# Évaluation d'une approche de classification conceptuelle

*Marie Chavent – Yves Lechevallier*

Mathématiques Appliquées de Bordeaux, UMR 5466 CNRS

Université Bordeaux 1 - 351, Cours de la libération

33405 Talence Cedex, France

chavent@math.u-bordeaux1.fr

INRIA- Institut National de Recherche en Informatique et en Automatique

Domaine de Voluceau- Rocquencourt B.P. 105

78153 Le Chesnay Cedex, France

Yves.Lechevallier@inria.fr

# Introduction

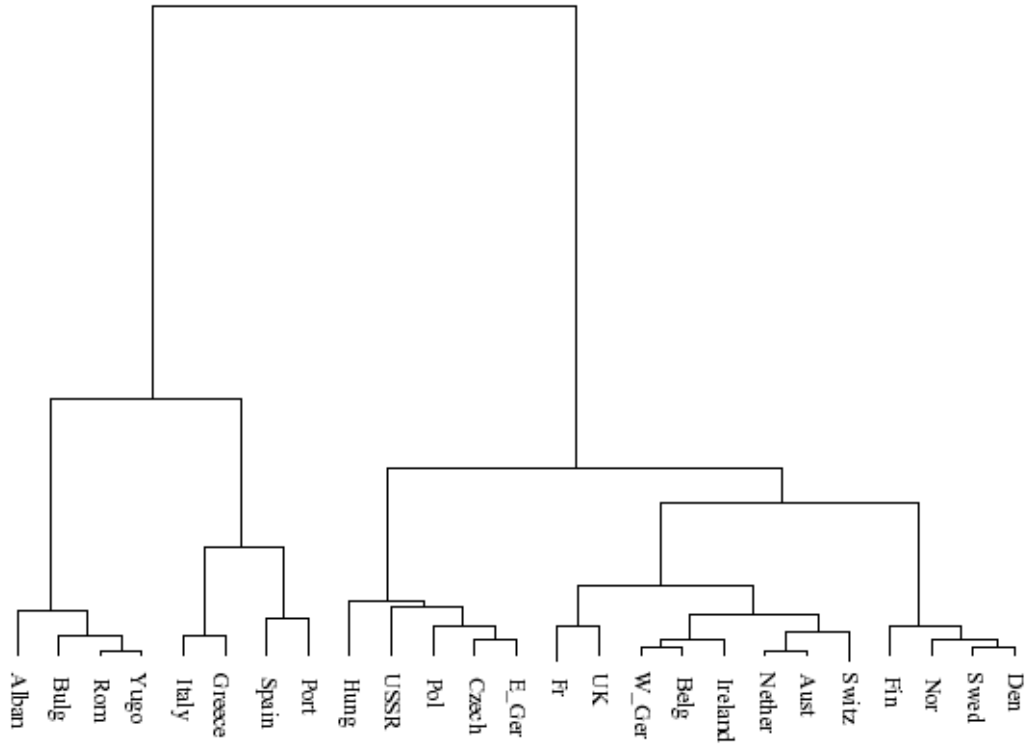
- L'objectif est d'évaluer la perte d'information au sens de l'inertie entre des méthodes de classification et notre approche de classification conceptuelle.
- Nous voulons répondre à la question suivante : l'aspect simpliste du processus monothétique d'une méthode conceptuelle implique t'il des partitions de moins bonne qualité au sens du critère de l'inertie ?
- Nous proposons de réaliser cette expérience sur 6 bases de l'UCI

