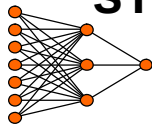


Stéphane Tufféry
Statisticien - Data Miner - Enseignant



DATA MINING - SCORING

STATISTIQUE DÉCISIONNELLE



APPLICATION AU CRM



19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

1

Plan du cours

- Qu'est-ce que le data mining ?
- A quoi sert le data mining ?
- Les 2 grandes familles de techniques
- Le déroulement d'un projet de data mining
- Coûts et gains du data mining
- Facteurs de succès - Erreurs - Consulting
- L'analyse et la préparation des données
- **Techniques descriptives de data mining**
- Techniques prédictives de data mining
- Logiciels de statistique et de data mining
- Informatique décisionnelle et de gestion
- CNIL et limites légales du data mining
- Le text mining
- Le web mining

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

2

Techniques descriptives

Points forts et points faibles

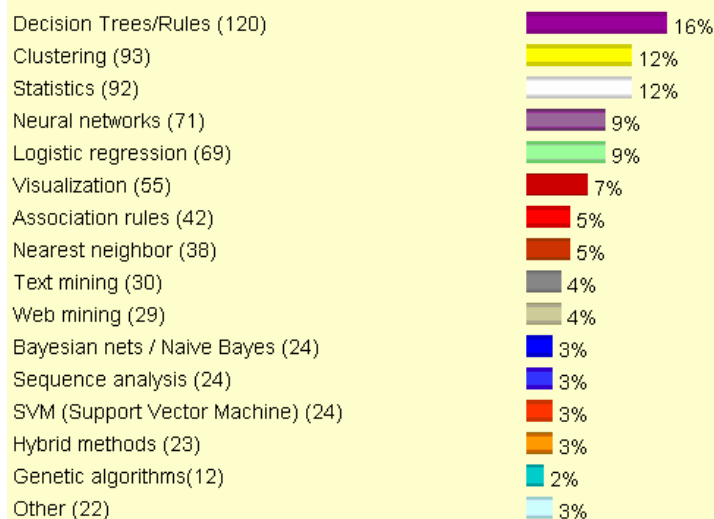
19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

3

Sondage sur www.kdnuggets.com (novembre 2003)

What data mining techniques you use regularly? [768 choices from 213 voters]



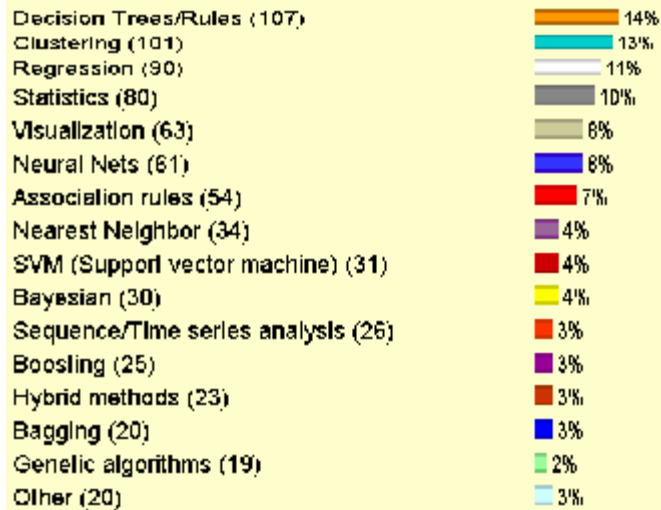
19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

4

Sondage sur www.kdnuggets.com (février 2005)

Data mining/analytic techniques you use frequently: [784 votes total]



19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

5

Techniques descriptives de data mining : La classification

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

6

Qu'est-ce que la classification ?

- Regrouper des objets en groupes, ou classes, ou familles, ou segments, ou clusters, de sorte que :
 - 2 objets d'un même groupe se ressemblent le + possible
 - 2 objets de groupes distincts diffèrent le + possible
 - le nombre des groupes est parfois fixé
- Méthode descriptive :
 - pas de variable cible privilégiée
 - décrire de façon simple une réalité complexe en la résumant
- Utilisation en marketing, médecine, sciences humaines...
- Les objets à classer sont :
 - des individus
 - des variables

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



7

Complexité du problème

- Le nombre de partitions (non recouvrantes) de n objets est le nombre de Bell :
$$B_n = \frac{1}{e} \sum_{k=1}^{\infty} \frac{k^n}{k!}$$
- Exemple : pour n = 4 objets, on a $B_4 = 15$, avec
 - 1 partition à 1 classe (abcd)
 - 7 partitions à 2 classes (ab,cd), (ac,bd), (ad,bc), (a,bcd), (b,acd), (c,bad), (d,abc)
 - 6 partitions à 3 classes (a,b,cd), (a,c,bd), (a,d,bc), (b,c,ad), (b,d,ac), (c,d,ab)
 - 1 partition à 4 classes (a,b,c,d)
- Exemple : pour n = 30 objets, on a $B_{30} = 8,47.10^{23}$
- $B_n > \exp(n) \Rightarrow$ Nécessité de définir des critères de bonne classification et d'avoir des algorithmes performants

