Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

Introducing prior knowledge into the clustering process

W. F. Castilho[1,2], H. A. do Prado[1,3] & M. Ladeira[4]
[1]*Catholic University of Brasilia, Brazil*
[2]*Federal Savings Bank, Brazil*
[3]*Brazilian Enterprise for Agricultural Research, Brazil*
[4]*University of Brasilia, Brazil*

Abstract

A cluster is a subset of elements from a population, defined as a high-density region separated from other clusters by low-density regions. By applying a clustering procedure, an analyst looks for a description about how data is structured in its dimension space. Then, the structure is explored in order to figure out interesting groups that can, eventually, be associated to classes. Since the late 90s, a significant effort has been undertaken to address the clustering approach for data mining. However, the contributions have focused mainly on algorithm optimization. An important aspect of the clustering process, namely, the use of prior knowledge to constrain the clusters definition, has rarely been contemplated. According to Hanson [1], "clustering of objects or events without a context, goal, or some information concerning the function of the derived clusters is not likely to be useful for real-world problems." Actually, clustering is neither a blind nor a strongly driven process. The most recent reference we could find related to the use of prior knowledge in clustering was Béjar et al. [2], who uses production rules based on expert experiments to constrain the process. In this paper we propose a meta-schema that takes into account the burden of prior knowledge, specifically the cohesion of variables in a domain theory, to constrain the way clusters are defined. The proposal is implemented through the introduction of specific changes to the K-Means algorithm, and the subsequent results are compared with the standard version of the same algorithm.

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

1 Introduction

Countless techniques and applications of multivariate analysis have been researched and published in recent times. An important set of those techniques comprises the clustering processes, especially their associated algorithms, which are generally part of what is denominated the Classification Theory. This is the process of identifying a new entity or observation with an established group of classes or categories. The essential attributes of each class are known, but uncertainty in a given attribution can exist. The last step in the progression of classification problems is the Clustering Analysis. Nothing, or virtually nothing, is known with respect to the structure of the classes. Starting from an available collection of entities, the clustering analysis seeks to discover a structure of categories that is appropriate to them. The objective is to find natural clustering of the entities based on some internal approach, classifying them into groups such that the degree of natural association is more extensive within the scope of the same group and smaller among members of different groups. The fundamental premises, therefore, center on the greater internal cohesion among the objects of a group and the maximum external isolation among the groups.

In chapter 2 of this paper, an analysis of the context of the application of the clustering techniques is offered. Chapter 3 discusses the large number of possible alternative groups in a clustering process and the high cost in terms of the processing of a non-heuristic solution. Chapter 4 addresses the influence, on the clustering process, of previous knowledge in relation to different aspects of the data domain structure. Chapter 5 presents the purpose of the informed clustering model, on the basis of the problem of patterns and the consideration of the variables, from the perspective of structures of previous knowledge, with respect to the data domain. Chapter 6 comments on the experiments and evaluation criteria of the results in terms of satisfactory clustering. In the final part, conclusions are presented.

2 Context of the clustering analysis

In about 1997, the processes and techniques of Clustering Analysis began to be researched with increasing intensity within the context of Knowledge Discovery from Databases (KDD), its underlying principles and applications. Using a descriptive approach, KDD researchers have attempted to uncover how a group of objects is organized within the space containing its dimensions by exploring that space in order to find interesting groups. A cluster can be conceived of as a subset of elements of a population E whose image in an Euclidean p-dimensional space I_p is formed by a group of points of high density, separated from other groups by regions I_p having a low density of points. Johnson [3] highlights, however, that the restriction required for the elements of E to be represented by a point in an Euclidean space can be unnecessary for some algorithms, especially for those that use similarity measures instead of distance measures.

According to Wrobel [4], currently, clustering processes have applications related to one of the more important parts of Artificial Intelligence: the formation

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

of concepts. Interesting applications in the form of clustering algorithms have appeared in the field of KDD, which can be seen as a collection of techniques for the discovery of knowledge achieved with the assistance of computers. KDD researchers have recently applied themselves to the task of description, which includes clustering and the discovery of associative rules. Using the descriptive approach, the researcher attempts to discover how a group of objects is organized within the space containing its dimensions, exploring that space in order to find interesting groups.