# Un exemple quantitatif : les « protein data »

| Country | Red Meat | White Meat | Eggs | Milk | Fish | Starchy Cereals | Foods | Nuts | Fruit/Veg. |
|---|---|---|---|---|---|---|---|---|---|
| Alban | 10.1 | 1.4 | 0.5 | 8.9 | 0.2 | 42.3 | 0.6 | 5.5 | 1.7 |
| Aust | 8.9 | 14 | 4.3 | 19.9 | 2.1 | 28 | 3.6 | 1.3 | 4.3 |
| Belg | 13.5 | 9.3 | 4.1 | 17.5 | 4.5 | 26.6 | 5.7 | 2.1 | 4 |
| Bulg | 7.8 | 6 | 1.6 | 8.3 | 1.2 | 56.7 | 1.1 | 3.7 | 4.2 |
| Czech | 9.7 | 11.4 | 2.8 | 12.5 | 2 | 34.3 | 5 | 1.1 | 4 |
| Den | 10.6 | 10.8 | 3.7 | 25 | 9.9 | 21.9 | 4.8 | 0.7 | 2.4 |
| Finl | 9.5 | 4.9 | 2.7 | 33.7 | 5.8 | 26.3 | 5.1 | 1 | 1.4 |
| E-Ger | 8.4 | 11.6 | 3.7 | 11.1 | 5.4 | 24.6 | 6.5 | 0.8 | 3.6 |
| Fr | 18 | 9.9 | 3.3 | 19.5 | 5.7 | 28.1 | 4.8 | 2.4 | 6.5 |
| Greece | 10.2 | 3 | 2.8 | 17.6 | 5.9 | 41.7 | 2.2 | 7.8 | 6.5 |
| Hung | 5.3 | 12.4 | 2.9 | 9.7 | 0.3 | 40.1 | 4 | 5.4 | 4.2 |
| Ireland | 13.9 | 10 | 4.7 | 25.8 | 2.2 | 24 | 6.2 | 1.6 | 2.9 |
| Italy | 9 | 5.1 | 2.9 | 13.7 | 3.4 | 36.8 | 2.1 | 4.3 | 6.7 |
| Nether | 9.5 | 13.6 | 3.6 | 23.4 | 2.5 | 22.4 | 4.2 | 1.8 | 3.7 |
| Nor | 9.4 | 4.7 | 2.7 | 23.3 | 9.7 | 23 | 4.6 | 1.6 | 2.7 |
| Pol | 6.9 | 10.2 | 2.7 | 19.3 | 3 | 36.1 | 5.9 | 2 | 6.6 |
| Port | 6.2 | 3.7 | 1.1 | 4.9 | 14.2 | 27 | 5.9 | 4.7 | 7.9 |
| Rom | 6.2 | 6.3 | 1.5 | 11.1 | 1 | 49.6 | 3.1 | 5.3 | 2.8 |
| Spain | 7.1 | 3.4 | 3.1 | 8.6 | 7 | 29.2 | 5.7 | 5.9 | 7.2 |
| Swed | 9.9 | 7.8 | 3.5 | 24.7 | 7.5 | 19.5 | 3.7 | 1.4 | 2 |
| Switz | 13.1 | 10.1 | 3.1 | 23.8 | 2.3 | 25.6 | 2.8 | 2.4 | 4.9 |
| UK | 17.4 | 5.7 | 4.7 | 20.6 | 4.3 | 24.3 | 4.7 | 3.4 | 3.3 |
| USSR | 9.3 | 4.6 | 2.1 | 16.6 | 3 | 43.6 | 6.4 | 3.4 | 2.9 |
| W-Ger | 11.4 | 12.5 | 4.1 | 18.8 | 3.4 | 18.6 | 5.2 | 1.5 | 3.8 |
| Yugo | 4.4 | 5 | 1.2 | 9.5 | 0.6 | 55.9 | 3 | 5.7 | 3.2 |