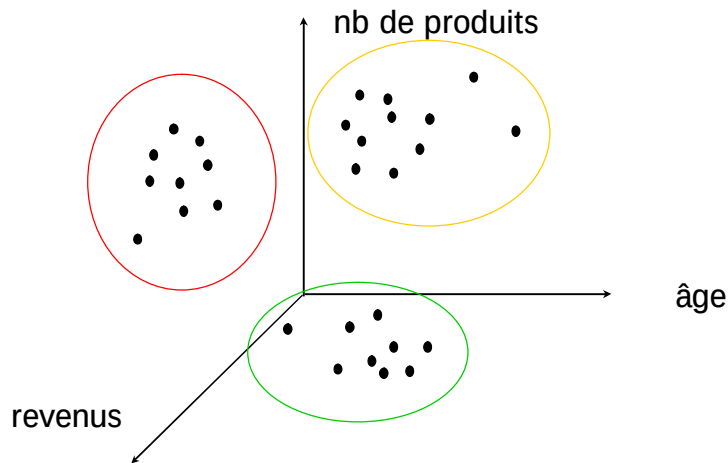
19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



8

La classification



19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

9

Terminologie : de nombreux synonymes

- **Classification**, ou classification automatique, terme généralement employé par les auteurs français
 - attention : il est employé dans un autre sens par les anglo-saxons (qui disent « classification » pour désigner la technique prédictive que les français appellent « classement »)
- **Segmentation** : terme employé en marketing (les « segments de clientèle ») et assez explicite
- **Typologie**, ou **analyse typologique**
- **Clustering** : terme anglo-saxon le plus courant
- **Taxinomie** ou **taxonomie** (biologie, zoologie)
- **Nosologie** (médecine)
- **Reconnaissance de forme non supervisée**
- ...

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

10

Structure des données à classer

- Soit une matrice rectangulaire dont :
 - lignes = individus
 - colonnes = variables
- Cette structure permet de classer individus ou variables
- Soit une matrice carrée de similarités, distances entre :
 - individus
 - ou variables (par exemple : la matrice des corrélations)
- Cette structure permet aussi de classer individus ou variables

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

11

Structure des classes obtenues

- Soit 2 classes sont toujours disjointes : méthodes de partitionnement
 - généralement, le nombre de classes est défini *a priori*
 - certaines méthodes permettent de s'affranchir de cette contrainte (analyse relationnelle, méthodes paramétriques par estimation de densité comme la proc SAS/MODECLUS)
- Soit 2 classes sont disjointes ou l'une contient l'autre : méthodes hiérarchiques
 - ascendantes (agglomératives)
 - descendantes (divisives)
- Soit 2 classes peuvent avoir plusieurs objets en commun (classes « empiétantes » ou « recouvrantes »)
 - analyse « floue », où chaque objet a une certaine probabilité d'appartenir à une classe donnée

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

12

Les différentes méthodes

- Méthodes hiérarchiques
 - ascendantes (agglomératives)
 - basées sur une notion de distance ou de densité
 - descendantes (divisives)
- Méthodes de partitionnement
 - centres mobiles, k-means et nuées dynamiques
 - k-modes, k-prototypes, k-représentants (k-medoids)
 - réseaux de Kohonen
 - méthodes basées sur une notion de densité
 - méthode « de Condorcet » (analyse relationnelle)
- Méthodes mixtes
- Analyse floue (fuzzy clustering)

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

13



Applications de la classification

- Marketing : découper la clientèle en segments dotés chacun d'une offre et d'une communication spécifique
- Commercial : répartir l'ensemble des magasins d'une enseigne en établissements homogènes du point de vue type de clientèle, CA, CA par rayon (selon type d'article), taille du magasin...
- Médical : déterminer des groupes de patients susceptibles d'être soumis à des protocoles thérapeutiques déterminés, chaque groupe regroupant tous les patients réagissant identiquement
- Sociologie : découper la population en groupes homogènes du point de vue sociodémographique, style de vie, opinions, attentes...

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

14



Classification générale de la clientèle

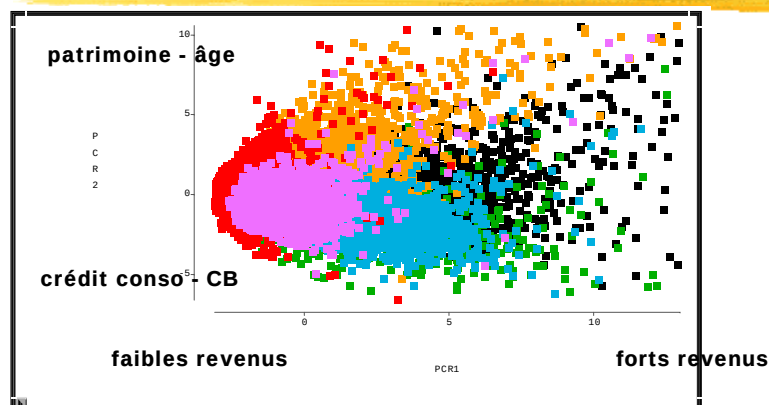
- Détection des profils :
 - clients dont on est la 2^{de} banque
 - clients avec faibles encours mais 1^{ère} banque chez nous
 - clients à tendance « épargne »
 - clients à tendance « crédit consommation »
 - clients à tendance « crédit habitat »
 - ...
- Utilisation pour :
 - les ciblage des actions commerciales
 - l'évaluation du potentiel commercial
 - l'affectation des clients aux différents types de commerciaux