3 Stirling number and the alternative number of groups

According to Prado et al [5], the possible number of ways in which N instances of a data collection can be segmented in K clusters grows remarkably in relation to N, and adheres to the Stirling number of the second type produced from the recurrence of the following definition:

$S(n,k) = 1$, for $k > 0$ and $n = k$

$S(n,0) = 0$ for $n \geq 0$

$S(n,k) = S(n-1,k-1) + kS(n-1,k)$, for $0 < k < n$

which leads to the following equation:

$$S(n,k) = \frac{1}{k!}\sum_{p=0}^{k}(-1)^{k-p}\binom{k}{p}p^n \qquad (1)$$

Considering the size of current databases, which are in general very large, the number of possible ways of segmentation can, then, reach a high magnitude. With a collection of relatively small objects, the number of possible clusters is already quite large. S(25,10), for example, is equal to 1,203,163,392,175,387,500. Because the number K of clusters in which an N population should be segmented is not known a priori, the number of possibilities would be the sum of very large numbers.

Although in a clustering process the structure of the classes cannot be determined a priori, the KDD analyst has a quota of previous knowledge in relation to the desirable and undesirable characteristics for a classification outline. He has enough previous information to choose an optimal classification structure among the different ways of segmenting an N population. What makes this an impracticable procedure, even in the light of the high computational power of available processing, is the greatness order of the number of alternative clusters that would have to be assessed to arrive at the optimal solution. The solution to the problem is, thus, in the heuristic. It doesn't strictly guarantee a convergence for an optimal solution, for it only evaluates a subset of the possibilities of alternative clusters. However, a satisfactory solution is reached through the application of a heuristic procedure or algorithm. An interesting

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

alternative would be the use of a base of previous knowledge over the database to restrict the algorithm towards the maximization of the objective function of the clustering process.

4 Influence of previous knowledge

According to Prado et al. [5], clustering processes can involve quite complex activities with considerable theoretical and technical consequences. Some scientists are focused on exploring space in search of areas having reaction patterns similar to a certain sign; service providers mapping homogeneous areas on the basis of criteria regarding the use of specific services. Those databases can grow significantly, at times reaching considerable dimensions. Among the several existing clustering techniques, most require countless exploratory scans on the database to identify the groups. This compromises the practical use of the application of the common clustering techniques in the context of very large databases and accelerated growth. In considering the operations for selecting characteristics and parameters, as well as for selecting appropriate algorithms, while also optimizing their performance for the exploration of large databases, we find that a challenging horizon for KDD clustering work emerges.

Hanson [1] highlights the need for the existence of a quota of previous knowledge as a requirement for the application of the clustering process. Consequently, we can conclude that the clustering process is not a "blind initiative" or a highly directed process. It is significantly influenced by the analyst: its goals, the perception of the underlying classes, and a great deal of additional information that affects the search for class structures.

In this context, the salient problem has to do with considering the previous knowledge in the definition of clusters. We find that this previous knowledge exists and influences the process. Therefore, it is clear that the process considers structures of previous knowledge. We can then think of the construction of a goal-outline based on previous knowledge capable of conditioning the results of a process of clustering analysis, where the pre-existing knowledge appropriately represented in the form of production rules can interact with the process, restricting it to the most appropriate clusters.

Therefore, the analyst's previous knowledge has significant influence on the clustering process, whether in terms of the decision regarding the number of clusters, the selection of the discriminating variables that are relevant to the domain under study, or the satisfactoriness of the cluster achieved. Without question this process takes into consideration pre-existing knowledge structures. Based on these ideas, we are able to imagine these structures, adequately represented interacting with the clustering process. It was in this context that Béjar et al. [2] presented the LINNEO system, proposing the use of production rules, through the use of a conceptual outline, as a way of searching for the most appropriate clustering. In his work, he presents a classification methodology aimed at discovering concepts related to structured domains, using them to organize hierarchies. To achieve this objective, LINNEO uses conceptual learning and classification techniques. The objective is to build knowledge bases

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

following the specialist's validation. Some techniques for improving results in the classification stages are used, on the basis of the specialist's partial knowledge (classification rules or causal or structural dependencies among attributes). He uses information to approach elements of the same domain to induce the formation of classes. Although this represents a significant effort, more research is necessary, however, to generalize the idea to other clustering techniques.

