Data Base Analysis using a Compact Data Set

Angel Ferrnando Kuri-Morales
Departamento de Computación
Instituto Tecnológico Autónomo de México
Mexico City, Mexico
akuri@itam.mx

***Abstract* — The exploitation of large data bases frequently implies the investment of large and, usually, expensive resources both in terms of the storage and processing time required. It is possible to obtain equivalent reduced data sets where the statistical information of the original data may be preserved while dispensing with redundant constituents. Therefore, the physical embodiment of the relevant features of the data base is more economical. We propose a method where we may obtain an optimal transformed representation of the original data which is, in general, considerably more compact than the original without impairing its informational content.**

Keywords - data bases, statistics, compaction

I. INTRODUCTION

Nowadays, commercial enterprises are importantly oriented to continuously improving customer-business (CRM) relationship. With the increasing influence of CRM 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 very large data bases 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 [1, 2]. To attempt direct clustering over a data base 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 one may use parallel computation, optimization of clustering algorithms, 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 work 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.

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

A. Paper Outline

The outline of the paper is as follows. First, we give an overview of the analysis of large databases in section 2; next we give an overview of the methodology we applied. In section 3 we briefly discuss the case study treated with the proposed methodology. Finally, we conclude in Section 4.

II. 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 nontrivial information 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).

A. Clustering in Large Databases

Clustering is a popular data mining task which consist 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 high importance as a data mining task, clustering has been the subject of multiple research efforts and has proven to be useful for many purposes [3].

Many techniques and algorithms for clustering have been developed, improved and applied [4], [5], [6]. Some of them try to ease the process on a large database as in [7], [8] and [9]. On the other hand, the so-called "Divide and Merge" [10] or "Snakes and Sandwiches" [11] 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 [12].

978-1-4799-5057-7/14 $31.00 © 2014 IEEE
DOI 10.1109/BigData.Congress.2014.41

B. 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 nonprobabilistic. 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 nonprobabilistic ones include methods such as convenience sampling, judgment sampling and quota sampling. There are many ways to select the elements from a data set and some of them are discussed in [13]. This field of research, however, continues to be an open one [14], [15].

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 [12]. However, small clusters are not always significant; such is the case of costumer 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 [16], [17], and these had resulted in a large number of methods and algorithms [18]. 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. [19]. Here we focus on the explicit determination of a functional which maximizes the resulting correlation coefficient while minimizing its standard error. This approach requires a general and efficient tool for model generation, as will be discussed in the sequel.

III. CASE STUDY

A data mining project was conducted for a very large multi-national company (one of the largest in the world) 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 several data bases. 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 one year period.

To test the working methodology the project team worked with a set of 12,000,000 customer's registers, consisting of a total of 118 variables per register.

A. Methodology

In order to ensure the reliability of the results, the following steps were taken:

i) Data analysis
ii) Programming language selection
iii) Categorical variable encoding
iv) Calculation of smallest equivalent sample
v) Seasonality analysis

In what follows we brieflys describe every step.

i) Data Analysis. In this step databases (DB) are filtered in order to eliminate deficiencies and/or limitations which may render them inconvenient for later analysis .
The DB is searched to detect:

[1] Possible design and/or input errors.
[2] Erroneous data types
[3] Inadequate numerical type
[4] Excessive numerical precision
[5] Insufficient numerical precision
[6] Variables with only one value
[7] Variables with too many values
[8] Outliers

ii) Programming language selection. Originally we planned to use a general purpose algorithmic language. However, because the Company made extensive use of SAS® (Statistical Analysis System) which is an integrated software product which allows the efficient manipulation of DBs and includes a set of auxiliary utilities, it was selected as the main development tool. It does, however, have certain limitations as a programming tool which were circumvented by writing a set of special routines in a lower level language. The original DBs were originally stored in an Oracle® environment which were then migrated to SAS.
Right from the start the design of two basic utilities was considered: a) Calculation of the entropy in a subset of the original (called "U") DB to determine the minimum number of objects which preserved the information in U (called "M"). Testing of a large enough set of experimental

probability distributions to determine the distribution equivalence between the pairs of variables in U and those in M.

iii) Categorical variable encoding. In order to apply the clustering algorithms (to be discussed in the sequel) we must be aware of the fact that these are restricted to working with numerical fields. In our DB, however, some of the fields were categorical (i.e. not representable by numbers). In order not to waste the potentially valuable information residing in such variables we wrote a program which: a)Identifies categorical variables, b) Determines the number of different possible values the variable make take (i.e. ethnia: White, Black, Asian, Indian, Caribbean, American, Polynesian). It is very important NOT to assign arbitrary numerical values (i.e. 1, 2, …, 7) since this practice may very surely induce patterns which are non-existent in the original data. Instead, we generate (we continue our example) 7 binary pseudo-variables; one per each instance of the category. Hence, the variable "Ethnic group" will be replaced by 7 synthetic (pseudo) variables as exemplified (the headings stand for White, Black, etc.) in Table I.

