Species clustering via classical and interval data representation

Marie Chavent

Université Bordeaux 1,
Institut de Mathématiques de Bordeaux, UMR CNRS 5251,
351 cours de la libération,
33405 Talence cedex, France,
chavent@math.u-bordeaux1.fr

Abstract. Consider a data table where n objects are described by p numerical variables and a qualitative variable with m categories. Interval data representation and interval data clustering methods are useful for clustering the m categories. We study in this paper a data set of fish contaminated with mercury. We will see how classical or interval data representation can be used for clustering the species of fish and not the fish themselves. We will compare the results obtained with the two approaches (classical or interval) in the particular case of this application in Ecotoxicology.

1 Introduction

Interval data representation can be very useful to study groups of objects described by quantitative variables. Describing a group of objects on each variable by an interval of values rather than by a mean value, allows to reflect the variability that underlies the observed measurement. Many data analysis techniques have been extended to treat such new data description (see for instance Bock and Diday (2000)). But a question frequently asked while applying these techniques is the following: 'Are the results obtained with intervals different than those obtained with means?'. Of course it is very difficult to answer this question not only because the data tables are different but also because the techniques are different. We will however try to answer this question in the particular case a real application in Ecotoxicology and in the context of clustering. This application is concerned by mercury contamination of fish in rivers of French Guyana (Chavent et al. (2000)). Our problem was to define a partition of the different species of fish according to their mercury concentrations in fives organs (gills, liver, intestine, stomach, kidney). A first partition was calculated with point-valued data and a second one with interval-valued data. The two partitions were compared not numerically (because no numerical comparison between the two partitions is possible) but according to an external partition (the diet of the species) and according to a fuzzy partition of the species (obtained by clustering the fish themselves).

Let consider the general case of a data table where n objects are described by p variables, one of them is qualitative with m categories and the $p-1$ others are quantitative. The problem is to find a partition in K clusters, not of the n objects, but of the m categories. In the application, the data table describes $n = 67$ fish of $m = 10$ different species by 5 quantitative variables (their mercury contaminations in fives organs). We present here three different approaches to find a partition of the 10 species into homogeneous clusters.

- clustering the 67 fish described by the five quantitative variables. It gives a fuzzy clustering of the 10 species,
- clustering the 10 species described by mean values on the five variables,
- clustering the 10 species described by intervals on the five variables.

2 The data

The data were collected by researchers of the EPOC[1] laboratory: 265 fish of 36 different species were catch in 1997 in several French Guyana rivers and a sample of 67 fish of 10 species and 3 diet were selected (see Table 1).

Carnivorous	Omnivorous	Detritivorous
Ageneiosus brevifilis (7)	Leporinus fasciatus (3)	Doras micropoeus (8)
Cynodon gibbus (7)	Leporinus frederici (3)	Platydoras costatus (10)
Hoplias amara (10)		Pseudoancistrus barbatus (7)
Potamotrygon hystrix (4)		Semaprochilodus varii (8)

Table 1. Diet and frequency of each species in the sample

The researchers of the EPOC laboratory measured for each of the 67 fish the length, the weight and the mercury concentration in μg/g in the muscle and in 5 organs. After several pre-treatments (descriptive statistics, variable selection....), we retained the data table described Table 2.

	species	diet	ln(liver/muscle)	...	ln(stomach/muscle)
1	Ageneiosus brevifili	Carnivorous	-0,12	...	NA
2	Cynodon gibbus	Carnivorous	1,59	...	0,22
3	Leporinus frederici	Omnivorous	-0,04	...	-1,77
⋮	⋮	⋮	⋮	...	⋮
66	Doras micropoeus	Detritivore	0,8	...	-0,89
67	Doras micropoeus	Detritivore	1,34	...	-1,45

Table 2. Extract of the data table

[1] UMR CNRS 5805 EPOC (Environnements et Paloenvironnements OCéaniques)

The five quantitative variables of this data table are based on the ratio between the mercury concentration in the five organs and the mercury concentration in the muscle. These ratios were used because of the positive correlation between the mercury concentration variables. In a second time, the skewness of the distributions of the ratios has motivated the logarithmic transformation.