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

15

Représentation des classes



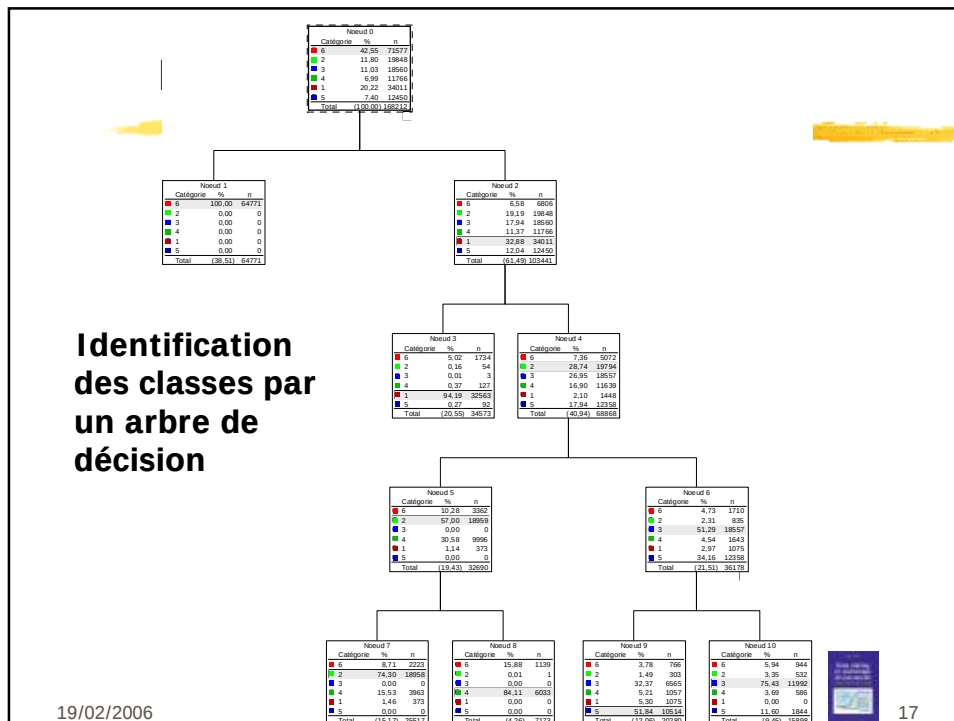
S1 : rouge
S2 : vert
S3 : rose

S4 : orange
S5 : bleu
S6 : noir

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

16



19/02/2006

17

Techniques de classification : La classification d'individus

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

18

Intérêt de la classification pour la prédiction

- Classification
- ⇓
- Homogénéité des groupes d'individus
- ⇓
- Moins de variables discriminantes
 - (éventuellement, valeurs manquantes complétées)
- ⇓
- + grande fiabilité des prédictions

Critères de bonne classification

- Détecter les structures présentes dans les données
- Permettre de déterminer le nombre optimal de classes
- Fournir des classes bien différenciées
- Fournir des classes stables vis-à-vis de légères modifications des données
- Traiter tous les types de variables (continues, ordinales et nominales)
- Traiter efficacement les grands volumes de données

Classification avec des variables continues

- Nécessité de standardiser les variables si elles ne sont pas toutes mesurées dans la même unité et ont des variances dissemblables
- Préférable d'isoler les « outliers » (individus hors-norme)
- Quand on a des variables catégorielles $\Rightarrow$ possibilité de se ramener à une classification de variables continues par une ACM

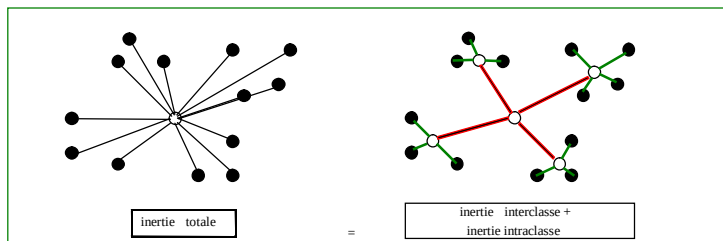
Inertie d'un nuage d'individus

- *Inertie totale* I_{TOT} de la population : moyenne (pondérée par le poids de chaque individu – généralement 1/effectif) des carrés des distances des individus au barycentre
- *Inertie intraclasse* I_A : somme des inerties totales de chaque classe
- *Inertie interclasse* I_R : moyenne (pondérée par la somme des poids de chaque classe) des carrés des distances des barycentres de chaque classe au barycentre global
- *Formule de Huygens* : $I_{TOT} = I_A + I_R$

$$\sum_{i \in I} p_i (x_i - \bar{x})^2 = \sum_{j \in \text{classes}} \left(\sum_{i \in I_j} p_i (x_i - \bar{x}_j)^2 \right) + \sum_{j \in \text{classes}} \left(\sum_{i \in I_j} p_i \right) (\bar{x}_j - \bar{x})^2$$