TABLE I. BINARY PSEUDO-VARIABLES

Ethnic Group		W	B	As	I	C	Am	P
Black		0	1	0	0	0	0	0
American		0	0	0	0	0	1	0
Asian		0	0	1	0	0	0	0
. . .		. . .	. . .	. . .	. . .	. . .	. . .	. . .

The pseudo-variables do not exactly represent the information in the variable they replace, since now we will have, for example, a variable representing those members of the (say) Assian ethnia, another for those representing the Caribbean ethnia and so on. But we do not have the variable "ethnic group". Nonetheless, this course of action allows us not to introduce in the DB spurious relations between the variables and apply the clustering algorithms without lossing potentially interesting information.

iv) Calculation of smallest equivalent sample. To reduce the search space we work 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 horizontally (reducing the number of tuples) and vertically (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. The following subsections discuss how we performed both reductions.

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. An exploration for correlated variables was performed over the original data. We calculated a correlation matrix for the 118 variables. We considered (after consulting with the experts) that those variables exhibiting a correlation factor equal or larger than 0.85 were redundant. Hence, from the original 118 variables only 73 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.

Horizontal Reduction. The data set is analyzed as a set of sequences of symbols composing messages in accordance with the statistical theory of communication. Each set of attributes represents a message and each attribute value corresponds to a symbol. In this light is possible to estimate the entropy contained on the message. The entropy is a measure of the average information in the symbols of a message. It is a tool which we used to ensure that the information of the sample and that of the population are alike. The entropy of the message can be approximated as the proportion of appearances of each symbol as follows:

$$H(X) = \sum_{i=1}^{m} - p_i \; log(p_i)$$

$$H(X) \approx \sum_{i=1}^{m}\left(\sum_{j=1}^{n} \frac{\delta(S_i, v_i)}{n} \right) log\left(\sum_{i=1}^{m} \frac{\delta(S_i, v_i)}{n} \right) \quad (1)$$

Where X is the message, p_i is the probability of occurrence of symbol i. m is the number of symbols, n is the number of data elements, S_i is the *i-th* symbol value and v_j is the *j-th* data value and

$$\delta(s, v) = \begin{cases} 0 & if \quad v \neq s \\ 1 & if \quad v = s \end{cases} \quad (2)$$

Our aim is to approximate the value of the population's entropy by the entropy of a properly selected sample to avoid accessing the full DB. The method consists of treating every attribute t to obtain a sample M_t, as follows. Initially M_t is empty. Then we proceed to extract randomly (uniform) selected elements of the population for attribute t iteratively and adding these elements to M_t. On each iteration i, the entropy is calculated and compared with the one of the previous iteration as follows

$$\Delta H(i)=H(i)-H(i-1) \quad (3)$$

As $\Delta H(i)$ becomes closer to a threshold parameter ε the entropy of M_t is asymptotically closer to the population's entropy as illustrated in Fig.1. At this point the size $|M_t|$ has

been determined. The value of ε is set by the user (we set *ε=0.0001*).

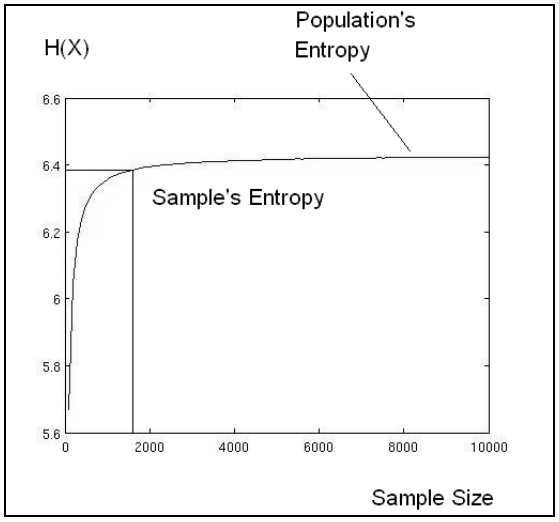

Figure 1. Entropy dynamics on different sample sizes.

