Evaluation d'une approche de classification conceptuelle

Marie Chavent *, Yves Lechevallier **

* Mathématiques Appliquées de Bordeaux, UMR 5466 CNRS
Université Bordeaux1, 351, Cours de la libération, 33405 Talence Cedex
chavent@math.u-bordeaux1.fr,
http://www.math.u-bordeaux1.fr/ chavent
**Institut National de Recherche en Informatique et en Automatique,
Domaine de Voluceau-Rocquencourt B.P.105, 78153 Le Chesnay Cedex
Yves.Lechevallier@inria.fr
http://www-rocq.inria.fr/axis/personnel/Yves.Lechevallier/yves.html

Résumé. L'objectif de ce travail est d'évaluer la perte d'information au sens de l'inertie entre des méthodes de partitionnement ou de classification hiérarchiques et une approche de classification conceptuelle. Nous voulons répondre à la question suivante : l'aspect simpliste du processus monothétique d'une méthode conceptuelle implique-t-il des partitions de moins bonne qualité au sens du critère de l'inertie ? Nous proposons de réaliser cette expérience sur 6 bases de l'UCI, trois de ces bases sont des tableaux de données quantitatives, les trois autres sont des tableaux de données qualitatives.

1 Introduction

Les méthodes conceptuelles de classification imposent une description monothétique des classes, cette contrainte forte devant a priori engendrer une perte de la qualité des partitions au sens des critères optimisés par les méthodes de partitionnement. Pour étudier ce phénomène nous avons comparé une méthode de classification conceptuelle appelée DIVCLUS-T qui est une méthode de classification descendante hiérarchique monothétique avec la méthode ascendante hiérarchique WARD et la méthode de partitionnement des k-means. Cette méthode conceptuelle s'applique à des données quantitatives et des données qualitatives. Cette comparaison a pu être effectuée car cette méthode conceptuelle optimise le même critère que les méthodes de WARD et des k-means. Ainsi comme WARD et les k-means, elle est basée sur la minimisation de l'inertie des classes mais à la différence de WARD et des k-means elle fournit par construction une interprétation simple et naturelle des classes. La question à laquelle nous allons chercher à répondre est la suivante : quel est le prix payé, en terme d'inertie, pour cette interprétation sous forme de règles des classes ?

Avant de présenter un peu plus en détail la méthode DIVCLUS-T et de comparer à partir de 6 bases de l'UCI les performances en terme d'inertie de DIVCLUS-T avec WARD et les k-means, nous donnons un petit exemple introductif. Il s'agit des données "protéines" de Hand et al. (1994) qui est un exemple classique de données quantitatives où 25 pays européens sont décrits en fonction de leur consommation en protéines dans neuf types d'aliments (viande

rouge, viande blanche, oeufs,...). La méthode Ward a d'abord été appliquée à ces données et le dendrogramme de la hiérarchie obtenue est représenté à la Figure 1.

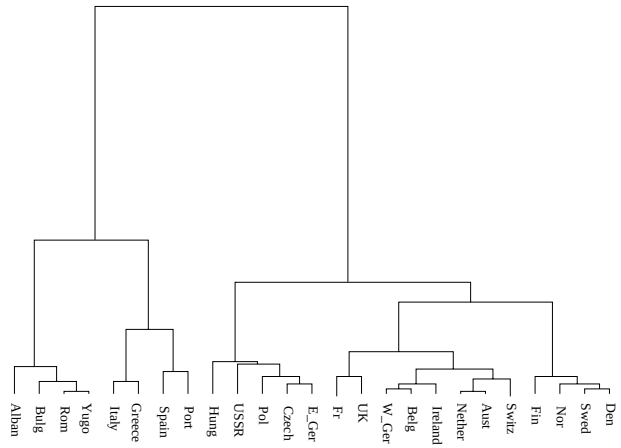

FIG. 1 – *Dendrogramme obtenu avec WARD pour les données proteines*

On obtient bien grâce à ce dendrogramme des classes de pays proches en terme de consommation en protéines mais si l'on souhaite avoir une interprétation facile de ces classes, une étape supplémentaire est alors nécessaire.

La méthode DIVCLUS-T a été appliquée au même jeu de données et le dendrogramme de la hiérarchie est représenté à la Figure 2.

