A monothetic clustering method [*]

Marie Chavent (*) (**)
(*) INRIA Rocquencourt, Action SODAS,
Domaine de Voluceau, B.P.105,
78153 Le Chesnay cedex, France
(**) Université de Paris IX Dauphine, Lise Ceremade,
Place du Maréchal De Lattre de Tassigny, 75775 Paris cedex 16, France
e-mail : Marie.Chavent@inria.fr

Abstract : The proposed divisive clustering method performs simultaneously a hierarchy of a set of objects and a monothetic characterization of each cluster of the hierarchy. A division is performed according to the within-cluster inertia criterion which is minimized among the bipartitions induced by a set of binary questions. In order to improve the clustering, the algorithm revises at each step the division which has induced the cluster chosen for division.

Key Words : Hierarchical clustering methods, Monothetic cluster, Inertia criterion

1. Introduction

The objective of cluster analysis is to group a set Ω of N objects into clusters having the property that objects in the same cluster are similar to another and different from objects of other clusters. In the pattern recognition literature (Duda and Hart, 1973) this type of problem is referred to as unsupervised pattern recognition. The most common clustering methods are partitioning, hierarchical agglomerative and hierarchical divisive ones.

A partition of Ω is a list $(C_1, \ldots, C_K)$ of clusters verifying $C_1 \cup \ldots \cup C_K = \Omega$ and $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$. The essence of partitioning is the optimization an objective function measuring the homogeneity within the clusters and/or the separation between the clusters. Algorithms of the exchange type are frequently used to find a local optimum of the objective function, because of the complexity of the exact algorithms. Well-known partitioning procedures are the Forgy's k-means and the ISODATA methods, described in Anderberg (1973), and the dynamical clustering method (Diday, 1974).

Agglomerative and divisive hierarchical clustering methods are different, in the type of structure they are searching, from partitioning. Indeed, a hierarchy of Ω is a family H of clusters satisfying $\Omega \in H$, $\{\omega\} \in H$ for all $\omega \in \Omega$ and $A \cap B \in \{\emptyset, A, B\}$ for all $A, B \in H$. A hierarchy can be represented in the form of a tree or dendogram, that shows how the clusters are hierarchically organized.

The general algorithm for agglomerative clustering starts with N clusters, each consisting of

[*] Pattern Recognition Letters 19 (1998) 989-996

one element of Ω, and merges successively two clusters on the basis of a similarity measure. Well-known agglomerative hierarchical methods are described in Everitt (1974).

Divisive hierarchical clustering reverses the process of agglomerative hierarchical clustering, by starting with all objects in one cluster, and dividing successively each cluster into smaller ones. Those methods are usually iterative and determine at each iteration the cluster to be divided and the subdivision of this cluster. This process is continued until suitable stopping rule arrests further division.

There is a variety of divisive clustering methods (Kaufman and Rousseeuw, 1990). A natural approach of dividing a cluster C of n objects into two non-empty subsets would be to consider all the possible bipartitions. In this, Edward and Cavalli-Sforza (1965) choose among the $2^{n-1} - 1$ possible bipartitions of C, the one having the smallest within-cluster sum of squares. It is clear that such complete enumeration procedure provides a global optimum but is computationally prohibitive.

Neverless, it is possible to construct divisive clustering methods that does not consider all bipartitions. MacNaughton-Smith (1964) proposed an iterative divisive procedure using an average dissimilarity between an object and a group of objects. Chidananda Gowda and Krishna (1978) proposed a disaggregative clustering method based on the concept of mutual nearest neighborhood. Other methods taking as input a dissimilarity matrix are based on the optimization of criterions like the split or the diameter of the bipartition (Guénoche, Hansen and Jaumard, 1991; Wang, Yan and Sriskandarajah, 1996). Probabilistic validation approach for divisive clustering has also been proposed (Har-even and Brailovsky, 1995).

Another family of divisive clustering methods is monothetic. A cluster is called monothetic if a conjunction of logical properties is both necessary and sufficient for membership in the cluster (Sneath and Sokal, 1973). Indeed, each division is carried out using a single variable and by separating objects possessing some specified values of this variable from those lacking them. Monothetic divisive clustering methods have first been proposed in the particular case of binary data (Williams and Lambert, 1959; Lance and Williams, 1968). Since then, monothetic clustering methods have mostly been developed in the field of unsupervised learning and are known as descendant conceptual clustering methods (Michalski, Diday and Stepp, 1981; Michalski and Stepp, 1983).