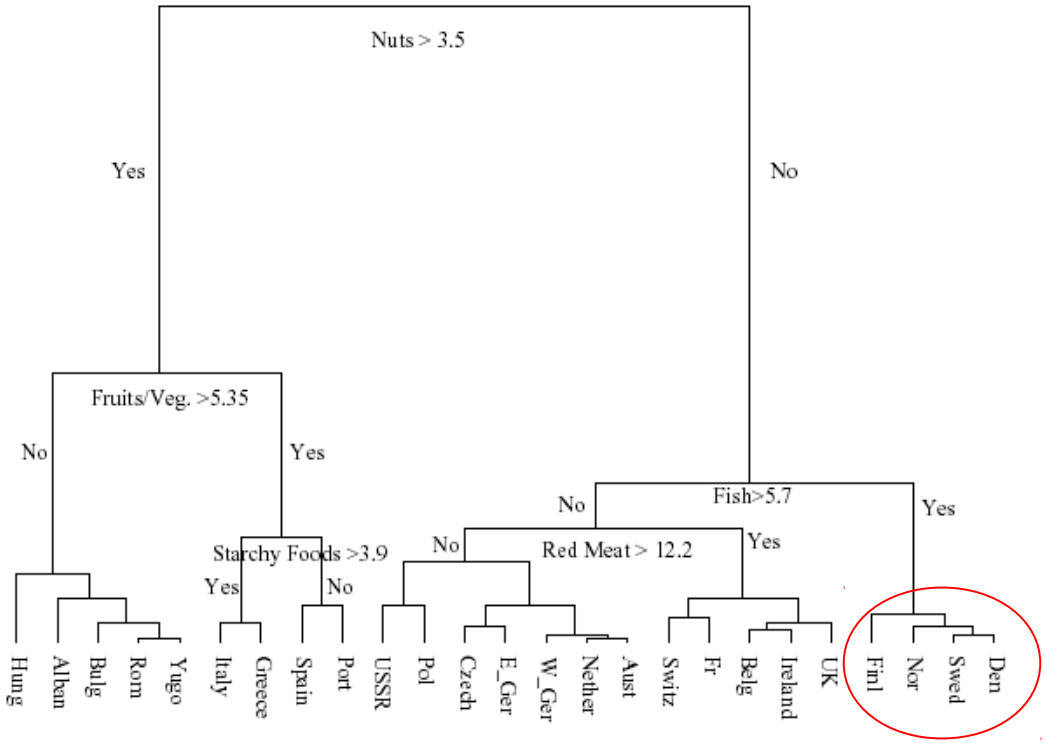
*A Handbook of Small Data Sets*, Hand, D.J. et al. (eds.) (1994)

# Dendrogramme obtenu avec Ward




Le dendrogramme ne donne pas d'indications sur l'interprétation des classes. Par exemple : {Fin, Nor, Swed, Den}

# Dendrogramme obtenu avec DIVCLUS-T




Avantage: Les classes ont une interprétation naturelle. Par exemple les objets la classe {Fin, Nor, Swed, Den} vérifient les propriétés [Nuts < 3,5] et [Fish > 5,7]

# Évaluation de la qualité des ces méthodes

Question: Les partitions obtenues avec WARD et DIV ont-elles des pourcentages d'inertie expliquée très différents ?

| 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 |

*Pourcentages d'inertie expliquée des partitions issues des dendrogrammes de WARD et de DIV .*

# DIVCLUS-T : algorithme divisif, méthode conceptuelle

Cette méthode (Chavent 1997, 1998) divise à chaque étape une classe en fonction d'une question binaire et du critère d'inertie.

A chaque étape, la méthode définit la question binaire qui minimise sur l'ensemble des questions binaires admissibles le critère de l'inertie intra-classe sur la bi-partition induite.

A chaque classe obtenue on associe une règle d'affectation ou un critère d'appartenance à partir de propriétés

# Questions binaires admissibles

variable continue

$[X > 3.5]$ ?

Variable qualitative

$[X \in \{m_1,...,m_h\}]$?

•Dans le cas d'une variable continue on évalue toutes coupures possibles c'est-à-dire au maximum *n-1* coupures.

• Pour une variable qualitative ordonnée *Y*, on évalue ainsi au maximum *m*-1 bipartitions

•Dans le cas d'une variable qualitative non ordonnée, on se heurte vite à un problème de complexité, le nombre de dichotomies du domaine d'observation étant alors égal à $2^{m-1}$-1.