5 Informed clustering

Within the context of the Process of Knowledge Discovery from databases (KDD), which involves stages such as the "goals definition," the "data selection," the "pre-processing," the "transformation," the "data mining," and the "interpretation and evaluation," clustering analysis as a potential methodology selected as the most adequate in the *Dataming* stage involves a series of relevant decisions under the analyst's responsibility that carry decisive consequences with respect to the final results. Among these decisions, the following warrant emphasis: the definition of the objectives of the analysis, the selection of variables, the definition of the number of groups, the choice of the algorithm or clustering technique, the standardization and consideration of variables, among others.

As we have seen, there is a quota of previous knowledge that influences the clustering process, guiding the analyst's decision-making process. It is clear, therefore, that the process considers structures of previous knowledge. The purpose of the informed clustering model is to build a meta-scheme based on previous knowledge, which can favorably condition the results of a clustering analysis process. In this article, the first in a series introducing the informed clustering mode, we discuss the consideration of the structural and causal dependencies among the attributes through the process of its standardization and consideration. These processes are applied on the basis of similarity or homogeneity so as to include greater previous knowledge of the structure or data domain in the aggregation algorithm

5.1 Data matrix and cluster homogeneity coefficient

Having defined the objectives of the clustering analysis and their target classes, we seek, on the basis of this information, to select the group of variables that is relevant and discriminating in the classification problem under consideration.

The E population of objects through which we intend to apply the clustering analysis, defined by a group of vectors $\vec{x}_i = (x_{i1}, x_{i2}, ..., x_{ip}) \quad i = \overline{1,n}$, is represented by a matrix $X_{(N \times P)}$, known as a data matrix. It represents the values of the variables for each of the objects in the E population. The generally agreed approach is to indicate the objects or instances in the lines and the variables or attributes in the columns. The variables of a clustering problem are the describers of the natural groups or underlying classes, formed by characteristics, attributes, and other properties. They are selected a priori starting from the analyst's

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

understanding with regard to their relevance for the classification problem under consideration.

$$X_{(NxP)} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \quad \text{Data Matrix} \qquad (2)$$

where n is the number of objects or instances of an E population; and p refers to the variable discriminants selected for the classification task.

The clustering analysis results in a defined structure of categories, profiles, or classes defined on the basis of the homogeneity coefficient or clustering function. By applying the homogeneity coefficient among the elements of the E population to the data matrix, we obtain the matrix of distances, generally denominated matrix of distances $Z_{(N \times N)}$, which is a triangular square matrix, whose principal diagonal elements are null..

$$Z_{(NxN)} = \begin{pmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & 0 \end{pmatrix} \quad \text{Matrix of Distances and Similarities} \qquad (3)$$

where n is the number of objects or instances of an E population; and d_{ij} is the result of the calculation of the homogeneity coefficient between the elements x_i and x_j.

We know intuitively that in the formation of clustering on the basis of a collection of observations it is necessary to discern the similarity or difference of elements for purposes of gathering them into the most homogeneous groups possible. This clustering approach is the homogeneity coefficient. Also denominated the similarity coefficient, it can be divided, initially, into the following categories: differences, similarities, and coefficients for variables of different types. The similarity measures assume values between 0 and 1, while the measures of dissimilarity or distance can assume any positive value. This constitutes, therefore, the fundamental difference between the two categories of measure of the homogeneity coefficient.

5.2 Information matrix and clustering function

Within the context of the KDD process, a selection is made, with the participation of a specialist, of the attributes considered relevant and discriminant in the clustering analysis on the basis of the information about the data structure and the application domain. Fig. 1 presents a general outline of the process of informed clustering.