Inertie et qualité d'une classification

- Une classe est homogène $\Leftrightarrow$ son inertie est faible
- Deux critères de bonne classification : grande I_R , petite I_A
- Ces deux critères sont équivalents d'après la *formule de Huygens* : $I_{TOT} = I_A + I_R$



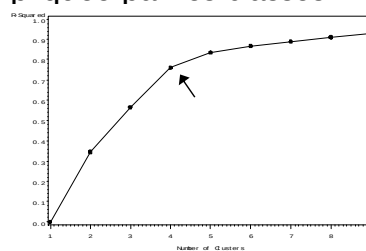
19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

23

Mesures de qualité d'une classification

- R^2 = proportion de la variance expliquée par les classes
 - rapport I_R / I_{TOT}
 - être le plus proche possible de 1 sans avoir trop de classes
 - s'arrêter après le dernier saut important
- Pseudo F = mesure la séparation entre toutes les classes
 - rechercher une grande valeur
 - avec n = nb d'observations
 - et c = nb de classes
 - on a pseudo F =
$$\frac{R^2}{c-1} \bigg/ \frac{1-R^2}{n-c}$$



19/02/2006

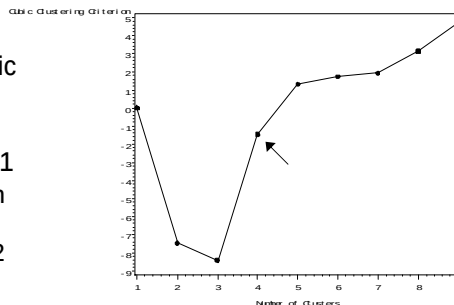
© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

24

Mesures de qualité d'une classification

- Cubic clustering criterion (CCC)
 - > 2 : bonne classification
 - entre 0 et 2 : classification peut-être OK mais à vérifier
 - < 0 : présence d'outliers gênants (surtout si $CCC < -30$)

un creux pour k classes suivi d'un pic pour $k+1$ classes indique une bonne classification en $k+1$ classes (surtout si on a une $\uparrow$ ou une $\downarrow$ douce à partir de $k+2$ classes)



- Ne pas utiliser CCC et pseudo F avec single linkage

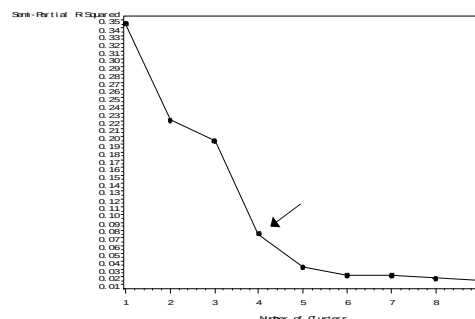
19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

25

Mesures de qualité d'une CAH

- R^2 semi-partiel (SPRSQ) = mesure la perte d'inertie interclasse provoquée en regroupant 2 classes
 - le but étant d'avoir une inertie interclasse maximum, on recherche un faible SPRSQ suivi d'un fort SPRSQ à l'agrégation suivante $\hat{o}$ un pic pour k classes et un creux pour $k+1$ classes indique une bonne classification en $k+1$ classes



19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

26

Classification hiérarchique ascendante

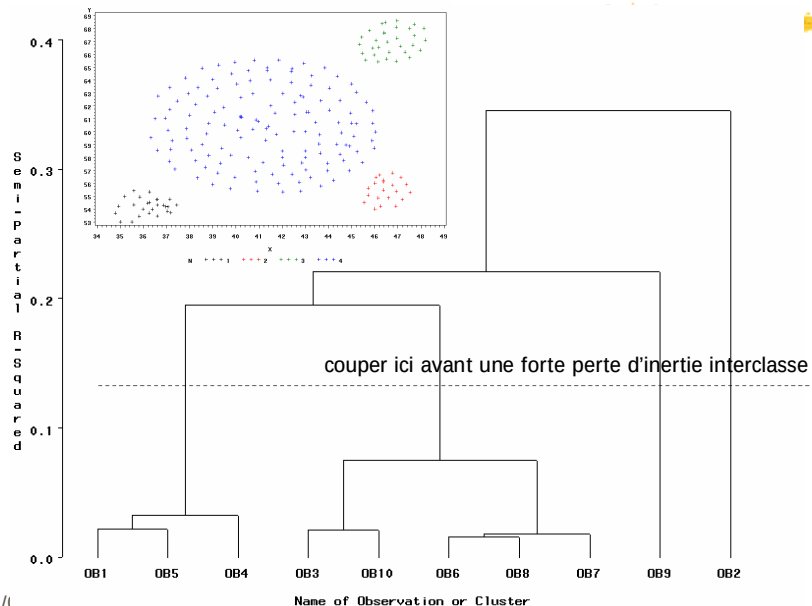
- Utilisables dès que l'on dispose d'une distance : dans un espace des individus ou des variables
- Schéma de l'algorithme :
 - 1) les classes initiales sont les observations
 - 2) on calcule les distances entre classes
 - 3) les 2 classes les plus proches sont fusionnées et remplacées par une seule
 - 4) on reprend en 2) jusqu'à n'avoir plus qu'une seule classe, qui contient toutes les observations
- Résultat sous forme d'un arbre appelé *dendrogramme* :
 - le niveau où l'on coupe l'arbre détermine le nb de classes
 - la hauteur d'une branche est proportionnelle à la perte d'inertie interclasse (R^2 semi-partiel)