# Critère d'évaluation

Soit $P_K=(P_1,\ldots,P_K)$ une partition en K classes

Critère d'évaluation $W(P_K)$ doit être additif $W(P_K)=\sum_{P_k \in P_K} w(P_k)$

Passage de la partition $P_K$ à la partition $P_{K+1}$ en divisant la classe C en deux classes $C_1$ et $C_2$

$$W(P_{K+1})=W(P_K)-w(C)+w(C_1)+w(C_2)$$

Calcul de l'écart

$$W(P_{K+1})-W(P_K)=-w(C)+w(C_1)+w(C_2)$$

# Critère d'évaluation

Le critère de l'inertie intra-classe est additif

La réduction du critère d'évaluation revient à maximiser le gain $\Delta(Q)$ associé à la question binaire $Q^*$ qui découpe la classe $C$ et deux classes $C_1$ et $C_2$

$$\Delta\left(Q^*\right) = \max_{Q \in B}\left(w(C) - w\left(C_1\right) - w\left(C_2\right)\right)$$

$B$ étant l'ensemble des questions binaires admissibles

# Algorithme divisif (récursif)

initialisation: $P_1$ partition en une classe

Étape t+1: Pour chaque classe C de la partition en k classes sélectionner la meilleure bi-partition ($C_1$,$C_2$) de C en fonction du critère de l'inertie intra-classes.

1) pour chaque variable $X$ , trouver la coupure $s$ qui maximise

$$\Delta(X,s/C)=|w(C) - w(C_1)-w(C_2)|$$

2) choisir la variable $X^*$ et la coupure $s^*$

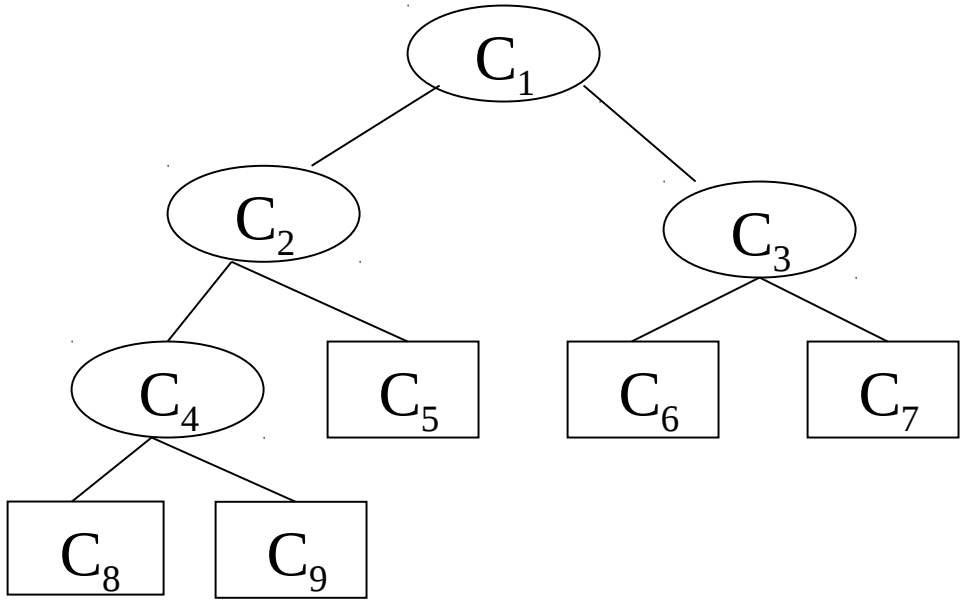
$$\Delta(X^*,s^*/C)=max\ \Delta(X,s/C)$$

3) diviser la classe C en ($C_1$,$C_2$) et construire la partition en k+1 classes

