

Empirical comparison of a monothetic divisive clustering method with the Ward and the k-means clustering methods.*

Marie Chavent¹ and Yves Lechevallier²

¹ Mathématiques Appliquées de Bordeaux, UMR 5466 CNRS,
Université Bordeaux1, 351, Cours de la libération, 33405 Talence Cedex, France

² Institut National de Recherche en Informatique et en Automatique,
Domaine de Voluceau-Rocquencourt B.P.105, 78153 Le Chesnay Cedex, France

Abstract. DIVCLUS-T is a descendant hierarchical clustering methods based on the same monothetic approach than segmentation but from an unsupervised point of view. The dendrogram of the hierarchy is easy to interpret and can be read as decision tree. We present DIVCLUS-T on a small numerical and a small categorical example. DIVCLUS-T is then compared with two polythetic clustering methods: the Ward ascendant hierarchical clustering method and the k-means partitional method. The three algorithms are applied and compared on six databases of the UCI Machine Learning repository.

1 Introduction

The aim of this paper is to present a descendant hierarchical clustering method called DIVCLUS-T and to compare this new method with two well-known clustering methods: the Ward ascendant hierarchical clustering method and the k-means partitional method.

Descendant hierarchical clustering algorithm consists in recursively splitting a cluster into two sub-clusters, starting from the main data set Ω . At each stage a cluster from the partition in k clusters obtained at the previous stage is chosen and split in order to find a new partition in $k + 1$ clusters which optimizes an adequacy measure.

In DIVCLUS-T the measure of heterogeneity of a cluster is the inertia. The bi-partitional algorithm and the choice of the cluster to split are based on the minimization of the within-cluster inertia. The complete enumeration of all the possible bi-partitions is avoided by using the same monothetic approach than Breiman et al. (1984) who proposed and used binary questions in a recursive partitional process, CART, in the context of discrimination. Here we use binary questions in the context of descendant hierarchical clustering. DIVCLUS-T is then a DIVisive CLUStering method and an unsupervised segmentation method where the output is not a classification tree but a CLUStering-Tree. Because the dendrogram can be read as a decision tree,

* Proceedings of the IFCS'2006, Springer, 83-90. 2006.

it provides simultaneously partitions into homogeneous clusters and a simple interpretation of those clusters.

In Chavent (1998) a simplified version of DIVCLUS-T was presented in the particular case of quantitative data. It was applied in Chavent et al. (1999) with another monothetic divisive clustering method (based on correspondence analysis) to a categorical data set of healthy human skin data and more recently to accounting disclosure analysis (Chavent et al. (2005)). A hierarchical divisive monothetic clustering methods based on the poisson processes has also been proposed in Pircon (2004). A complete presentation of DIVCLUS-T for numerical and for categorical data as well as those algorithm and its complexity are given in Chavent et al. (2006).

Having a simple interpretation of the clusters is an advantage of the monothetic approach. By contrast the monothetic approach should induce partitions of worst quality (according to the within-cluster inertia). The aim of this paper is then to compare the quality of partitions performed with the monothetic method DIVCLUS-T with the quality of partitions performed with two polythetic methods (WARD and the k-means) based on the same measure of adequacy i.e. the within-cluster inertia. After a small presentation of the monothetic descendant hierarchical clustering method on two simple examples (a numerical and a categorical), we compare the quality of the partitions performed with DIVCLUS-T, WARD and the k-means algorithms on six datasets of the UCI Machine Learning repository (Hettich et al. (1998)). Because those three methods are based on the minimization of the within-cluster inertia, we will compare the proportion of the total inertia explained by the partitions performed by those three algorithms on the six databases.

2 Two examples

In the first example the divisive method is applied to a well-known numerical dataset: the protein consumption data table (Hand et al. 1994).