In the field of discriminant analysis, monothetic divisive methods have also been widely developed. However, those methods are different from clustering in which the clusters are inferred from data. Indeed, a partition of Ω is pre-defined and the problem concerns the construction of a systematic way of predicting the class membership of a new object. In the pattern recognition literature, this type of classification is referred to as supervised pattern recognition. Divisive methods of this type are usually known as tree structured classifier like CART (Breiman, Friedman, Olshen and Stone, 1984) or ID3 (Quinlan, 1986). Recently, Ciampi (1994) insisted on the idea that trees offer a natural approach for both class formation (clustering) and development of classification rules (discrimination).

The clustering method proposed in this paper was developed in the framework of symbolic data analysis (Diday, 1995), which aims at bringing together data analysis and machine learning. More precisely, we propose a monothetic hierarchical clustering method performed in the spirit of CART from an unsupervised point of view. We have restricted the presentation of this method to the particular case of quantitative data. At each stage, the division of a cluster is performed according to the within-cluster inertia criterion (section **??**). This criterion is minimized among bipartitions induced by a set of binary questions (section **??**). Moreover, clusters are not systematically divided but one of them is chosen according to a specific criterion (section **??**). The divisions are stopped after a number of iterations given as input by the user, usually interested in few clusters partitions. The output of this divisive clustering method is an indexed hierarchy. It is also a decision tree (section **??**). The Ruspini's data are given as a first illustration of this method (section **??**). We propose a modification of the algorithm in order to soften the property shared by both agglomerative and divisive hierarchical methods, that efficient early partition cannot be corrected at a later stage. It consists in revising, after the division of a cluster, the previous division which has induced the cluster itself (section **??**). Before the conclusion (section **??**), the method is performed on Fisher's iris dataset (section **??**).

2. The inertia criterion

Let N be the number of objects in Ω. Each object is described on p real variables $Y_1, \ldots, Y_p$ by a vector $\mathbf{x}_i \in \mathbf{R}^p$ and weighted by a real value p_i $(i = 1, \ldots, N)$. Indeed, the analyst will prefer sometimes to weight the objects differently. For instance, countries could be weighted according to the size of their population. But usually, the weights are equal to 1 or equal to $\frac{1}{n}$.
The inertia I of a cluster C_k is an homogeneity measure equal to :

$$I(C_k) = \sum_{\mathbf{x}_i \in C_k} p_i d_M^2(\mathbf{x}_i, \overline{\mathbf{x}_k}) \tag{1}$$

where d_M is the Euclidean distance (M is a symmetric matrix positively defined) :

$$\forall \mathbf{x}, \mathbf{y} \in \mathbf{R}^p,\ d_M^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^t M (\mathbf{x} - \mathbf{y}) \tag{2}$$

and $\overline{\mathbf{x}_k}$ is the center of gravity of the cluster C_k :

$$\overline{\mathbf{x}_k} = \frac{1}{\mu_k} \sum_{\mathbf{x}_i \in C_k} p_i \mathbf{x}_i \tag{3}$$

$$\mu_k = \sum_{\mathbf{x}_i \in C_k} p_i \tag{4}$$

The within-cluster inertia W of a K-clusters-partition $P_K = (C_1, \ldots, C_K)$ is equal to :

$$W(P_K) = \sum_{k=1}^{K} I(C_k) \tag{5}$$

According to the Huygens Theorem, minimizing the within-cluster inertia of a partition (e.g. the homogeneity within the clusters) is equivalent to maximizing the between-cluster inertia (e.g. the separation between the clusters). This equals to :

$$B(W_K) = \sum_{k=1}^{K} \mu_k d_M^2(\overline{\mathbf{x}_k}, \overline{\mathbf{x}}) \tag{6}$$

3. Bipartitioning a cluster

Let C be a set of n objects. We want to find a bipartition (C_1, C_2) of C such that the within-cluster inertia is minimum. In the Edward and Cavalli-Sforza method (1965) one chooses the optimal bipartition (C_1, C_2) among the $2^{n-1} - 1$ possible bipartitions. It is clear that the amount of calculation needed when n is large will be prohibitive.

In our approach, to reduce the complexity, we divide C according to a *binary question* (Breiman, Friedman, Olshen and Stone, 1984) of the form "$Y_i \leq c$?" where $Y_i : \Omega \rightarrow \mathbf{R}$ is a real variable and $c \in \mathbf{R}$ is called the *cut point*.
The bipartition (C_1, C_2) induced by the binary question is defined as follows. Let ω be an object in C. If $Y_i(\omega) \leq c$ then $\omega \in C_1$ else $\omega \in C_2$. Those objects in C answering "yes" go to the left descendant cluster and those answering "no" to the right descendant cluster (Fig. **??**).