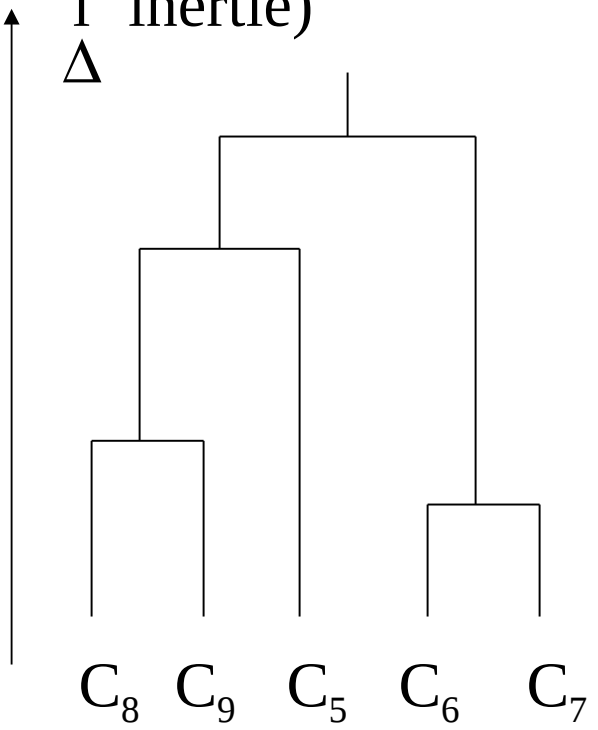
Arrêter quand t+1 est égal à K

# Arbre de décision/ Hiérarchie indicée

Pas d'ordre de découpage




Ordre de construction donnée par l'indice de WARD (variation de l 'inertie)



# Exemple qualitatif

Ce tableau de données comprend 27 races de chiens décrites par 7 variables qualitatives (Saporta, 1990).