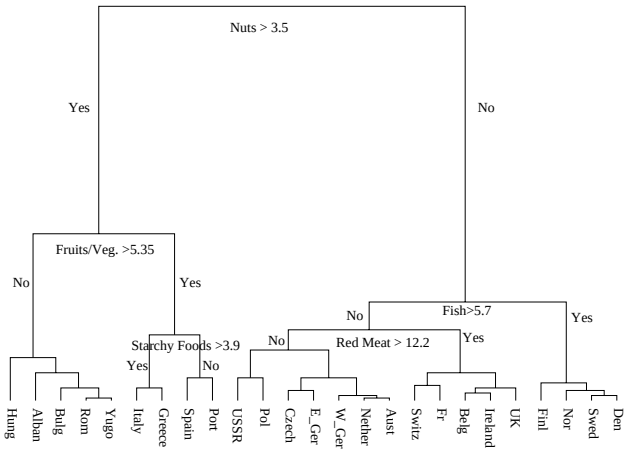

FIG. 2 – *Dendrogramme obtenu avec DIVCLUS-T pour les données protéines*

On note que l'on retrouve avec DIVCLUS-T la classe des pays nord européens {Fin, Nor, Swed, Den} déjà trouvée par Ward mais avec en plus une interprétation naturelle de cette

classe : [Nuts >3.5] et [Fish > 5.7]. Imposer ce type de caractérisation monothétique des classes dont on cherche par ailleurs à minimiser l'inertie devrait impliquer pour les partitions issues de DIVCLUS-T des pourcentages d'inertie expliquée moins bons que pour les partitions issues de WARD. Sur cet exemple nous avons donc avons calculé les pourcentages d'inertie expliquée des partitions de 2 à 10 classes des deux dendrogrammes.

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

TAB. 1 – *Pourcentages d'inerties expliquées des partitions de 2 à 10 classes*

On note dans le Tableau 1 que les partitions de DIVCLUS-T sont meilleures que celles de WARD de 2 à 4 classes, puis WARD prend le dessus jusqu'à 7 classes et enfin les pourcentages sont identiques à partir de 9 classes (car les partitions sont identiques). La perte d'inertie due à la contrainte sur la description des classes est probablement compensée par le fait que DIVCLUS-T est un algorithme descendant qui trouve les partitions en peu de classes dans ses premières itérations tandis que WARD est ascendant est trouve donc ces mêmes partitions dans ses dernières itérations.

2 La méthode DIVCLUS-T

La méthode de classification monothetique DIVCLUS-T a d'abord été proposée dans le cadre plus général de l'Analyse des Données Symbolique (Chavent (1997)) et avait alors été implémentée sous le nom DIV dans le logiciel SODAS (Bock et Diday (2000)). Elle avait également été présentée de manière succincte dans Chavent (1998) pour des données quantitatives et dans Chavent et al. (1999) pour des données qualitatives. Dans Chavent et al. (1999) la méthode, appelée DIVOP à l'époque, était présentée dans le cadre d'une application en dermatologie conjointement avec une autre méthode divisive monothétique appelée DIVAF, basée sur l'analyse des correspondances multiples. Une méthode divisive de type monothétique utilisant le processus de Poisson a également été proposée par Pircon (2004). Récemment, une méthode de classification divisive proche de DIVCLUS-T a été implémentée dans la dernière version du logiciel SPAD sous le nom ICT pour Interactive Clustering Tree (Rakotomalala et LeNouvel (2006)).

Ici, le nom DIVCLUS-T a été choisi comme acronyme de DIVisive CLUstering Tree. Cette méthode procède comme toute méthode descendante hiérarchique par divisions successives et s'articule autour des 3 points suivants :

- Les divisions s'arrêtent après k étapes. On obtient donc le "haut" du dendrogramme c'est à dire les partitions de 2 à $k+1$ classes.
- A chaque étape cette méthode choisit de diviser la classe telle que la nouvelle partition ainsi obtenue soit d'inertie intra-classe minimum. Pour des données qualitatives, l'inertie est calculée avec la distance du χ^2 sur le tableau disjonctif complet. Le critère d'inertie intra-classe étant additif cela revient à choisir la classe telle que la variation de l'inertie obtenue en la divisant soit maximum. Dans WARD on agrège à chaque étape les

deux classes minimisant ce même critère de variation de l'inertie. Dans WARD et dans DIVCLUS-T on utilise donc le même critère pour indicer la hiérarchie et donc évaluer la hauteur des paliers dans le dendrogramme. Dans DIVCLUS-T le choix de la classe à diviser est nécessaire puisque l'on ne continue pas nécessairement les divisions jusqu'à l'obtention des singletons.