Figure 1: "Is *height* ≤ 172 ?"

For each variable Y_i, there will be at most $n-1$ different bipartitions (C_1, C_2) induced by the above procedure. Indeed, whatever the cut point c between two consecutive observations $Y_i(\omega)$ may be, the bipartition induced is the same. In order to ask only $n-1$ questions to generate all these bipartitions, we decide to use the $n-1$ cut points c, chosen as the middle of two consecutive observations $Y_i(\omega) \in \mathbf{R}$. Indeed, if the n observations $Y_i(\omega)$ are different, there are $n-1$ cut points on Y_i. If there are p variables, we choose among the $p(n-1)$ corresponding bipartitions (C_1, C_2), the bipartition having the smallest within-cluster inertia.

4. Choice of the cluster

Let $P_K = (C_1, \ldots, C_K)$ be a K-clusters-partition of Ω. At each stage, a new (K+1)-clusters-partition is obtained by dividing a cluster $C_k \in P_K$ into two new clusters C_k^1 and C_k^2. The purpose is to choose the cluster $C_k \in P_K$ so that the new partition,

$$P_{K+1} = P_K \cup \{C_k^1, C_k^2\} - \{C_k\}$$

has minimum within-cluster inertia.
We know that :

$$W(P_{K+1}) = W(P_K) - I(C_k) + I(C_k^1) + I(C_k^2)$$

In this, minimizing $W(P_{K+1})$ is equivalent to choosing the cluster $C_k \in P_K$ so that the difference between the inertia of C_k and the within-cluster inertia of its bipartition (C_k^1, C_k^2) is maximum. The criterion used to determine the cluster that will be divided is then equal to :

$$\Delta(C_k) = I(C_k) - I(C_k^1) - I(C_k^2) \tag{7}$$

Of course, it means that the bipartitions of all the clusters of the partition P_K have been defined previously. At each stage, the bipartitions of the two new clusters C_k^1 and C_k^2 are defined and used in the next stage.

5. The stopping rule and the output

The divisions are stopped after a number L of iterations and L is given as input by the user, usually interested in few clusters partitions. Indeed, the last partition obtained in the last iteration is a $L+1$-clusters-partition. The issue of stopping the divisions before obtaining the total hierarchy ($L = N$) is to ensure that the partitions of smallest within-cluster inertia of the total hierarchy are still in the hierarchy obtained after L iterations. This property is verified because the clusters are not systematically divided but one cluster is chosen according to the criterion Δ given in (**??**) which ensures that the partition induced by this division has minimum within-cluster inertia. However, this stopping rule doesn't solve the issue of determining the number of clusters in the dataset (Milligan and Cooper, 1985).

The output of this divisive clustering method is a hierarchy H which singletons are the $L + 1$ clusters of the partition obtained in the last iteration of the algorithm. Each cluster $C_k \in H$ is indexed by $\Delta(C_k)$. Because Δ is a non-decreasing mapping,

$$C_k \subset C_{k'} \Rightarrow \Delta(C_k) \leq \Delta(C_{k'}) \tag{8}$$

there will be no inversions in the dendogram of the hierarchy.

This hierarchy is also a decision tree. The L clusters are the leaves and the nodes are the binary questions selected by the algorithm. Each cluster is characterized by a rule defined according to the binary questions leading from the root to the corresponding leaves.

6. A simple example

The dataset is 75 points of $\mathbf{R}^2$ (Ruspini, 1970). We find successively a partition in 2,3 and 4 clusters ($L = 3$).

At the first stage, the method induces $2(75 - 1) = 148$ bipartitions. We choose among the 148 bipartitions (C_1, C_2), the one of smallest within-cluster inertia. It has been induced by the binary question "Is $Y_1 \leq 75.5$?". Notice that the number of subdivisions has been reduced from $2^{75} - 1 = 3,77 \times 10^{22}$ to 148.

At the second stage, we have to choose whether we divide C_1 or C_2. Here, we choose the cluster C_1 and its bipartition (C_1^1, C_1^2) because $\Delta(C_1) > \Delta(C_2)$. The binary question is "Is $Y_2 \leq 54$?".

At the third stage, we choose the cluster C_2 and its bipartition (C_2^1, C_2^2). The binary question is "Is $Y_2 \leq 75.5$?".

Finally, the divisive algorithm gives the 4 clusters represented Fig. **??**.