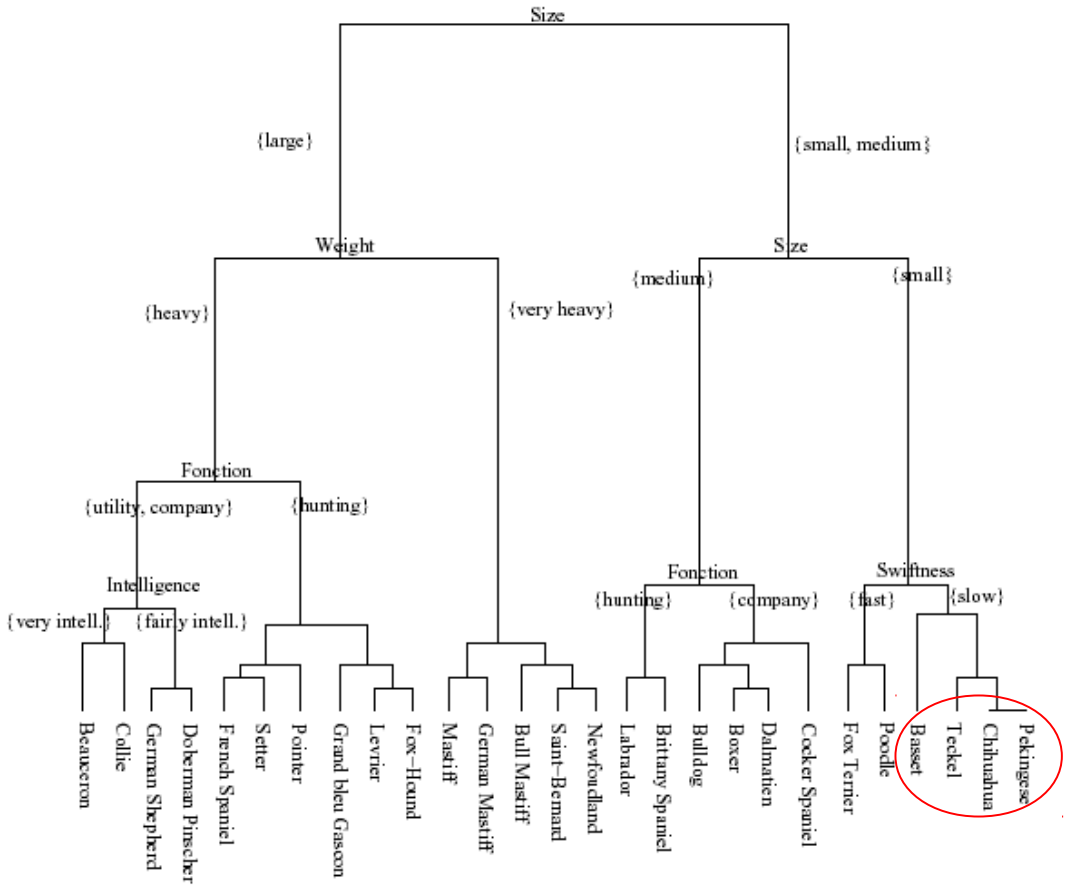
*Size* ={small, medium,large} …

2 variables binaires d'où 2 bi-partitions à évaluer et 5 variables à 3 catégories d'où un total de 17 question binaires à évaluer.

La question binaire sur la variable *size* créant la bi-partition avec 15 chiens ayant la modalité « large » et 12 chiens ayant la modalité « small » ou « medium » donne la bi-partition ayant la plus petite inertie intra-classe.

Cette question binaire crée la première segmentation de notre population.

# Exemple qualitatif



# Comparaison empirique

Prix à payé pour cette représentation monothètique des classes ?

Six jeux de données de l'UCI Machine Learning Repository
***http://www.ics.uci.edu/~mlearn/MLRepository.html***

| 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) |

DIVCLUS-T, WARD et les centres mobiles après WARD, trois méthodes basées sur la minimisation de l'inertie intra-classe, ont été appliquées aux six jeux de données. K-means avec 100 initialisations a été utilisé sur les tableaux quantitatifs.
pour les partitions de 2 à 15 classes obtenues avec ces trois méthodes, le pourcentage d'inertie expliquée a été calculé

# Comparaison empirique

Le pourcentage d'inertie expliquée par un partition *P* est

$$E(P) = 100.\frac{(1-W(P))}{T}$$

*E(P)* mesure la part d'inertie expliquée par *P*
Minimiser l'inertie intra-classe est équivalent à maximiser l'inertie expliquée.

*E(P)*= 0 pour la partition en une classe

*E(P)*=100 pour la partition des singletons

***Comparer deux partitions***

La partition *P* est meilleure que la partition *P'* si

*E(P) > E(P')*

# Données quantitatives

| | 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 |

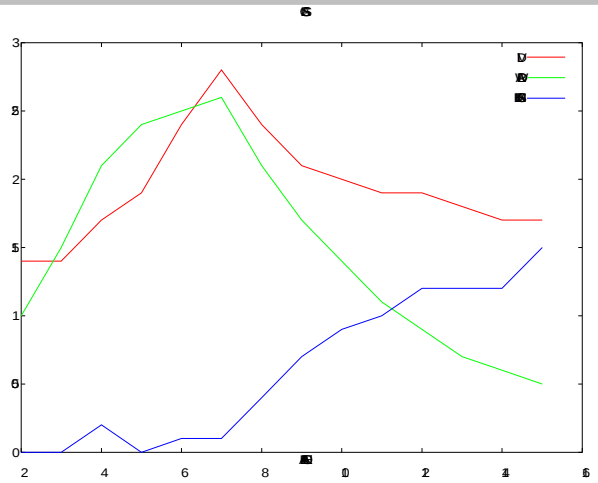
DIVCLUS-T est meilleur que WARD quand le nombre de classes est petit