The dendrogram of the hierarchy built with DIVCLUS-T is given Figure 1 and the dendrogram of the hierarchy built with WARD is given Figure 2. We notice that the dendrogram Figure 1 differs from the dendrogram Figure 2 in the monothetic description of each level. For instance we can read that the south located countries of the cluster {Italy, Greece, Spain, Port} are characterized by their Nuts and Fruits/Vegetable consumption ($Nuts > 3.5$) whereas the north European countries of the cluster {Fin, Nor, Swed, Den} are characterized by their Fish consumption ($Fish > 5.7$).

In order to compare the quality of the two hierarchies which are nearly close, we have also compared the heterogeneity of the k -cluster partitions. We see Table 1 that the proportion of the explained inertia is better for the partitions of DIVCLUS-T from 2 to 4 clusters and better (or equal) for the partitions of WARD from 4 to 10 clusters. A reason could be that few clusters

partitions are obtained in the first stages of descendant hierarchical clustering whereas they are obtained in the last stages of ascendant hierarchical clustering.

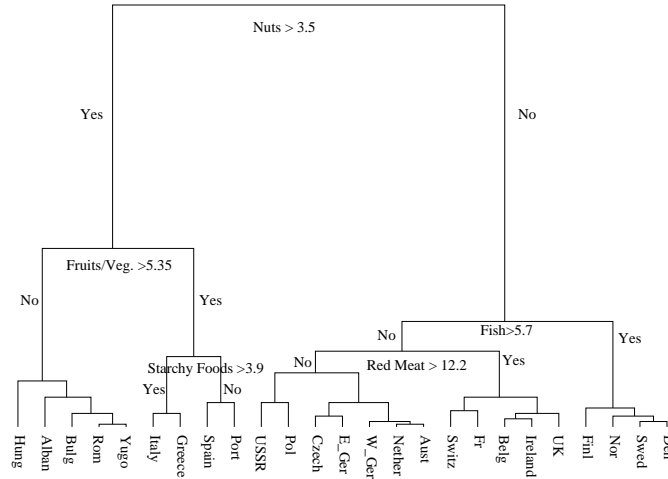


Fig. 1. DIVCLUS-T dendrogram for protein data

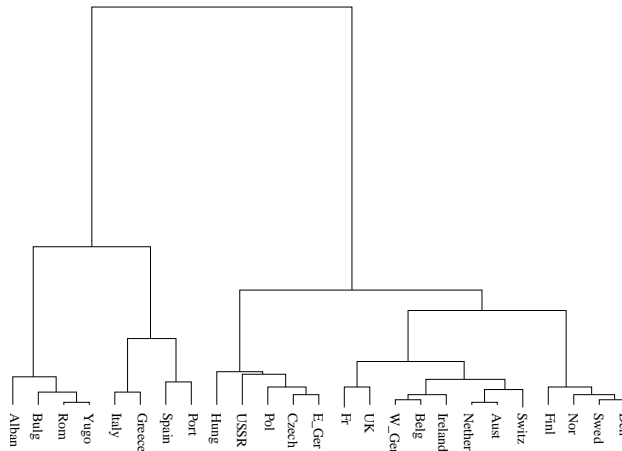


Fig. 2. WARD dendrogram for protein data

In the second example the divisive method is applied to a categorical dataset where 27 races of dogs are described by 7 categorical variables. The dendrogram of the hierarchy and the 7 first binary questions are given Figure 3.

k	2	3	4	5	6	7	8	9	10
DIVCLUS-T	37.1	50.6	59.2	65.5	71.2	73.5	79.3	81.6	84
WARD	34.7	48.5	58.5	66.7	72.4	75.5	79	81.6	84

Table 1. Proportion of the inertia explained by the k -clusters partitions of DIVCLUS-T and WARD on the protein data

