimension Reduction Techniques. US Department of Energy, Lawrence Livermore National Laboratory.

Guha, S., Rastogi, R., Shim, K., 1998. CURE: an efficient clustering algorithm for large databases. In: ACM Proceedings: International Conference on Management of Data, Washington, USA, pp. 73–84.

Hair, J.F., Anderson, R.E., Tatham R. L., BlackW.C., 1999. Análisis Multivariante, fifth ed. Pearson, Prentice-Hall, Madrid (Chapter 4).

Jagadish, H.V., Lakshmanan, L.V., Srivastava, D., 1999. Snakes and sandwiches: optimal clustering strategies for a data warehouse. In: ACM Proceedings: International Conference on Management of Data, Philadelphia, USA, pp. 37–48.

Jain, K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. ACM Computing Surveys 31 (3), 264–323.

Kleinberg, J., Papadimitriou, C., Raghavan, P., 2004. Segmentation problems. Journal of the ACM (JACM) 51 (2), 263–280.

Kuri, A., Rodriguez, F., 2007. A search space reduction methodology for large databases: a case study. In: Perner, P. (Ed.), Advances in Data Mining—Theoretical Aspects and Applications, lncs 4597. Springer, Heidelberg, p. 2007.

Liu, H., Motoda, H., 2002. On issues of instance selection. Data Mining and Knowledge Discovery 6 (2), 115–130.

Palmer, C. R., Faloutsos, C., 2000. Density biased sampling: an improved method for data mining and clustering. In: ACM Proceedings: International Conference on Management of Data, Texas, USA, pp. 82–92.

Palpanas, T., 2000. Knowledge discovery in data warehouses. ACM SIGMOD Record 29 (3), 88–100.

Perner, P., 2002. Data Mining on Multimedia Data. Springer.

Peter, W., Chiochetti, J., Giardina, C., 2003. New unsupervised clustering algorithm for large datasets. In: ACM Proceedings: International Conference on Knowledge Discovery and Data Mining, Washington, USA, pp. 643–648.

Raymong, T. N., Jiawei, H., 1994. Efficient and effective clustering methods for spatial data mining. In: ACM Proceedings: International Conference on Very Large Data Bases, San Francisco, USA, pp. 144–155.

Vu, K., Hua, K. A., Cheng, H., Lang, S., 2006. A non-linear dimensionality-reduction technique for fast similarity search in large databases. In: ACM Proceedings: International Conference of Management of Data, Chicago, USA, pp. 527–538.

Zhang, D., Zhou, Z., Chen, S., 2007. Semi-supervised dimensionality reduction. In: SIAM Proceedings: International Conference on Data Mining.

Zhu, X., Wu, X., 2006. Scalable representative instance selection and ranking. In: IEEE Proceedings: International Conference on Pattern Recognition, pp. 352–355.

Angel Kuri-Morales was born in Mexico City. He is author of two text books and more than 60 articles published in international magazines and conferences; member of the National System of Researchers (SNI). He won an international prize during the International Congress on Evolutionary Computation in 2000. He has been included in "Who is Who in the World" in 1988, 1998, 2000, 2002 and 2003. He has been president of several International Congresses, and invited

speaker in many national and international scientific events. He is a Distinguished Lecturer of the Association for Computing Machinery (ACM) and member of the Scientific Committee of the World Scientific and Engineering Academy and Society (WSEAS). Currently, he is the member of the Board of IBERAMIA, President of the Mexican Society for Artificial Intelligence and Professor in the Autonomous Technological Institute of Mexico (ITAM).

Fátima Rodríguez-Erazo was born in El Salvador. She graduated with honors as Informatics Systems Engineer from the University of El Salvador (UES) in 2003. In 2002, she received the recognition from the Salvadorian Association of Engineers and Architects as one of the best students of Engineering from El Salvador. She studied a master's degree in Computer Science and Engineering in the National Autonomous University of Mexico (UNAM) and graduated with honors in 2007.