Figure 1 represents the 67 fish in the first factorial plane calculated with these five quantitative variable. Each fish is numbered according to its species (from 1 to 10).

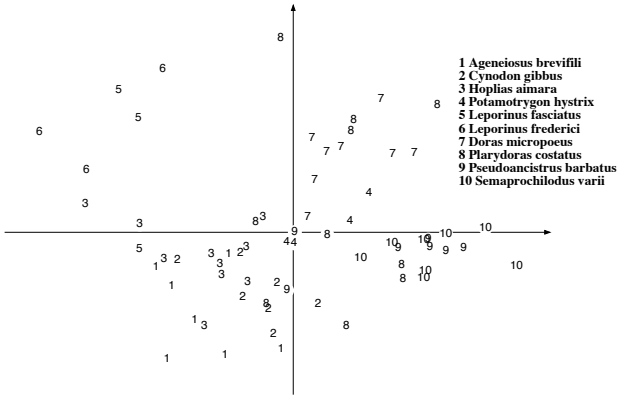

Fig. 1. The 67 fish in the first factorial plane, numbered from 1 to 10 according to their species.

We notice on this figure that the fish of the same species are rather close in the first factorial plane. As we will see in the next section when clustering the fish, those in the same species are mostly in same clusters. The partition of the 67 fish will then define a kind of fuzzy partition of the 10 species.

3 Fuzzy partition of the species

A partition in 4 clusters of the 67 fish described by the five quantitative variables of Table 2, was performed by Ward hierarchical clustering. The Table 3 gives the proportion of fish of each species in each cluster. The diet of the species is also indicated. We notice that all the fish of the three carnivorous species are in cluster1, and that this cluster contains no fish from another species. Obviously, a clustering of the 10 species should put the three carnivorous species in the same cluster. It means also that the carnivorous fish have the same kind of behavior in term of mercury concentration. In the same way, we see that cluster2 contains only omnivorous fish, and the of the omnivorous fish are almost in cluster2 (only one of the three Leporinus fasciatus

fish is in cluster1). The two omnivorous species should then be in the same cluster in a partition of the species. Five species of detritivorous are in two different clusters and two species of detritivorous are difficult to assign to one of the four clusters. This result is not surprising because a doubt remains concerning the diet of these species.

	cluster1	cluster2	*cluster3*	*cluster4*	Diet
Ageneiosus brevifili	**100**	0	0	0	**carnivorous**
Cynodon gibbus	**100**	0	0	0	**carnivorous**
Hoplias aimara	**100**	0	0	0	**carnivorous**
Doras micropoeus	0	0	*100*	0	*detritivorous*
Leporinus fasciatus	33.33	66.67	0	0	omnivorous
Leporinus frederici	0	100	0	0	omnivorous
Pseudoancistrus barbatus	14.29	0	0	*85.71*	*detritivorous*
Semaprochilodus varii	0	0	0	*100*	*detritivorous*
Platydoras costatus	20	0	40	40	detritivorous ?
Potamotrygon hystrix	50	0	25	25	detritivorous ?

Table 3. Proportion of fish of each species in each cluster and the diet of the species

Table 4 gives the crisp partition of 8 of the 10 species deduced from Table 3. The two species Platydoras costatus and Potamotrygon hystrix are not assigned to one of those clusters.

cluster1 (carnivorous)	cluster2 (omnivorous)	*cluster3 (detritivorus)*	*cluster4 (detritivorus)*
Ageneiosus brevifili	Leporinus fasciatus	*Doras micropoeus*	*Pseudoancistrus barbatus*
Cynodon gibbus	Leporinus frederici		*Semaprochilodus varii*
Hoplias aimara			

Table 4. Crisp partition of 8 of the 10 species

4 Classical data description and divisive clustering

In order to describe the 10 species with the 5 mercury concentration variables, a new data table was constructed. The fish of the same species were aggregated by calculating the mean value on each variable and Table 5 is the resulting classical data table.