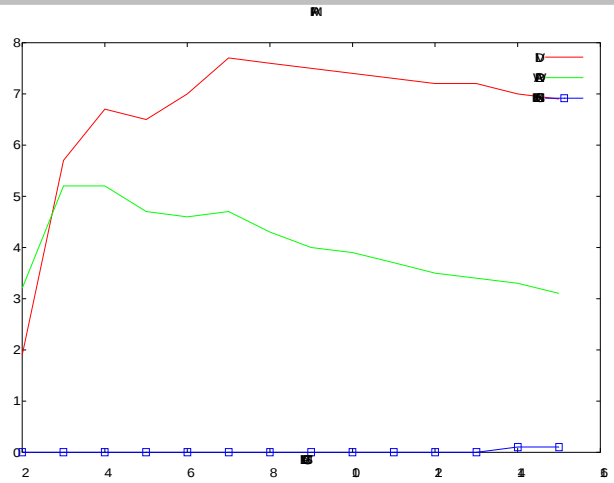
# Données qualitatives

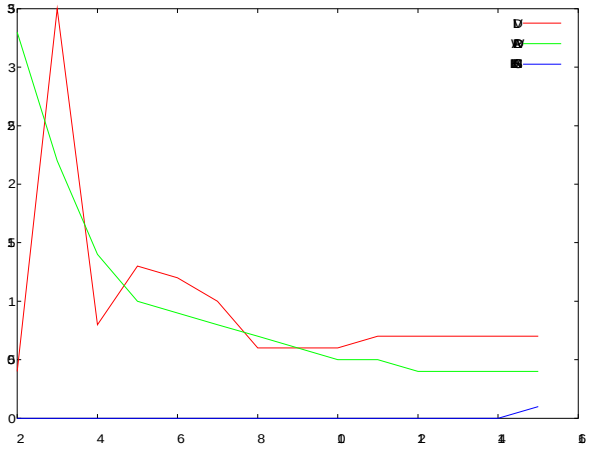
| | Zoo | | | Solar Flare | | | CMC | | |
|---|---|---|---|---|---|---|---|---|---|
| K | DIV | WARD | W+km | DIV | WARD | W+km | DIV | WARD | W+km |
| 2 | 23.7 | 24.7 | 26.2 | 12.7 | 12.6 | 12.7 | 8.4 | 8.2 | 8.5 |
| 3 | 38.2 | 40.8 | 41.8 | 23.8 | 22.4 | 23.8 | 14.0 | 13.1 | 14.8 |
| 4 | 50.1 | 53.7 | 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 |

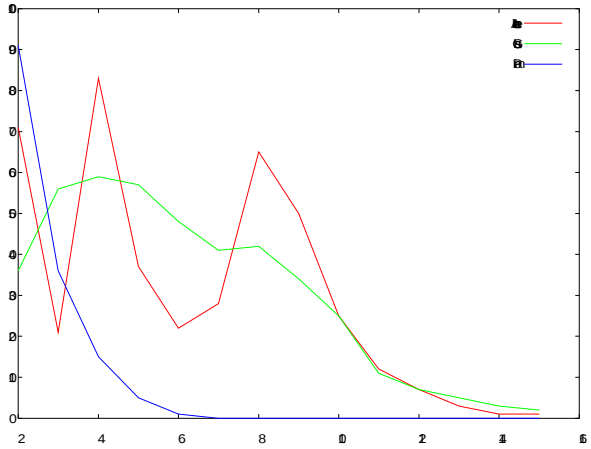
Pour les données zoo DIVCLUS-T est moins que bon WARD.