– L'algorithme de bi-partitionnement d'une classe à n éléments en deux sous-classes n'évalue pas l'inertie intra-classe des $2^{n-1} - 1$ bi-partitions possibles pour en retenir la meilleure, mais évalue ce critère sur l'ensemble de toutes les bi-partitions induites par l'ensemble de toutes les questions binaires. On utilise donc ici l'approche monothétique des arbres de décisions et de régression (Morgan et Sonquist (1963), Breiman et al. (1984)) mais dans un cadre non supervisé. Les différences sont nombreuses. En particulier il n'y a pas de variable à expliquer et pas d'élagage.

Pour des données quantitatives la méthode DIVCLUS-T est en $o(Kpn(log(n) + p))$ où K est le nombre de classes de la partition la plus fine, n est le nombre d'objets et p le nombre de variables. La méthode WARD est en $o(pn^2)$. DIVCLUS-T est donc plus performant pour des petites valeurs de K. La méthode des k-means quant à elle est en $o(KpnT)$ où T est le nombre maximum d'itérations. DIVCLUS-T reste meilleure seulement dans le cas particulier où $log(n) + p < T$.

Pour des données qualitatives on a, comme pour les arbres de décision, un problème de complexité lorsque le nombre de modalités est trop grand. En effet, le nombre de partitions en deux classes d'un ensemble de m modalités augmente exponentiellement avec m. Deux approches sont alors possibles :

– construire une hiérarchie des modalités (en représentant chaque modalité par la descrption moyenne des objets qui la possèdent et en pondérant cet objet moyen par l'effectif de la modalité). On retient alors une partition en $k < m$ classes, k étant suffisamment petit pour que l'on puisse parcourir toutes les partitions en deux classes de ces k groupes de modalités
– définir q ordres sur les m modalités en utilisant l'ordre des modalités sur les q composantes principales issues de l'Analyse Factorielle des Correspondances Multiples.

3 Evaluation sur six bases de l'UCI

Nous avons voulu répondre à la question suivante : l'aspect rigide et simpliste du processus monothétique de DIVCLUS-T implique-t-il des partitions beaucoup moins bonnes en terme d'inertie intra-classe ? Pour donner un premier élément de réponse à cette question, nous avons comparé empiriquement le pourcentage d'inertie expliquée des partitions de 2 à 15 classes obtenues avec DIVCLUS-T, WARD et les k-means, sur 6 jeux de données de l'UCI Machine Learning repository (Hettich et al. (1998)). Ces six bases, trois quantitatives et trois qualitatives, sont décrites dans le tableau 2.

Le tableau 3 donne les résultats pour les trois bases quantitatives et les trois méthodes (colonne DIV pour DIVCLUS-T, colonne WARD pour WARD, colonne W+km pour les centres mobiles sur la partition de WARD et la colonne km pour les centres mobiles en conservant le meilleure solution de 100 initialisations au hasard). Pour les données GLASS, on note que DIVCLUS-T est parfois meilleur que WARD (pour 4 classes), parfois moins bon (pour 2, 3 classes et de 12 à 15 classes) ou encore parfois équivalent (de 5 à 11 classes). Pour les don-

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

TAB. 2 – *Description des 6 jeux de données*

nées PIMA, DIVCLUS-T est meilleur ou équivalent à WARD jusqu'à 4 classes puis WARD prend le dessus à partir de 5 classes. Pour les données ABALONE qui est la plus grande base (4177 objets), DIVCLUS-T est meilleur que WARD en 2 et 4 classes (WARD est meilleur pour trois classes) et fournit des résultats proches ensuite. Finalement sur ces trois jeux de données DIVCLUS-T semble plus performant en terme d'inertie expliquée pour les partitions en peu de classes (ce qui n'est pas surprenant puisque DIVCLUS-T descend et que WARD monte) et pour les bases plus volumineuses (ce qui n'est pas non plus surprenant puisque lorsque le nombre d'objets augmente, le nombre de bi-partitions évaluées à chaque étape augmente également).