The descendant hierarchical clustering method DIV (Chavent (1997)) was then applied to this data table and after three divisions, a four clusters partition of the 10 species was obtained (see Figure 2). This partition is not satisfactory according to the diet partition and according to the partition

species	ln(liver/musc)	ln(kidn/musc)	ln(gills/musc)	ln(intest/musc)	ln(stom/musc)
Ageneiosus brevifili	-0,39	-0,25	-1,54	-0,89	-1,25
Cynodon gibbus	1,05	0,24	-1,61	-1,29	-1,06
Hoplias aimara	0,26	0,764	-1,73	-1,36	-1,55
Doras micropoeus	1,72	2,11	-2,21	-0,78	-0,90
Leporinus fasciatus	-0,82	-0,28	-2,81	NA	-1,93
Leporinus frederici	-0,47	-0,65	-2,87	-1,61	-1,55
Pseudoancistrus barbatus	2,29	-1,00	NA	0,38	-0,24
Semaprochilodus vari	3,43	1,49	-1,64	0,02	-0,25
Platidoras costatus	1,58	1,51	-1,98	-0,28	-1,00
Potamitrigon hystrix	1,15	1,25	NA	-0,13	-0,76

Table 5. Point-type description of the 10 species

obtained by clustering the fish (Table 4). The two omnivorous species (Leporinus fasciatus, Leporinus frederici) are not in the same cluster and the two clusters of detritivorous species (Doras micropoeus against Pseudoancistrus barbatus and Semaprochilodus varii) highlighted Table 3 and Table 4, do not appear in this partition.

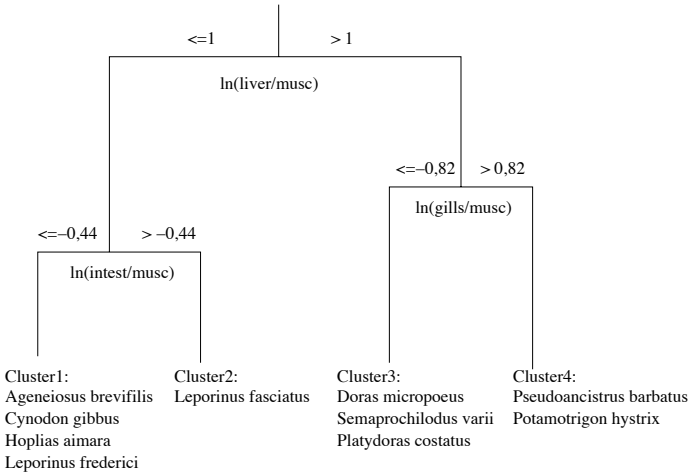

Fig. 2. Dendrogram of the upper hierarchy for classical data description.

The question was then: is this unsatisfactory result due to way the data were aggregated or to the clustering method itself? On order to answer this question, we applied an other clustering method, the Ward ascendant hierarchical clustering method, to the same data table. Figure 3 represents the 10 species described in Table 5, in the first factorial plane. Each species is numbered according to its cluster in the 4-clusters partition obtained with Ward. In this partition the two species (Leporinus fasciatus, Leporinus frederici) are in the same cluster. The inappropriate separation of these two species by DIV was perhaps due the monothetic constraint of this method. The three carnivorous species (Hoplias aimara, Cynodon gibbus, Potamotrygon hystrix) are

well gathered in one cluster but the separation of the detritivorous species Doras micropoeus from the two other detritivorous species Pseudoancistrus barbatus and Semaprochilodus varii, again do not appear in this partition.

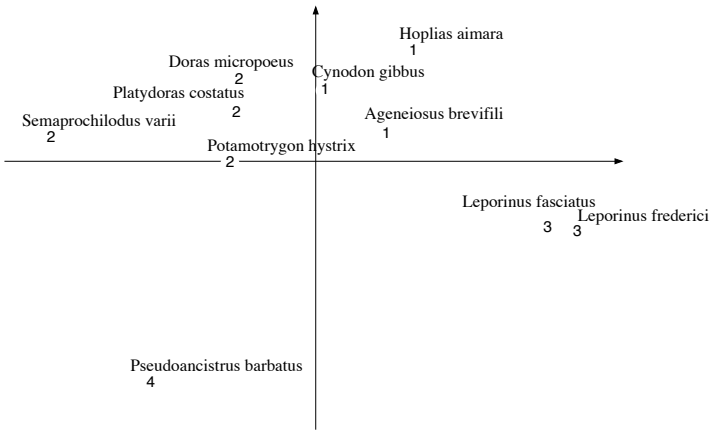

Fig. 3. First factorial plane of the 10 species (aggregated by the mean), numbered from 1 to 4 according to its cluster in the Ward partition.