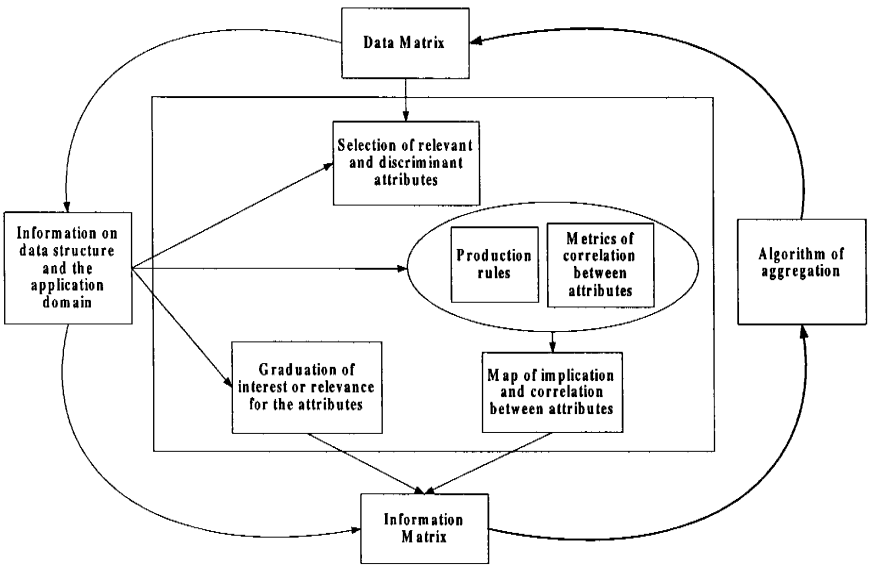


Figure 1: Informed clustering process.

In adopting the measure of dissimilarity, particularly the Euclidean distance, which will be used in the clustering algorithm, as a clustering function, our first concern is the standardization of the attributes or variables, which can be accomplished by adopting equation (4) and considering two objects identified by p and q, where $\sigma_{(i)}^2$ is the variant of the attribute.

$$d(p,q) = \sqrt{\sum_{i=1}^{n} \frac{(x_{pi} - x_{qi})^2}{\sigma_{(i)}^2}}. \tag{4}$$

An idea for pondering the distance that is highly convenient for purposes of the implementation derives from the "Mahalanobis distance", in which the information concerning the correlation metrics between attributes can be included by means of the covariance matrix, as set out in equation (5), wherein, as a result of $P_{(NxN)}$ the following particularities present themselves.

- $P = [V]^{-1}$, where V is the variance matrix, that is, $V = \left|diag\left(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2\right)\right|$ we arrive at the former case of standardization through variance;
- $P = [C]^{-1}$, where C is the covariance matrix, we arrive at the actual Mahalanobis distance;

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

- $P = I$, where P is the Identity Matrix, we arrive at the standard Euclidean Distance.

$$d(P,Q) = \sqrt{(x(P) - x(Q))' P(x(P) - x(Q))} \tag{5}$$

The Information Matrix (φ) is given, then, by equation (6), where δ is the matrix calculated from the graduation vector of the attributes' relevance and the map of their implication and correlation.

$$\varphi = [\delta(V + C)]^{-1} \tag{6}$$

6 Experiments

The clustering analysis was applied to the table of nutritional composition of foods deriving from the ENDEF (National Study of Family Expenses) table [6]. The aim of the initial experiments was to cluster the foods, comparing the results of the k-means algorithm using the standard Euclidean Distance with the same algorithm using the distance pondered by the attributes correlation information and a graduation scheme of relevance or interest defined on three levels: calories (cal), weight 3; minerals (mg) and vitamins (mg), weight 2; proteins (mg), lipids (mg), and glicids (mg), weight 1.

In the K-means algorithm that was implemented, a dissimilarity coefficient called "the sum of residual square" (SQRes) was used as an inverse measure of the internal homogeneity of the groups. Given a sequence of partitions in an E population of n objects up to a group of k groups, the sum of the residual square inside the j^{th} group is thus defined:

$$SQ\,\mathrm{Re}\,s(j) = \sum d^2(o_i(j); \bar{o}(j)) \quad [1 \le i \le n_j] \tag{7}$$

Where $o_i(j)$ represents i^{th} object of a partition or j group; $\bar{o}(j)$ represents the center of the j group, which is the point corresponding to the calculations of the average of its members' coordinates; d^2 represents the square of the *i* object's Euclidean Distance, of the j group, to its center.

The sum of the partition's residual squares is made up of the sum of the sums of residual squares within each group.

$$SQ\,\mathrm{Re}\,s = \sum SQ\,\mathrm{Re}\,s(j) \quad [1 \le j \le k] \tag{8}$$

The experiment found that the SQRes is very much affected by the definition of the number of clusters of the clustering process. It also found that the choice of more distant seeds reduces the number of interactions in the relocation phase of the algorithm that is aimed at a more satisfactory configuration.

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

A sample of 25 kinds of foods was selected, classified as fruits. From the experiments with the configuration of 3, 4, 5, 7, and 9 clusters, the most satisfactory in terms of SQRes minimization was the 7-cluster sample. Table 1 presents the results of the K-means standard and of the informed version with the structural correlations and attribute dependencies. Table 2 presents the distribution of fruits in the clusters, where the informed algorithm works with the defined levels of relevance.