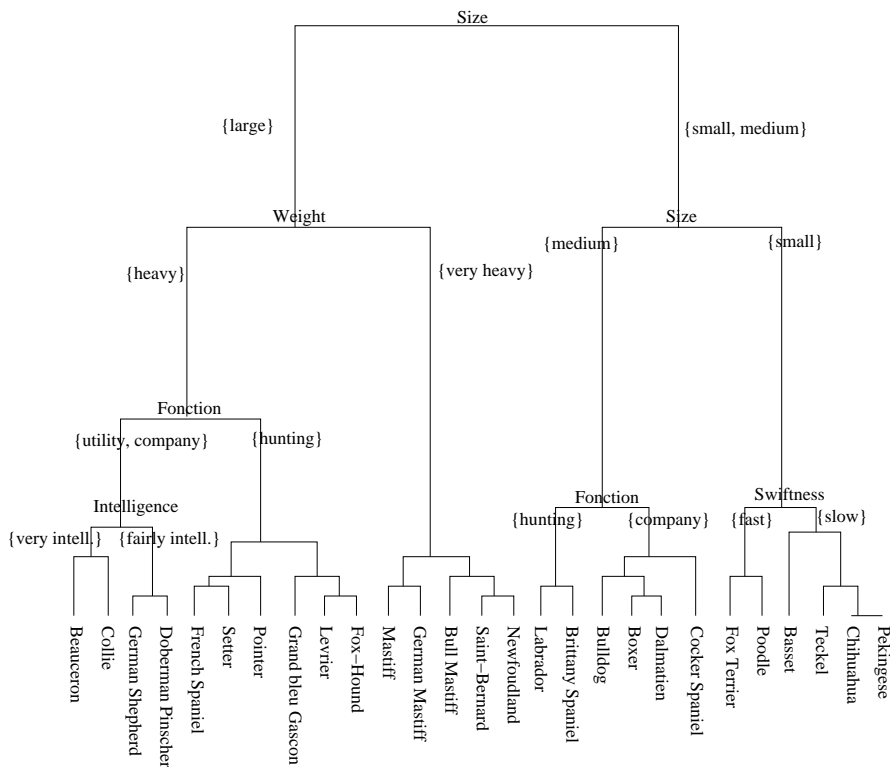


Fig. 3. DIVCLUS-T dendrogram for dogs data

At the first stage, the divisive clustering method performs a bi-partition of the 27 dogs. There are 17 different binary questions and 17 bi-partitions to evaluate: two variables are binary (and induce two different bi-partitions) and the five other variables have 3 categories and induce then 5×3 different bi-partitions. The question “Is the size large?” which induces the bi-partition of smallest within-cluster inertia is then chosen. For each sub-cluster the “best” bi-partition is then performed in the same way. The inertia variation obtained by splitting the 15 “large” dogs is slightly smaller than the one obtained by splitting the 12 “small or medium” dogs. This latter cluster is then divided. This process is repeated here until getting singleton clusters or clusters of identical dogs. The Pekingese and the Chihuahua for instance have exactly

the same description and can not then be divided. Finally the divisions are stopped after 25 iterations.

Because WARD can only be applied to quantitative data we have applied WARD on the 12 principal components performed by Multiple Factorial Analysis. The dendrogram obtained with WARD is identical to the one obtained with DIVCLUS-T.

3 Comparison with WARD and the k-means

We have applied DIVCLUS-T, WARD and the k-means algorithms on 3 numerical and 3 categorical datasets of the UCI Machine Learning repository (Hettich et al. (1998)). A short description of the 6 databases is given Table 2.

Name	Type	Nb objects	Nb variables(nb categories)
Glass	numerical	214	8
Pima Indians diabete	numerical	768	8
Abalone	numerical	4177	7
Zoo	categorical	101	15(2) + 1(6)
Solar Flare	categorical	323	2(6) + 1(4) + 1(3) + 6(2)
Contraceptive Method Choice (CMC)	categorical	1473	9(4)

Table 2. Databases descriptions

The quality of the partitions built by those three clustering methods on those 6 datasets can be compared with the proportion of explained inertia criterion. This criterion, noted E , takes its values between 0 and 100 (percent). It is equal to 0 for the singleton partition and it is equal to 100 for the partition reduced to one cluster (Ω). Because E decreases with the number of clusters k of the partition, it can be used only to compare partitions having the same number of clusters. Of course a partition P is “better” (for the inertia criterion) than a partition P' if $E(P) > E(P')$.