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

27

Classification hiérarchique ascendante



19/i

28

Les différentes distances utilisées 1/3

- Distance moyenne entre 2 observations $a \in A$ et $b \in B$
 - tend à produire des classes de variances égales
 - « saut moyen » ou « average linkage »
- Distance entre les barycentres de A et B
 - plus robuste aux outliers mais moins précise
 - « distance des barycentres » ou « centroid method »
- Distance maximale entre 2 observations $a \in A$ et $b \in B$
 - tend à produire des classes de diamètres égaux
 - très sensible aux outliers !
 - peu utilisée
 - « saut maximum » ou « critère du diamètre » ou « complete linkage »

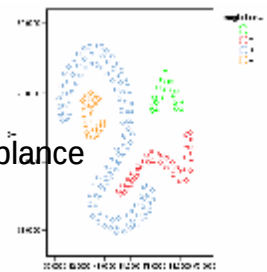
19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

29

Les différentes distances utilisées 2/3

- Distance minimale entre 2 observations $a \in A$ et $b \in B$
 - sait détecter les classes allongées/irrégulières voire sinueuses
 - sensible à « l'effet de chaîne » : 2 classes bien distinctes mais reliées par une série de points isolés vont se retrouver regroupées
 - moins adaptée pour détecter les classes sphériques
 - bonnes propriétés théoriques
 - fréquemment utilisée
 - « saut minimum » ou « single linkage »
- Méthode par maximisation de la vraisemblance
 - peu usitée
 - temps de traitement en n^3



19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

30

Les différentes distances utilisées 3/3

- Critère de Ward (baisse d'inertie interclasse résultant de la fusion des 2 classes)
 - correspond à l'objectif d'avoir la plus forte inertie interclasse possible $\Rightarrow$ avoir la plus faible baisse d'inertie en fusionnant 2 classes
 - la distance de 2 classes A et B, de barycentres a et b, et d'effectifs n_A et n_B vaut :

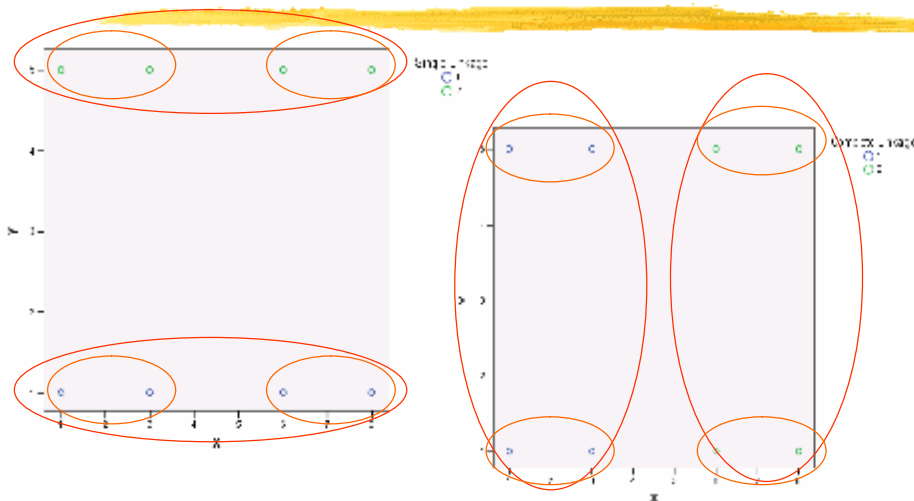
$$d(A, B) = \frac{d(a, b)^2}{\frac{1}{n_A} + \frac{1}{n_B}}$$
 - tend à produire des classes sphériques et de mêmes effectifs – peu efficace sur les classes allongées
 - très sensible aux outliers
 - méthode la plus utilisée (fonctionne bien sur les pbs réels)