Figure 2: The 4-clusters partition

According to the dendogram of the hierarchy given figure **??**, the four clusters are characterized by four rules. For instance cluster C_1^1 is characterized by the following rule:

$$\text{If } [Y_1(\omega) \leq 75,5] \text{ and } [Y_2(\omega) \leq 54] \text{ then } \omega \in C_1^1.$$

Figure 3: The dendogram of the indexed hierarchy

This dendogram can be read as a decision tree and the rules can be read as classification rules of new objects to one of the four clusters.

7. Revising a binary question

The purpose is to enable the analyst to revise at each division of a cluster the binary question which has induced the cluster itself.

Let $\mathcal{C}$ be a cluster which has been divided in two clusters C and $\overline{C}$ according to the binary question "Is $Y_1 \leq c_1$?". Then C is chosen to be divided in two clusters C_1 and C_2 according to the binary question "Is $Y_2 \leq c_2$?".

Figure 4: Revising a binary question

At this stage, the binary question "Is $Y_1 \leq c_1$?" is revised by modifying the cut point c_1. We choose a new cut point c' among all possible cut points on Y_1, such that the 3-clusters-partition $(C'_1, C'_2, \overline{C'})$ induced by "Is $Y_1 \leq c'$?" and "Is $Y_2 \leq c_2$?" has minimum within-cluster inertia (figure **??**).

For instance, figure **??** gives the 3-clusters-partition of 320 points of $\mathbf{R}^2$ simulated from four 2-dimensional Gaussian distributions. The points have been divided first according to the binary question "Is $Y_2 \leq 10, 9$?" and then according to the binary question "Is $Y_1 \leq 8$?".

The first cut point 10.9 is then modified in order to find, with the second binary question "Is $Y_1 \leq 8$?", the 3-clusters-partition of minimum within-cluster inertia. The new cut point is 12.1 (figure **??**).

8. The Fisher's iris dataset

The above clustering method has been examined with the well-known Fisher's iris dataset. The length and breadth of both petals and sepals were measured on 150 flowers. There are three varieties of iris: Setoa, Versicolor and Virginia. There are 50 iris of each variety.

Figure 5: The two 3-clusters-partitions

Of course, the knowledge of this pre-defined 3-clusters-partition is not used in our unsupervised clustering procedure which is performed only with four quantitative variables : the petal width (PeWi), the petal length (PeLe), the sepal width (SeWi) and the sepal length (SeLe).
First, we have used the Euclidean distance d_M, with $M = I$, the identity matrix. Figure **??** gives the dendogram of the hierarchy and the 3-clusters-partition (C_1, C_2, C_3) obtained after two divisions of the dataset. The first cluster is composed of 53 iris including 50 Setoa, 3 Versicolor and no Virginia . Wholly, the 3-clusters-partition contains 19 iris misclassified.
The first binary question "Is PeLe ≤ 3.4" is then revised in order to improve the within-cluster inertia of the 3-clusters-partition. Figure **??** gives the dendogram of the hierarchy obtained with the revised binary question "Is PeLe ≤ 2.45". We can notice that the mis-classifications have been reduced to 16. Indeed, the 50 Versicolor are all in C_2.

Figure 6: Before the revision

Figure 7: After the revision

Dynamical clustering and Ward agglomerative hierarchical clustering methods have also been performed on the same dataset. The same distance was used. The partitions obtained with the two clustering methods contained the same number of mis-classifications since 16 iris were misclassified.
Secondly, we have used the normalized Euclidean distance d_M, with $M = D_{1/U_i^2}$ where U_i is the length between the maximum and the minimum value for the variable Y_i. The figure **??** gives the dendogram of the hierarchy obtained with this distance and we notice a reduction of the number of mis-classifications from 16 to 10 iris. It confirms the influence of the choice of the distance in the result of a clustering. Then, before the second division, we have normalized the Euclidean distance, according to the four length U_i computed locally in the cluster which

Figure 8: Global normalization

Figure 9: Local normalization

was divided. The figure **??** gives the dendogram of the hierarchy obtained with the locally normalized Euclidean distance. We can notice that the number of mis-classifications is now reduced to 6. It corresponds to an error rate of 0.04.

In their comparative study of the performance of different classifiers with Fisher's iris dataset, Weiss & Kulikowski (1991) give for the CART decision tree an error rate equal to 0.04. In this, we obtain with the Fisher's iris dataset comparable results with both unsupervised and supervised approaches. However, the goal of the proposed clustering method and the CART algorithm are different since we aim at inferring clusters from the data and CART algorithm aims at discovering classification rules.