We have built the partitions from 2 to 15 clusters for the three numerical databases (see Table 3) and for the three categorical databases (see Table 4). For each database the two first columns give the proportion of explained inertia of the partitions built with DIVCLUS-T and WARD. The third column (W+km) gives the proportion of explained inertia of the partitions built with the k-means (km) when the initial partition is performed with WARD (W). As already stated two proportions of explained inertia can be compared only for partitions of the same database and having the same number of clusters. For this reason we will never compare two values in two different rows and two values of two different databases.

First we compare the results obtained on the three numerical databases (Table 3). For the Glass and the Pima databases the proportions of explained inertia in columns DIV and WARD are about the same for the few clusters

K	Glass			Pima			Abalone		
	DIV	WARD	W+km	DIV	WARD	W+km	DIV	WARD	W+km
2	21.5	22.5	22.8	14.8	13.3	16.4	60.2	57.7	60.9
3	33.6	34.1	34.4	23.2	21.6	24.5	72.5	74.8	76.0
4	45.2	43.3	46.6	29.4	29.4	36.2	81.7	80.0	82.5
5	53.4	53.0	54.8	34.6	34.9	40.9	84.2	85.0	86.0
6	58.2	58.4	60.0	38.2	40.0	45.3	86.3	86.8	87.8
7	63.1	63.5	65.7	40.9	44.4	48.8	88.3	88.4	89.6
8	66.3	66.8	68.9	43.2	47.0	51.1	89.8	89.9	90.7
9	69.2	69.2	71.6	45.2	49.1	52.4	91.0	90.9	91.7
10	71.4	71.5	73.9	47.2	50.7	54.1	91.7	91.6	92.4
11	73.2	73.8	75.6	48.8	52.4	56.0	92.0	92.1	92.8
12	74.7	76.0	77.0	50.4	53.9	58.0	92.3	92.4	93.0
13	76.2	77.6	78.7	52.0	55.2	58.8	92.6	92.7	93.3
14	77.4	79.1	80.2	53.4	56.5	60.0	92.8	93.0	93.7
15	78.5	80.4	81.0	54.6	57.7	61.0	93.0	93.2	93.9

Table 3. Continuous databases

partitions. As expected (because DIVCLUS-T is descendant and WARD is ascendant) when the number of clusters increases WARD tends to become better than DIVCLUS-T. In the third column (W+km) the k-means algorithm is performed on the WARD partition (taken as initial partition) and the proportion of explained inertia is then necessarily greater than the one in the second column WARD.

For the Abalone database which is bigger than the two others (4177 objects), DIVCLUS-T is better than Ward for the partitions in 2 and 4 clusters. Afterwards the results obtained with the three methods are very close. A reason for having better results of DIVCLUS-T on the abalone dataset is perhaps the greater number of objects in this database. Indeed the number of bi-partitions considered for optimization at each stage increases with the number of objects. We can then expect to have better results with databases having more objects. For instance DIVCLUS-T will search at the first stage the bi-partition of smallest within-cluster inertia among nearly 7×4177 bi-partitions for the Abalone database and among nearly 8×214 bi-partitions for the Glass database.

With the three categorical databases (Table 4) we obtain the same kind of results. For the Solar Flare and the CMC databases the proportions of explained inertia obtained with DIVCLUS-T are greater than the those obtained with WARD for the partitions until respectively 11 and 9 clusters. The proportion in the DIV column is also sometimes greater than the one in the W+km column (for the partition in 6 clusters of the Solar Flare database for instance). For the Zoo database the results obtained with DIVCLUS-T are slightly less good than the one obtained with WARD and then than the one obtained with the "Ward+k-means" strategy. A is maby that all the variables in the Zoo database are binary and as already stated the quality of the