This process is performed on every attribute. Once the M_i's have been calculated then the overall sample size M is the largest one, thus ensuring the proper representativeness for all attributes.

Sample validation. To further attest to the validity of any possible sample of size |M|, we must ensure that their entropies have the same or larger value than the one calculated before. Making use of the information available, a simple non parametric Monte Carlo test may be applied [20] to validate the entropy preservation on each variable. This test should confirm the null hypothesis *H0: H(X)* $\leqslant$ *Hc(X)* where *H(X)* is the value of the maximum entropy obtained from the sample size calculation process and *Hc(X)* is the calculated entropy of any sample of size |M|. The test consists of the generation of a set of samples of size |M|. If we set a confidence level of 95%, then at least 95% of the samples must be consistent with the null hypothesis, proving that the samples do comply. If the test is not passed then the sample size should be increased. Because of computational cost, we only perform tests between couples of attributes. In strict sense, to validate the preservation of the all patterns and relations, multivariate tests of high order should explored. Nevertheless, a mathematical model may be calculated to estimate the value of an arbitrary attribute as a function of another, also arbitrarily selected. We take advantage of SAS's ease in fitting data to mathematical models. We defined a set of 36 test models (a partial list is shown in Table II). Every model can be evaluated on data from different samples. We calculate the approximation L_2 error for polynomial *P*, where *P(Xi)=Yi+e*. Where *Xi*, *Yi* are the values of the i-th value of the attributes and *e* is the L_2 approximation error on that data value. This error is defined as: $e = \sum \left(P(X_i) - Y_i\right)^2$. If the sample preserves all the couples' relations then the approximation error should be close on every sample. We calculate the ratio $r = e_{max}/e_{min}$. *r* must be close to 1 if the approximation errors are similar. A sample could be rejected if $r > 1+\gamma$ (where γ's value is determined by the user). This analysis should be applied to every couple to determine whether to accept or increase |M|. Once M has been validated we are statistically certain that a clustering algorithm operated on it will yield similar results as if it were applied to the original database.

TABLE II. EVALUATED REGRESSION MODELS

Model	Equation
Linear	$y = a + bx$
Quadratic	$y = a + bx + cx^2$
nth Order Polynomial	$y = a + bx + cx^2 + dx^3 + \ldots$
Exponential	$y = ae^{bx}$
Modified Exponential	$y = ae^{b/x}$
Logarithm	$y = a + b\ln x$
Reciprocal Log	$y = \frac{1}{a + b\ln x}$
Vapor Pressure Model	$y = e^{a+b/x+c\ln x}$
Power	$y = ax^b$
Modified Power	$y = ab^x$
Shifted Power	$y = a(x-b)^c$
Geometric	$y = ax^{bx}$
Root	$y = ab^{1/x}$
Hoerl Model	$y = ab^x x^c$
Modified Hoerl Model	$y = ab^{1/x} x^c$
Hiperbolic	$y = a + b/x$
Heat Capacity	$y = a + bx + c/x^2$
Gaussian Model	$y = \frac{a + bx}{1 + cx + dx^2}$

Several functions resulting from paired variables did yield similar regressive fits. We note that, because of space limitations, we are unable to discuss the entire set; however, very similar remarks do apply in all cases. Interestingly, the self-regressive correlation coefficient in all cases is better than 0.93 indicating the very high quality of the fit. Hence, we rest assured that all samples display statistically significant equivalence. For different couples we obtain best fit with *different* models. For example, $[(ab+cx^d)/(b+x^d)]$ *(MMF)* for couple 1; $(a+bx+cx^2+dx^3+ex^4)$ (4^{th} degree polynomial) for couple 2 and $[(a+bx)/(1+cx+dx^2)]$

(rational function) 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.

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 "non-linear", which resolves the apparent contradiction.

The probability of displaying results as discussed by chance alone is less than 10^{-12}. We must stress the fact that this analysis is only possible because we were able to numerically characterize each of the subsets in 36 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.

v) Seasonality. The last issue we had to consider had to do with seasonal tendencies of time-dependent variables. They add no information to the process but may induce a certain amount of numerical instability to the models. We Illustrate the fact in Fig. 2 and Fig. 3. Data before and after removing seasonal tendencies is shown. We calculated a moving average of 12 weeks, as illustrated in Fig. 2.