5 Interval data description and divisive clustering

In a second time, the fish of the same species were aggregated by calculating an interval of values on each variable. Table 6 is the resulting interval data table calculated with the DB2SO method (see Stephan (1998)) and the SODAS software (see for instance Diday and Esposito (2003)).

species	ln(liver/musc)	ln(kidn/musc)	ln(gills/musc)	ln(intest/musc)	ln(stom/musc)
Ageneiosus brevifili	[-0.80:0.34]	[-1.50:0.35]	[-1.88:-1.21]	[-1.45:-0.48]	[-1.49:-1.05]
Cynodon gibbus	[0.12:1.59]	[-0.51:1.18]	[-1.91:-1.44]	[-1.75:-0.68]	[-1.61:0.22]
Hoplias aimara	[-0.44:0.90]	[-0.17:1.60]	[-1.98:-1.53]	[-2.17:-0.71]	[-2.36:-0.93]
Doras micropoeus	[1.34:2.12]	[1.47:2.69]	[-2.38:-2.21]	[-1.99:0.39]	[-1.45:-0.24]
Leporinus fasciatus	[-0.98:-0.58]	[-0.32:0.35]	[-3.00:-2.63]	NA	[-2.11:-2.76]
Leporinus frederici	[-0.82:-0.04]	[-0.95:-0.19]	[-3.27:-2.55]	[-1.74:-1.42]	[-2.03:-0.55]
Pseudoancistrus barbatus	[1.26:2.84]	[-1.00:1.00]	NA	[-0.31:0.68]	[-0.71:0.12]
Semaprochilodus vari	[2.70:3.96]	[1.11:1.91]	[-1.79:-1.40]	[-0.91:0.52]	[-0.74:0.22]
Platidoras costatus	[0.41:2.42]	[-0.02:2.75]	[-2.90:-1.27]	[-1.22:0.38]	[-1.41:-0.49]
Potamitrigon hystrix	[0.66:2.01]	[0.77:2.15]	NA	[-0.50:0.23]	[-0.80:-0.69]

Table 6. Interval type description of the 10 species

The divisive clustering method DIV for interval data description (Chavent (1997)), was applied to the 10 species described in Table 6. After three divisions, a four clusters partition of the 10 species was obtained (see Figure

4). This partition is more in adequation with the fuzzy partition obtained by clustering the fish (Table 4) than those obtained with the classical descriptions. The two omnivorous species are alone in one cluster. The three carnivorous species are alone in one cluster and the two clusters of detritivorous species (Doras micropoeus against Pseudoancistrus barbatus and Semaprochilodus varii) are found. The two detritivorous species Platydoras costatus and Potamitrigon hystrix that were not assign clearly to one cluster in the fuzzy partition (Table 3), are put together with the detritivorous species Doras Micropoeus. For all these reasons, this partition is more satisfactory than the one obtained with the classical data representation.

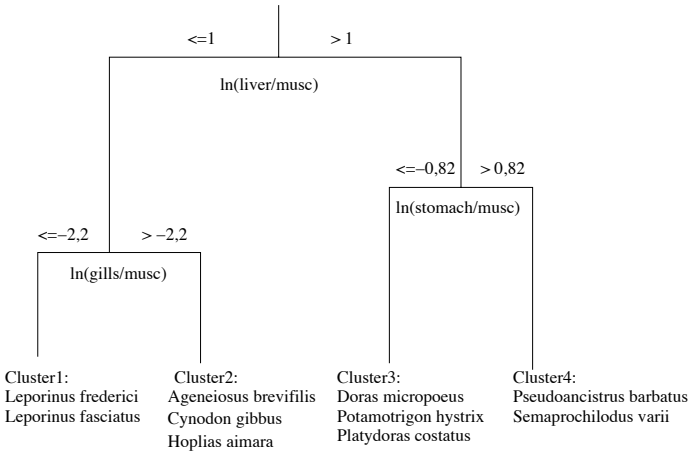

Fig. 4. Dendrogram of the upper hierarchy for interval data description.