K	Zoo			Solar Flare			CMC		
	DIV	WARD	W+km	DIV	WARD	W+km	DIV	WARD	W+km
2	23.7	24.7	26.2	12.7	12.6	12.7	8.4	8.2	8.5
3	38.2	40.8	41.8	23.8	22.4	23.8	14.0	13.1	14.8
4	50.1	53.7	54.9	32.8	29.3	33.1	18.9	17.3	20.5
5	55.6	60.4	61.0	38.2	35.1	38.4	23.0	21.3	24.0
6	60.9	64.3	65.1	43.0	40.0	42.7	26.3	24.9	27.7
7	65.6	67.5	68.4	47.7	45.0	47.6	28.4	28.1	29.8
8	68.9	70.6	71.3	51.6	49.8	52.1	30.3	30.7	32.7
9	71.8	73.7	73.7	54.3	53.5	54.6	32.1	33.4	35.2
10	74.7	75.9	75.9	57.0	57.1	58.3	33.8	35.5	37.7
11	76.7	77.5	77.5	59.3	60.4	61.7	35.5	37.5	40.1
12	78.4	79.1	79.1	61.3	62.9	64.4	36.9	39.4	41.5
13	80.1	80.6	80.6	63.1	65.2	65.7	38.1	41.0	42.9
14	81.3	81.8	81.8	64.5	66.2	67.7	39.2	42.0	44.2
15	82.8	82.8	82.8	65.8	68.6	69.3	40.3	43.1	44.9

Table 4. Categorical databases

results (in term of inertia) may depend on the number of categories and of variables.

4 Conclusion

Imposing the monotheticity of the clusters in the hierarchical process like in DIVCLUS-T is of course an advantage in term of interpretation of the results. We have seen that the dendrogram has the advantage to give a very simple interpretation of the levels of the hierarchy. Of course this advantage has to be balanced with a relative rigidity of the clustering process. Simple simulations should be able to show easily that DIVCLUS-T is unable to find correctly clusters of specific shapes. But what are the shapes of the clusters in real datasets ? We have seen on the six databases of the UCI Machine Learning repository that the proportions of explained inertia of the partitions performed with DIVCLUS-T are very comparable to those obtained with the Ward or the k-means algorithms, particularly for the few clusters partitions (at the top of the dendrograms). A more complete comparative study of those three clustering methods remain necessary in particular in order to better understand the influence of the number of objects, categories and variables in the quality of the results, combined with a study of their stability.

References

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A. and STONE, C.J. (1984): *Classification and regression Trees*, C.A:Wadsworth.

- CHAVENT, M. (1998): A monothetic clustering method. *Pattern Recognition Letters*, 19, 989-996.
- CHAVENT, M., GUINOT, C., LECHEVALLIER Y. and TENENHAUS, M. (1999): Méthodes divisives de classification et segmentation non supervisée: recherche d'une typologie de la peau humaine saine. *Revue Statistique Appliquée*, XLVII (4), 87-99.
- CHAVENT, M., DING, Y., FU, L., STOLOWY and H., WANG, H. (2005): Disclosure and Determinants Studies: An extension Using the Divisive Clustering Method (DIV). *European Accounting Review*, to publish.
- CHAVENT, M., BRIANT, O. and LECHEVALLIER, Y. (2006): *DIVCLUS-T: a new descendant hierarchical clustering method*. Internal report U-05-15, Laboratoire de Mathématiques Appliquées de Bordeaux.
- HAND, D.J., DALY, F., LUNN, A.D., McCONWAY, K.J. and OSTROWSKI, E. (eds.) (1994), *A Handbook of Small Data Sets*, London: Chapman & Hall.
- HETTICH, S., BLAKE, C.L. and MERZ, C.J. (1998): *UCI Repository of machine learning databases*, <http://www.ics.uci.edu/mlearn/MLRepository.html>. Irvine, CA: University of California, Department of Information and Computer Science.
- PIRCON, J.-Y. (2004): *La classification et les processus de Poisson pour de nouvelles méthodes de partitionnement*. Phd Thesis, Facultés Universitaires Notre-Dame de la Paix, Belgium.