The moving average values were directly subtracted from the original data. The graph corresponding to the resulting trend-free data is shown in Fig. 3.

B. Clustering Phase

Once the search space is reduced the clustering phase is reached. We want the number of clusters to be determined automatically (without applying any aprioristic rules). Hence, the "best" number (*N*) of clusters is derived from

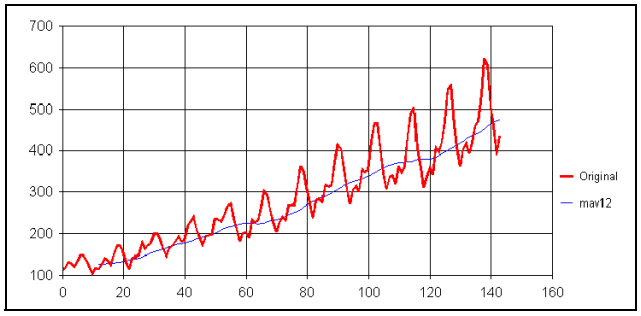

Figure 2. Original data displaying seasonal tendencies.

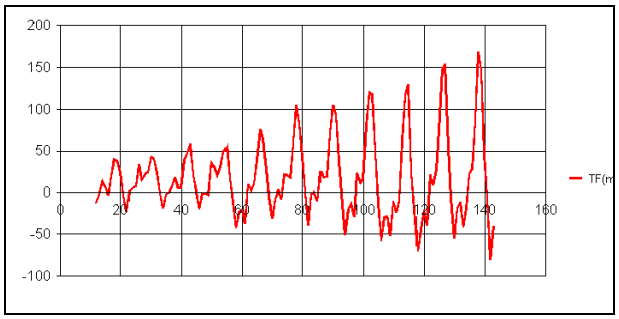

Figure 3. Data with tendencies removed.

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: a) Consecutively obtain the clusters (via a Fuzzy C Means (FCM) algorithm) assuming *n* clusters for *n=2, 3, ..., k*; where "*k*" represents the largest acceptable number of clusters. Determine the "optimal" number of clusters according to "elbow" criterion [21, 22]. b) Find the clusters with a self organizing map to find the optimal segmentation. The reduced minable view was processed. FCM was used on the processed data and the elbow criterion was applied 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 (see below). 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 III displays part of the numeric data values of PC and PE. These coefficients were calculated with formulas 4 and 5.

$$PC = \sum_{k=1}^{K} \sum_{i=1}^{c} \frac{(\mu_{ik})^2}{K} \quad (4)$$

$$PE = -\frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{c} \mu_{ik} \ln(\mu_{ik}) \quad (5)$$

TABLE III. NUMERICAL DATA FOR ELBOW CRITERION

Clusters	2	3	4	5	6	7	8
Partition	0.879	0.770	0.642	0.560	0.498	0.489	0.413
Entropy	0.204	0.436	0.639	0.812	0.982	1.036	1.220

In Fig. 4 the elbow behavior is illustrated.

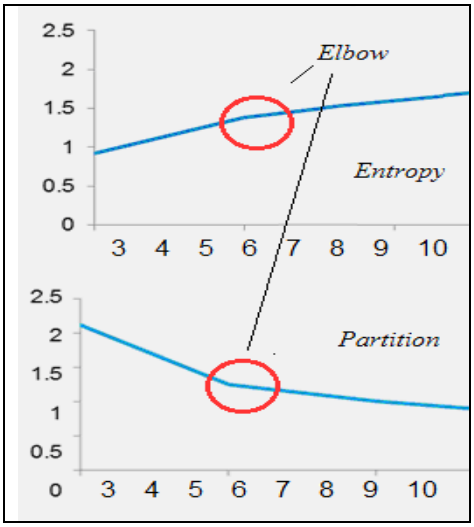

Figure 4. Elbow criterion for Partition and Entropy Coefficients.

From it we infer that the number of clusters associated with de DB is 6.