Table 1: Sum of the residual square (SQRes).

Phase	Standard	Informed
Location	38.224,04	47.551,28
Relocation	32.140,12	30.999,48

Table 2: Resulting clusters.

Standard K-means	Informed K-means
Cluster 1: passion fruit; avocado	Cluster 1: passion fruit; melon;
Cluster 2: banana;	Cluster 2: grape;
Cluster 3: cashew;	Cluster 3: lemon; strawberry; tangerine;
Cluster 4: lemon; watermelon	Cluster 4: kaki; apricot; mango;
Cluster 5: damascus; cherry; pine cone; jabuticaba; strawberry; tangerine	Cluster 5: Jabuticaba; apple; watermelon;
Cluster 6: orange; apple; papaya;	Cluster 6: pineapple; cashew; guava;
Cluster 7: kaki; mango; guava; melon; pineapple; plum; fig; jack fruit; pear; grape.	Cluster 7: avocado; plum; banana; cherry; fig; pine cone; jack fruit; orange; papaya; pear.

The informed algorithm led to a more satisfactory configuration. Thus, cashew fruit, pineapple, and guava, previously located in different groups, ended up in Cluster 6 - a high average content of Vitamin C (166 g), followed by Cluster 3 (average of 51.33 g) made up of lime, strawberries, and tangerines. Mango was attracted to Cluster 4 - with a high average Vitamin A content (226.67 g) and consisting of apricots, followed by Cluster 1 (average of 93 g), composed of passion fruit and melon. Cluster 7 was shown to contain the highest caloric average (71 cal), along with the highest average content of lipids (1.82 g) and phosphorous (26.80 g). Grape distinguished itself in Cluster 2 with totals of 68 cal and high amounts of glicids (16.70 g). Cluster 5, with averages of 41 cal and 10.33 g of phosphorous, attracted apple, jatubicaba, and watermelon.

7 Conclusions and future work

The purpose of informed clustering is to use the information on data structure and on the application domain, not only to optimize the process, but also to work with the restrictions originating from the data space or from the knowledge of the specialist. This is what is known as the Domain Theory, which guides the

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

inductive process. The inclusion of the knowledge on correlation and structural and causal dependencies between attributes in the homogeneity coefficient has been studied. That allows us to minimize the weight of the attributes or even to discard them, thereby establishing levels of relevance from a map of implication and correlation. The specialist may also want to define a graduation scale of relevance or interest of the attributes in relation to the desired configuration for the clustering. Initial experiments have shown that this consideration, which derives from the specialist's interest, has an impact on the traditional definition of the calculated function based on the standard Euclidean Distance. Studies are being performed to better define the evaluation criteria of a more satisfactory clustering configuration. Studies and experiments are being conducted on the influence of the production rules on the homogeneity coefficient and on the ways for translating these in the information matrix. Another idea being studied and implemented is the utilization of production rules for the definition of former classes as hypotheses to be developed during the relocation phase of the algorithm for purposes of finding the best possible configuration.

References

[1] Hanson, S. J., Conceptual Clustering and Categorization: Bridging The Gap Between Induction and Causal Models. In: Kodratoff, Y. & Michalski, R. (Eds.), *Machine Learning: An Artificial Intelligence Approach*, Morgan: San Mateo, CA, 1990.

[2] Béjar, J., Cortes, U. & Poch, M., *LINNEO+: A Classification Methodology for Ill-structured Domains.* Technical Report LSI-93-22-R. Facultad d'Informática de Barcelona, 1993. http://www.lsi.upc.es/~bejar/reports.

[3] Johnson, S. C. *Hierarchical Clustering Schemes.* Psychometria, vol. 32, n° 3, pp. 241-254, 1997.

[4] Wrobel, S., *Concept Formation and Knowledge Revision.* Dordrecht, The Netherlands: Kluwer, pp. 240, 1994.

[5] Prado, H.A., Engel, P.M. & Hirtle, S.C., Clustering Algorithms for Data Mining. *Revista Tecnologia da Informação,* Brasilia, DF, v. 2, n. 1, pp. 51-58, 2000.

[6] Sá, N. G., *Nutrition and Diet – table*, Nobel: Rio de Janeiro, 2001.