	Glass				Pima				Abalone			
K	DIV	WARD	W+km	km	DIV	WARD	W+km	km	DIV	WARD	W+km	km
2	21.5	22.5	22.8	22.8	14.8	13.3	16.4	16.5	60.2	57.7	60.9	60.9
3	33.6	34.1	34.4	35.0	23.2	21.6	24.5	29.0	72.5	74.8	76.0	76.0
4	45.2	43.3	46.6	46.6	29.4	29.4	36.2	36.2	81.7	80.0	82.5	82.6
5	53.4	53.0	54.8	54.7	34.6	34.9	40.9	40.9	84.2	85.0	86.0	86.1
6	58.2	58.4	60.0	60.7	38.2	40.0	45.3	45.3	86.3	86.8	87.8	87.9
7	63.1	63.5	65.7	65.7	40.9	44.4	48.8	48.9	88.3	88.4	89.6	89.6
8	66.3	66.8	68.9	68.2	43.2	47.0	51.1	51.2	89.8	89.9	90.7	90.9
9	69.2	69.2	71.6	70.5	45.2	49.1	52.4	53.2	91.0	90.9	91.7	91.8
10	71.4	71.5	73.9	72.4	47.2	50.7	54.1	55.1	91.7	91.6	92.4	92.4
11	73.2	73.8	75.6	74.7	48.8	52.4	56.0	56.7	92.0	92.1	92.8	92.8
12	74.7	76.0	77.0	76.6	50.4	53.9	58.0	58.4	92.3	92.4	93.0	93.1
13	76.2	77.6	78.7	77.2	52.0	55.2	58.8	59.7	92.6	92.7	93.3	93.4
14	77.4	79.1	80.2	78.2	53.4	56.5	60.0	61.1	92.8	93.0	93.7	93.7
15	78.5	80.4	81.0	79.3	54.6	57.7	61.0	62.1	93.0	93.2	93.9	93.9

TAB. 3 – *Données quantitatives*

Pour les trois bases qualitatives (tableau 4) on obtient le même type de résultats. Pour les données Solar Flare et CMC, DIVCLUS-T est meilleur que WARD jusqu'à respectivement 10 et 8 classes. Pour les données Zoo, DIVCLUS-T reste toujours en dessous de WARD. C'est peut-être du aux fait que les variables sont binaires et que pour des données qualitatives, le nombre de bi-partitions évaluées à chaque étape augmente avec le nombre de modalités.

3.1 Echantillonages

Afin de mieux évaluer ces résultats nous avons créé pour chacune des trois bases quantitatives d'UCI (Glass, Pima et Abalone) 100 échantillons de taille 150 pour Glass, 500 pour

	Zoo			Solar Flare			CMC		
K	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

TAB. 4 – *Données qualitatives*

Pima et pour Abalone. Sur chacun de ces échantillons nous avons appliqué les quatre méthodes de classification utilisées sur la base complète. Les premières colonnes des tableaux 5 et 6 donnent la moyenne des écarts entre la meilleure solution et la solution obtenue par la méthode. Par exemple dans la colonne DIV nous avons la moyenne des écarts entre DIV et la meilleure solution. Cette moyenne est souvent égale à zéro pour la méthode des centres mobiles (km) ce qui montre que cette méthode est presque toujours la meilleure. Ceci est aussi vrai pour la stratégie WARD+km sur la base Abalone mais sur la base Pima l'écart est un peu différent de zéro. Avec la base Glass la stratégie km est la meilleure quand le nombre de classes est inférieur à 7, la stratégie WARD+km devient la meilleure stratégie quand le nombre de classes est supérieur à 7.

Les dernières colonnes donnent, en pourcentage, le nombre de fois où la solution obtenue par DIV est meilleure que la solution obtenue par WARD. Pour les bases Abalone et Glass la méthode DIV est meilleure que la méthode WARD lorsque le nombre de classes est petit, on observe cela aussi pour la base Pima mais uniquement lorsque le nombre de classes est égal à 2. Cet indicateur montre assez clairement que pour un nombre de classes assez petit la méthode DIV est plus efficace que la méthode WARD.

	Glass					Pima				
K	DIV	WARD	W+km	km	Pourc	DIV	WARD	W+km	km	Pourc
2	1.4	1.0	0.1	0.0	36.0	1.9	3.2	0.3	0.0	91.0
3	1.4	1.5	0.3	0.0	56.0	5.7	5.2	0.2	0.0	36.0
4	1.7	2.1	0.3	0.2	59.0	6.7	5.2	0.3	0.0	15.0
5	1.9	2.4	0.3	0.0	57.0	6.5	4.7	0.2	0.0	5.0
6	2.4	2.5	0.3	0.1	48.0	7.0	4.6	0.3	0.0	1.0
7	2.8	2.6	0.4	0.1	41.0	7.7	4.7	0.4	0.0	0.0
8	2.4	2.1	0.3	0.4	42.0	7.6	4.3	0.4	0.0	0.0
9	2.1	1.7	0.1	0.7	34.0	7.5	4.0	0.4	0.0	0.0
10	2.0	1.4	0.1	0.9	25.0	7.4	3.9	0.4	0.0	0.0
11	1.9	1.1	0.0	1.0	11.0	7.3	3.7	0.3	0.0	0.0
12	1.9	0.9	0.0	1.2	7.0	7.2	3.5	0.3	0.0	0.0
13	1.8	0.7	0.0	1.2	5.0	7.2	3.4	0.2	0.0	0.0
14	1.7	0.6	0.0	1.2	3.0	7.0	3.3	0.2	0.1	0.0
15	1.7	0.5	0.0	1.5	2.0	6.9	3.1	0.2	0.1	0.0