C. Data Analysis

Once the number of clusters has been decided, we used Kohonen's Self-Organizing Maps (SOM) to find the definitive centers of the clusters. Hence, we defined a SOM consisting of 6 neurons whose coordinates in the problem's space correspond to the centers of each of the 6 clusters.
The last phase of our analysis implied the use of SAS miner and the theoretically determined best number of clusters, as shown in Fig. 5. The pie chart shows the percentage of elements grouped in each cluster. 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.

Validation. The clustering effort reported here was undertaken simultaneously by independent groups in at least

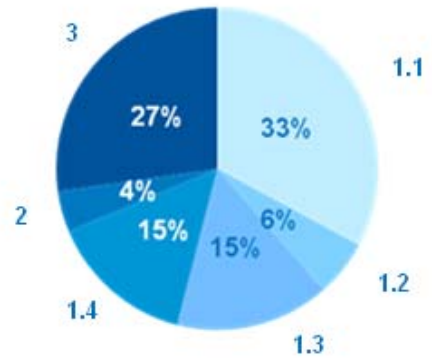

Figure 5. Pie Chart of the clustering result .

five different countries where the Company has ongoing operations. The different proposed methodologies were subject to worldwide practical tests for over eight months and, when the results were assessed, the conclusions derived from M were the more accurate and economical of all the alternatives. As a result, the strategy reported herein was adopted as a standard for different regions of the Company's operations.

Intuitive Interpretation of the Results. One of the most difficult tasks when performing a kind of statistical clustering as the one reported here, is to achieve the ease of understanding so that the clusters convey some intuitive sense to the end user. In this regard, we conducted a principal component's analysis and determined the eigenvectors. From these we extrated a set of simple rules which, although not exactly precise, allow us to establish a more intuitive view for each of the clusters. These are illustrated in Fig. 6.

SEGMENT	INCOME	AVERAGE	MORT.	NOM.	CC	HIPO	TOK
1.1	500-3,500	<1,500	125-300	NO	NO	--	<3,500
	3,500-10,000	<10,000	NO	NO	NO	--	<10,500
1.2	5,000-13,000	<45,000	250-650	6-48	3,800-19,000		
		<4,500	300-600	5-27.5	NO		<17,000
1.3	<500	<1,000	250-650	2-20	NO		<100
1.4	11,000-20,000	<110,00	150-750	2.5-27.5	5,000-22,000	<5,000	<30,000
2	14,000-53,000	<320,000	336-1,000	34-97	15,000-89,000	<20,000	<190,000
	10,000-45,000	<105,000	190-700	10-80	<13,000	<500	<65,000
3	<500	--	125-800	5-50	<25,000	<1,500	

Figure 6. A set of simple rules for the clusters.

Finally, several statistics were obtained to characterize the clusters. As an example, consider the data en Fig. 7. From the socio-demographic life cycle we obtained the corresponding histogram for the customers of the Company.

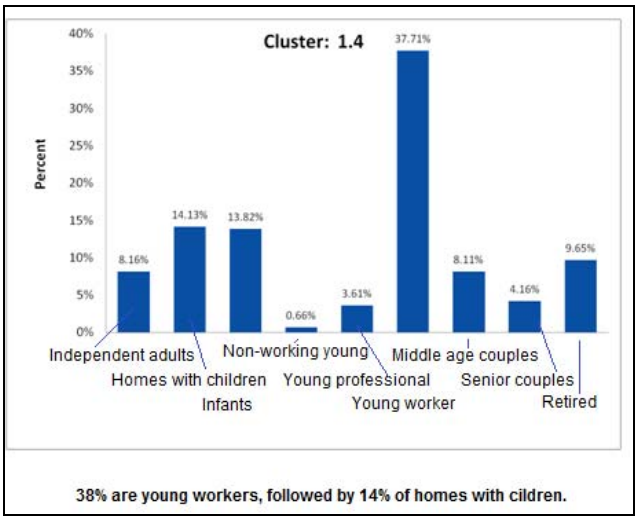

Figure 7. Socio-demographic characterization of Cluster 1.4..

IV. CONCLUSIONS

We have discussed a method which allows us to determine the number of clusters and their characteristics without appealing to expert knowledge of the information present in the data base under consideration. We have shown that it is, in general, possible to find a representative, usually much smaller, data set with which to work in more efficient (i.e. faster, easier to maintain) way instead of working with the original, usually much larger database. This is achieved without detriment of the statistical conclusions derived from the smaller DB. That is to say, we have shown how to replace a large data base with a smaller but statistically equivalent one. We have discussed the steps to determine the compact data base. From the compact data base we have successfully trained a Fuzzy C-Means set of plausible clusters and measured the information behavior of the different alternatives from which we found the number of clusters to be 6. We have trained a SOM and determined the centers of the 6 clusters. From these we have conducted the interpretation of every cluster from a non-algorithmic more intuitive point of view.