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

31

Influence du choix de la distance

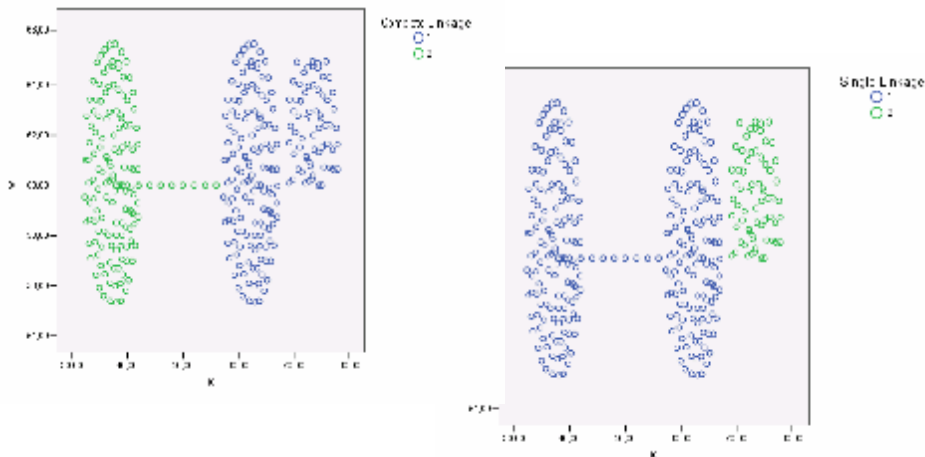


19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

32

Effet de chaîne



19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

33

Les méthodes non paramétriques en CAH (estimation de densité)

- 3 étapes :
 - estimation de la densité selon une certaine méthode :
 - plus proches voisins (la densité en un point x = nb d'observations dans une sphère centrée sur x , divisé par le volume de la sphère)
 - méthode des noyaux uniformes
 - méthode hybride de Wong (utilise les k-means dans une analyse préliminaire)
 - on définit une distance d_p entre 2 classes comme inversement proportionnelle à la densité au milieu de ces 2 classes (ou $d_p = \infty$ si les 2 classes sont non adjacentes)
 - CAH selon la méthode du saut minimum appliquée à d_p
- La densité de probabilité est un paysage vallonné dont les montagnes sont les classes et les fonds des vallées sont les frontières

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

34

Illustration de la CAH

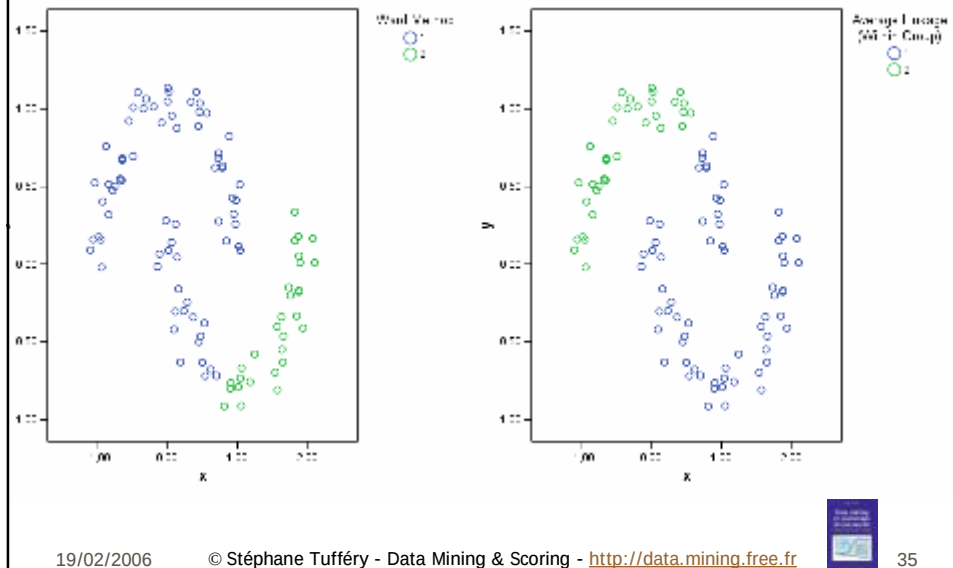


Illustration de la CAH (suite)