Figure 5 gives an idea of the variation of the fish of the 10 species. Rectangles were drawn on the Figure 1 in order to represent the min-max variation of the fish of each species (numbered from 1 to 10) in each dimension of the first factorial plane. This figure helps understanding the partition obtained with DIV and the interval data description. The Semaprochilodus varii and the Pseudoancistrus barbatus for instance are in the same cluster because of the similarity between their positions and between their dispersions. The rectangle Platydoras castatus (8) shows an broad variability of the fish of this species. It was assigned to the same cluster than the rectangle Doras micropoeus (7) but this important variability questions on the signification of the proximity between the two species. The fuzzy partition of the species gives more precise results concerning the difficulty of clustering this species.

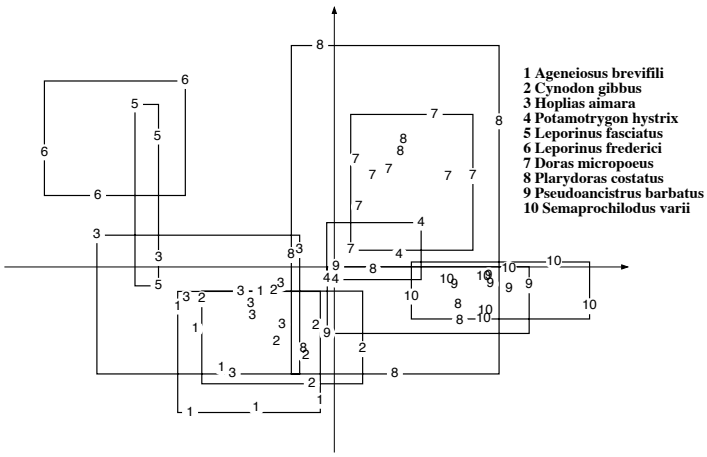

Fig. 5. Min-max variation of the fish of each species in the two dimensions of the first factorial plane.

6 Conclusion

This case study in Ecotoxicology was a good illustration of the use of interval data representation for clustering aggregated data. We proposed a three steps methodology: clustering the 67 fish to find a fuzzy partition of the species, clustering the species with point-type descriptions and clustering the species with interval-type descriptions. We compared the three partitions and we concluded that the partition obtained with the interval-type description is more in adequation with the diet of the species and with the fuzzy partition. This is a good result in a particular case showing the interest of interval data representation. Concerning the Ecotoxicological application, this study highlighted the discriminant power of the diet in term of mercury concentration and the existance of two clusters of detritivorous species.

Acknowledgments
The author thanks A. Boudou and R. Maury-Brachet (UMR CNRS 5805 EPOC laboratory of University Bordeaux 1) for providing data and collaborating on this application.

References

BOCK, H.-H. and DIDAY, E. (eds.) (2000): *Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data*. Springer Verlag, Heidelberg.

CHAVENT, M. (1997), *Analyse des données symboliques, une méthode divisive de classification*. PhD thesis of Paris IX-Dauphine University.

CHAVENT, M. (1998): A monothetic clustering method. *Pattern Recognition Letters, 19, 989-996.*

DIDAY, E., ESPOSITO, F. (2003): An introduction to symbolic data analysis and the SODAS software. *Intell. Data Anal. 7(6): 583-601.*

CHAVENT, M., LACOMBLEZ, C., BOUDOU A., MAURY-BRACHET R. (2000): Contamination par le mercure et classification d'espéces en Ecotoxicologie: approche classique, approche symbolique. *La revue Modulad, Décembre 2000, 19–32.*

STEPHAN, V. (1998): *Construction d'objets symboliques par synthèse des résultats de requêtes SQL.* PhD thesis of Paris IX-Dauphine University.