The methodology is currently at use in the Company worldwide.

ACKNOWLEDGMENT

We wish to acknowledge the continued support and enthusiasm for our work from the Asociación Mexicana de Cultura, A.C.

REFERENCES

1. Palpanas, T.: Knowledge Discovery in Data Warehouses. ACM SIGMOD record. Vol. 29, Issue 3 (2000) 88 – 100.
2. Silva, D. R., Pires, M. T.: Using Data Warehouse and Data Mining Resources for Ongoing Assessment of Distance Learning . IEEE ICALT Proceedings (2002).
3. Jain, K., Murty, M. N., Flynn, P.J.: Data Clustering: A Review. ACM Computing Surveys, 31(3), (1999) 264-323.
4. Berkhin, P.: Survey of Clustering Data Mining Techniques. Accrue Software Inc. (2002)
5. Kleinberg, J., Papadimitriou, C., Raghavan, P.: Segmentation Problems. Journal of the ACM, Vol. 51, No. 2, (2004) 263-280.
6. Guha, S., Rastogi, R., Shim, K.: CURE: An efficient clustering algorithm for Large Databases. ACM SIGMOD Proceedings (1998). 73 -84.
7. Peter, W., Chiochetti, J., Giardina, C.: New unsupervised clustering algorithm for large datasets. ACM SIGKDD Proceedings (2003). 643-648.
8. Raymong, T. N., Jiawei H.: Efficient and Effective Clustering Methods for Spatial Data Mining. 20th International Conference on Very Large Data Bases, (1994). 144-155.
9. Cheng, D., Kannan, R., Vempala, S., Wang, G.: A Divide-and-Merge Methodology for Clustering. ACM SIGMOD Proceedings (2005). 196 – 205.
10. Jagadish, H.V., Lakshmanan, L.V., Srivastava, D.: Snakes and Sandwiches: Optimal Clustering Strategies for a Data Warehouse. ACM SIGMOD Proceedings (1999) 37-48
11. Palmer, C. R., Faloutsos, C.: Density Biased Sampling: An Improved Method for Data Mining and Clustering. ACM SIGMOD Record (2000). 82-92.
12. Liu, H., Motoda, H.: On Issues of Instance Selection. Data Mining and Knowledge Discovery, Vol. 6, Number 2. Springer (2002) 115-130.
13. Zhu, X., Wu, X.: Scalable Representative Instance Selection and Ranking. Proceedings of the 18th IEEE international conference on pattern recognition, (2006) 352-355.
14. Brighton, H., Mellish, C.: Advances in Instance Selection for Instance-Based Learning Algorithms. Data Mining and Knowledge Discovery, Vol. 6, (2002) 153-172.
15. Vu, K., Hua, K. A., Cheng, H., Lang, S.: A Non-Linear Dimensionality-Reduction Technique for Fast Similarity Search in Large Databases. ACM SIGMOD Proceedings, (2006) 527-538.
16. Zhang, D., Zhou, Z., Chen, S.: Semi-Supervised Dimensionality Reduction. Proceedings of the SIAM International Conference on Data Mining, (2007).
17. Fodor, I.K. A survey of dimension reduction techniques. U. S. Department of Energy, Lawrence Livermore National Laboratory, (2002).
18. Hair, J. F., Anderson, R. E., Tatham R. L., Black W. C.: Análisis Multivariante. 5th edn. Pearson Prentice Hall, Madrid (1999). 11-15.
19. Delmater, R., Hancock, M.: Data Mining Explained: A Manager's Guide to Customer-Centric Business Intelligence. Digital press (2001). Chapter 6.
20. Zhu, L.: Nonparametric Monte Carlo Tests and Their Applications. Springer Science+Business Media, Inc. (2005).
21. Slagle, J. R., Chang, C. L. and Heller, S.: A Clustering and data-reorganization algorithm, IEEE Trans. on Systems, Man and Cybernetics, (1975)
22. Bezdek, J. C.: Cluster Validity with Fuzzy Sets. Journal of Cybernetics, issue 3 (1974). 58-72.)