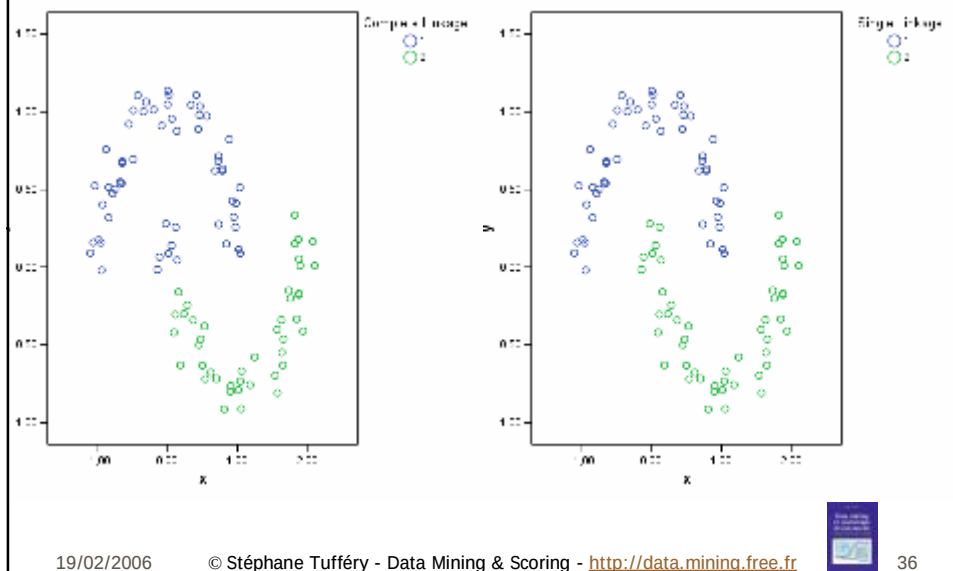
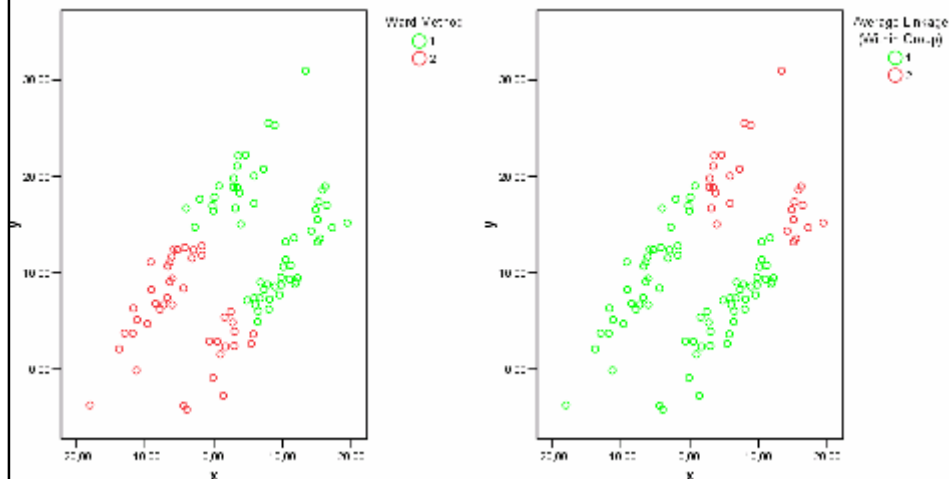


Illustration de la CAH (autre cas)

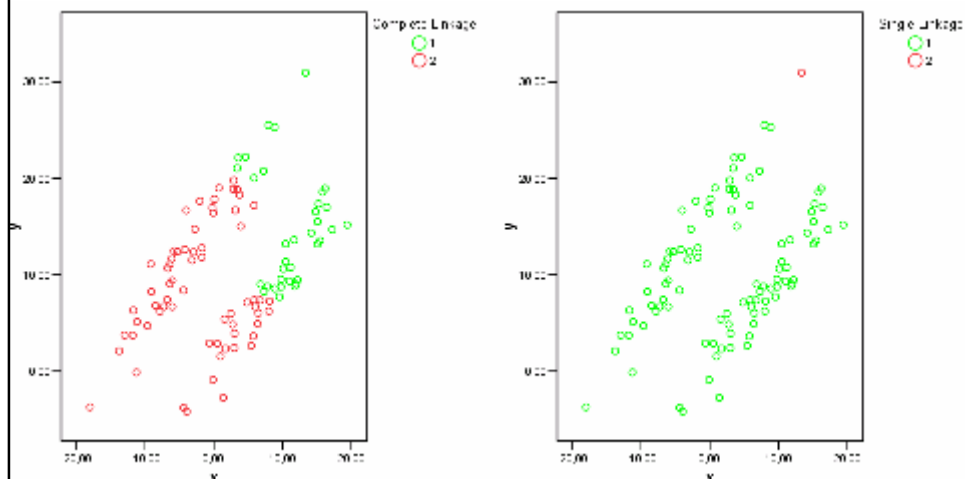


19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

37

Illustration de la CAH (fin)



19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

38

Avantages de la CAH

- Permet de classer : des individus, des variables, des moyennes de classes obtenues en sortie d'un algorithme des centres mobiles
 - si on classe des moyennes, on améliore les résultats si on connaît non seulement les moyennes des classes, mais aussi les inerties intraclasse et les effectifs des classes
- S'adapte aux diverses formes de classes, par le choix de la distance
- Permet de choisir le nombre de classes de façon optimale, grâce à des indicateurs de qualité de la classification en fonction du nombre de classes
 - R^2 semi-partiel et pseudo T^2