9. Conclusion

The proposed clustering method has the advantages to be simple and to give simultaneously a hierarchy and a simple interpretation of its cluster. Moreover, it deals easily with very large datasets. Indeed, is possible to construct the hierarchy on a sample of the dataset, and to use the classification rules to assign the rest of the objects. This method has also given good results on the Fisher's iris dataset and on other real applications where it has been compared with the dynamical clustering method and the Ward agglomerative hierarchical method (Chavent, 1997). However, dividing a cluster according to a single variable can also be a deficiency in some situations. As for CART algorithm, in situations where the cluster structure depends on combinations of variables, the divisive method will do poorly at discovering the structure.

A perspective would be on the one hand to use a local stopping rule (Milligan and Cooper, 1985; Har-even and Brailovsky, 1995) for deciding if a cluster should be divided into two subclusters and on the other hand to divide a cluster according to a metric locally defined in the cluster itself.

References

Anderberg, M.R. (1973). *Cluster analysis for applications.* Academic Press, New York.

Breiman, L., J.H. Friedman, R.A. Olshen and C.J. Stone (1984). *Classification and regression Trees.* C.A:Wadsworth.

Chavent, M. (1997). *Analyse des Données Symboliques. Une méthose divisive de classification.* PhD Thesis, Université Paris-IX Dauphine, France.

Chidananda Gowda, K. and G. Krishna (1978). Disaggregative Clustering Using the Concept of Mutual Nearest Neighborhood. IEEE *Transactions on Systems, Man, and Cybernetics* 8, 888-895.

Ciampi, A. (1994). Classification and Discrimination : the RECPAM Approach. *In proc. of* COMPSTAT *'94*, 129-147.

Diday, E. (1974). Optimization in non-hierarchical clustering. *Pattern Recognition* 6, 17-33.

Diday, E. (1995). Probabilist, possibilist and belief objects for knowledge analysis. *Annals of Operations Research* 55, 227-276.

Duda, R.O. and Hart, P.E. (1973). *Pattern Classification and Scene Analysis.* Wiley, New York.

Edwards, A.W.F. and L.L. Cavalli-Sforza (1965). A method for cluster analysis. *Biometrics* 21, 362-375.

Everitt, B. (1974). *Cluster Analysis.* Social Sciences Research Council, Heineman Educational Books.

Guénoche, A., P. Hansen and B. Jaumard (1991). Efficient algorithms for divisive hierarchical clustering. *Journal of Classification* 8, 5-30.

Har-even, M. and V.L. Brailvosky (1995). Probabilistic validation approach for clustering. *Pattern Recognition* 16, 1189-1196.

Kaufman, L. and P.J. Rousseeuw (1990). *Finding groups in data.* Wiley, New York.

Lance, G.N. and W.T. Williams (1968). Note on a new information statistic classification program. *The Computer Journal* 11, 195-197.

MacNaughton-Smith, P. (1964). Dissimilarity analysis : A new technique of hierarchical subdivision. *Nature* 202, 1034-1035.

Michalski, R.S. and E. Diday, R. Stepp (1981). A recent advance in data analysis : Clustering objects into classes characterized by conjunctive concepts. *Progress in Pattern Recognition*, L.N. Kanal and A. Rosenfeld (eds), North Holland, 33-56.

Michalski, R.S. and R. Stepp (1983). Learning from observations : Conceptual clustering. *Machine Learning: An Artificial Intelligence Approach*, R.S. Michalsky, J. Carbonell and T. Mitchell (eds), 163-190.

Milligan, G.W. and M.C. Cooper (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50, 159-179.

Quinlan, J.R. (1986). Induction of decision trees. *Machine Learning* 1, 81-106.

Ruspini, E.M. (1970). Numerical Methods for Fuzzy Clustering. *Information Science* 2, 319-350.

Sneath, P.H. and R.R. Sokal (1973). *Numerical Taxonomy.* Freeman and company, San Francisco.

Weiss, S.M. and C.A. Kulikowski (1991). *Computer systems that learn : Classification and prediction methods from statistics, neural network, machine learning, and expert systems.* San Mateo, Calif:Morgan Kaufmann.

Williams, W.T. and J.M. Lambert (1959). Multivariate methods in plant ecology. *Journal of Ecology* 47, 83-101.

Wang, Y. and H. Yan, C. Sriskandarajah (1996). The weighted Sum of Split and Diameter Clustering. *Journal of Classification* 13, 231-248.

List of Figures