TAB. 5 – *Simulation sur les données Glass et Pima*

	Abalone				
K	DIV	WARD	W+km	km	Pourc
2	0.4	3.3	0.0	0.0	71.0
3	3.5	2.2	0.0	0.0	21.0
4	0.8	1.4	0.0	0.0	83.0
5	1.3	1.0	0.0	0.0	37.0
6	1.2	0.9	0.1	0.0	22.0
7	1.0	0.8	0.1	0.0	28.0
8	0.6	0.7	0.1	0.0	65.0
9	0.6	0.6	0.0	0.0	50.0
10	0.6	0.5	0.0	0.0	25.0
11	0.7	0.5	0.0	0.0	12.0
12	0.7	0.4	0.0	0.0	7.0
13	0.7	0.4	0.0	0.0	3.0
14	0.7	0.4	0.0	0.0	1.0
15	0.7	0.4	0.0	0.1	1.0

TAB. 6 – *Simulation sur les données Abalone*

4 Conclusion

DIVCLUS-T est une méthode monothétique qui a l'avantage par rapport aux méthodes polythétiques telles que WARD et les k-means, de donner une interprétation très simple et immédiate des classes et un arbre hiérarchique facile a lire et à comprendre par l'utilisateurs. Il est en outre normal que cette contrainte sur l'interprétation des classes, imposée dans le processus de classification, implique une perte de qualité au niveau du critère d'inertie. Il n'est donc pas surprenant que DIVCLUS-T soit généralement moins performant que WARD ou les k-means. En revanche, nous avons noté sur les exemples des 6 bases de l'UCI que le comportement de DIVCLUS-T reste tout à fait raisonnable en terme d'inertie, surtout pour les partitions en peu de classes.

Lorsqu'un utilisateur veut obtenir une partition en un nombre de classes relativement important, par exemple pour réduire le nombre d'objets, WARD et les centres mobiles sont certainement plus performants que DIVCLUS-T. Mais lorsque l'utilisateur s'intéresse aux partitions en peu de classes et à leur interprétation, alors DIVCLUS-T semble être une alternative intéressante aux méthodes classiques.

Références

Bock, H.-H. et E. Diday (Eds.) (2000). *Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data*. Studies in classification, data analysis and knowledge organisation. Heidelberg : Springer Verlag.

Breiman, L., J. H. Friedman, R. A. Olshen, et C. J. Stone (1984). *Classification And Regression Trees*. New York: Chapman and Hall.

Chavent, M. (1997). *Analyse des données symboliques. Une méthode divisive de classification*. Ph. D. thesis, Université Paris-IX Dauphine.

Chavent, M. (1998). A monothetic clustering method. *Pattern Recognition Letters 19*, 989–996.

Chavent, M., C.Guinot, Y. Lechevallier, et M. Tenehaus (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.

Hand, D., F. Daly, K. McConway, D. Lunn, et E. O. (eds.) (1994). *A Handbook of Small Data Sets*. CHAPMAN & HALL.

Hettich, S., C. L. Blake, et C. J. Merz (1998). *UCI Repository of machine learning databases, http://www.ics.uci.edu/ mlearn/MLRepository.html*. CA: University of California, Department of Information and Computer Science: Irvine.

Morgan, J. et J. Sonquist (1963). Problems in the analysis of survey data, and proposal. *J. Aler. Statist. Assoc. 58*, 415–434.

Pircon, Y. (2004). *La classification et les processus de Poisson pour de nouvelles méthodes de partitionnement*. Thèse de doctorat, Facultés Universitaires Notre-Dame de la Paix.

Rakotomalala, R. et T. LeNouvel (2006). Interactive clustering tree–une méthode de classification descendante adaptée aux grands ensembles de données. *RNTI à paraître*, 87–99.

Summary

The main aim of this work is to evaluate the loss of information in term of inertia between partitioning or hierarchical clustering methods and a conceptual clustering approach. We want to answer the following question: does the constraints of the monothetic process in a conceptual clustering method leads to partitions of worse quality i.e. of bigger within cluster inertia ? We propose to carry out this experiment on 6 bases of the UCI, three of these bases are tables are quantitative data tables, the three others are qualitative.