# Simulation sur les données quantitatives

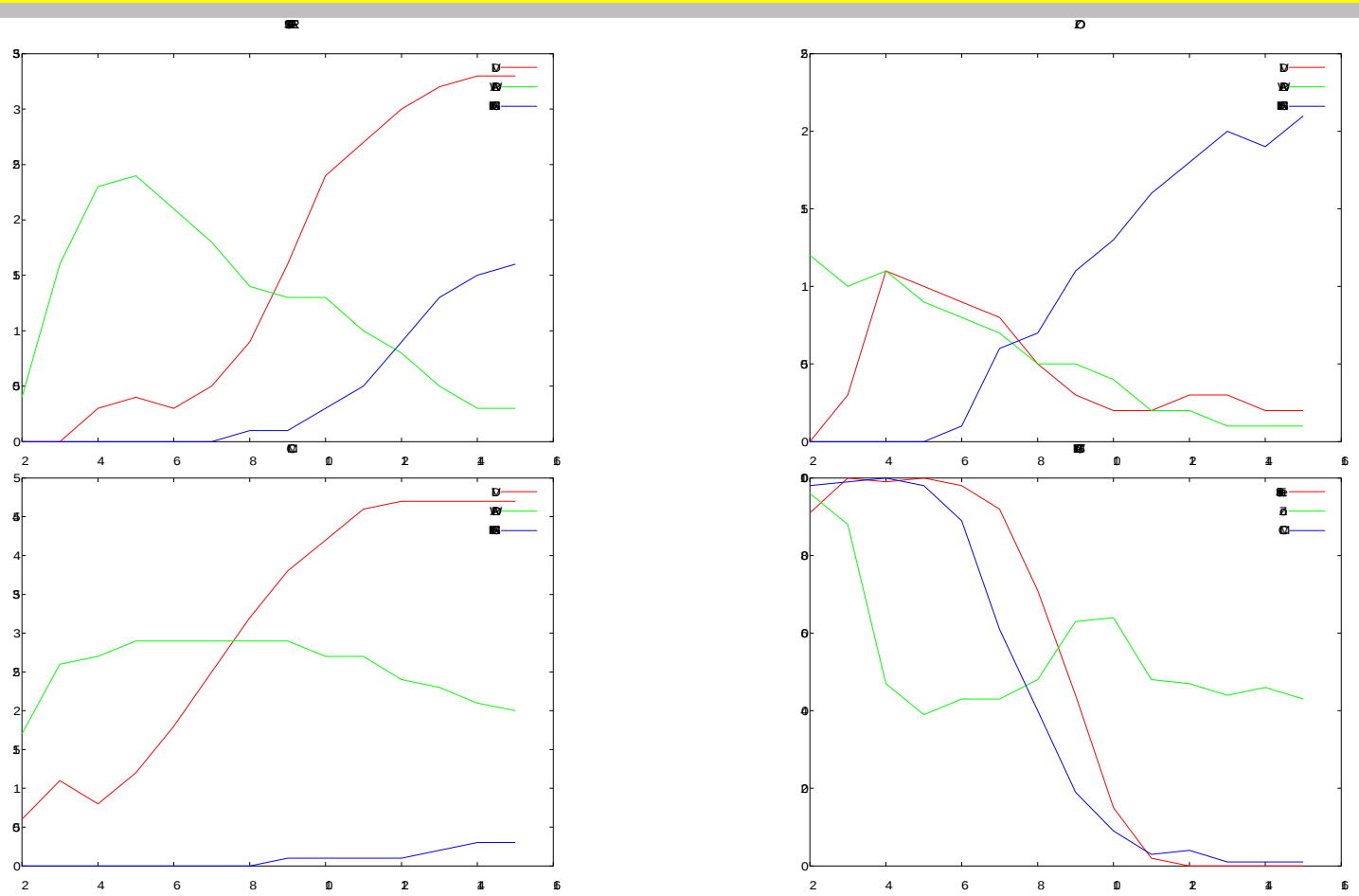








# Simulation sur les données qualitatives



# Conclusion

Comparaison avec WARD et les centres mobiles

Bon comportement de DIVCLUS-T en terme d'inertie sur les 6 jeux de données surtout pour les partitions en un petit nombre de classes.

Complexité

DIVCLUS-T est performant pour des données quantitatives. Dans le cas qualitatif des solutions existent pour définir un ordre sur les modalités et réduire ainsi la complexité

# Conclusion

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