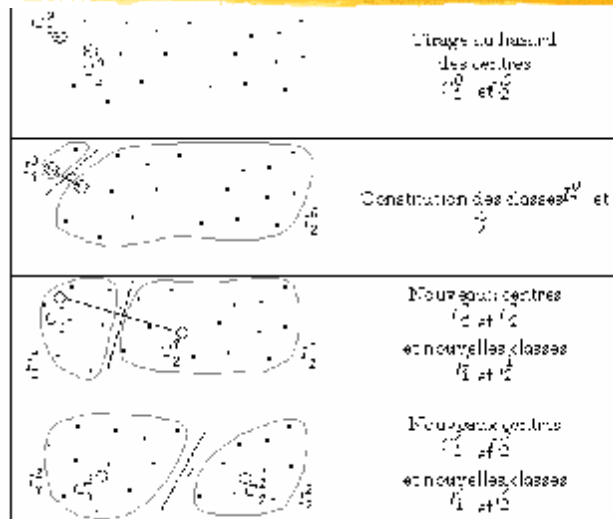


Méthode des centres mobiles

- 1) On choisit k individus $c_1, \dots, c_k$
- 2) On regroupe les autres individus autour des centres définis à l'étape 1, en sorte que le groupe de c_i soit constitué des individus plus proches de c_i que de tout autre centre
- 3) On remplace les individus de l'étape 1 par les barycentres des groupes définis à l'étape 2
- 4) Comme l'étape 2
- ...
- n) On s'arrête quand l'inertie intraclasse, qui décroît d'une partition à la suivante, ne diminue plus sensiblement (en général $n \leq 10$)



Centres mobiles : illustration



Source : Ludovic Lebart – Analyse des données appliquée - 2002

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

41

Variante

- **k-means**
 - le barycentre de chaque groupe est recalculé à chaque nouvel individu introduit dans le groupe, au lieu d'attendre l'affectation de tous les individus
 - la convergence est parfois possible en une seule itération
⇒ plus grande rapidité
- **Nuées dynamiques**
 - chaque classe n'est plus représentée par son barycentre (éventuellement extérieur à la population), mais par un sous-ensemble de la classe, appelé *noyau*, qui, s'il est bien composé (des individus les plus centraux, par exemple), sera plus représentatif de la classe que son barycentre

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

42

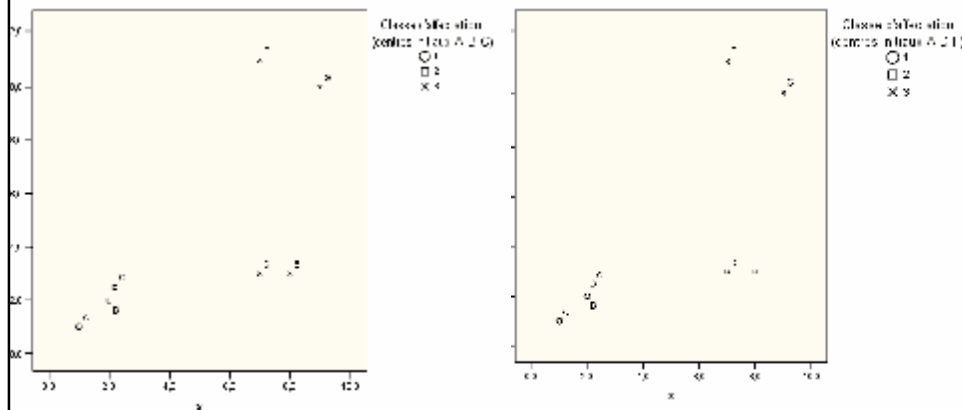
Avantages des centres mobiles

- Rapidité (complexité en n)
 - on ne compare pas toutes les observations entre elles
 - mais par rapport aux centres de classes
- Permet de détecter les outliers et de les isoler ou non
 - les outliers sont des classes réduites à un élément
 - on peut les supprimer de l'ensemble des centres initiaux et utiliser une option (« strict=s » dans la proc FASTCLUS de SAS) qui attribue le n° de classe $-k$ à toute observation plus proche de la k^e classe que des autres, mais séparée de cette classe par une distance supérieure au seuil s indiqué
 - choix du seuil en regardant les rayons (`_radius_`) des classes d'effectifs élevés et en choisissant s à peine supérieur
- Préférer la CAH en dessous de 100 observations à classer

Inconvénients de ces techniques

- Centres mobiles :
 - obligation de fixer *a priori* le nombre de classes
 - possibilité de s'aider d'une ACP
 - dépendance au choix des centres initiaux
 - ne détecte bien que les formes convexes
 - surtout sphériques de même taille
- Classification ascendante hiérarchique :
 - complexité algorithmique non linéaire (en n^2 ou n^3 , parfois $n^2 \log(n)$)
 - l'algorithme des voisins réciproques permet de passer de n^3 en n^2 , en regroupant judicieusement plus de 2 observations – pas seulement les 2 plus proches – à chaque itération
 - 2 observations placées dans des classes différentes ne sont jamais plus comparées
 - pas d'optimisation globale $\Rightarrow$ même défaut que les arbres de décision

Influence du choix des centres initiaux

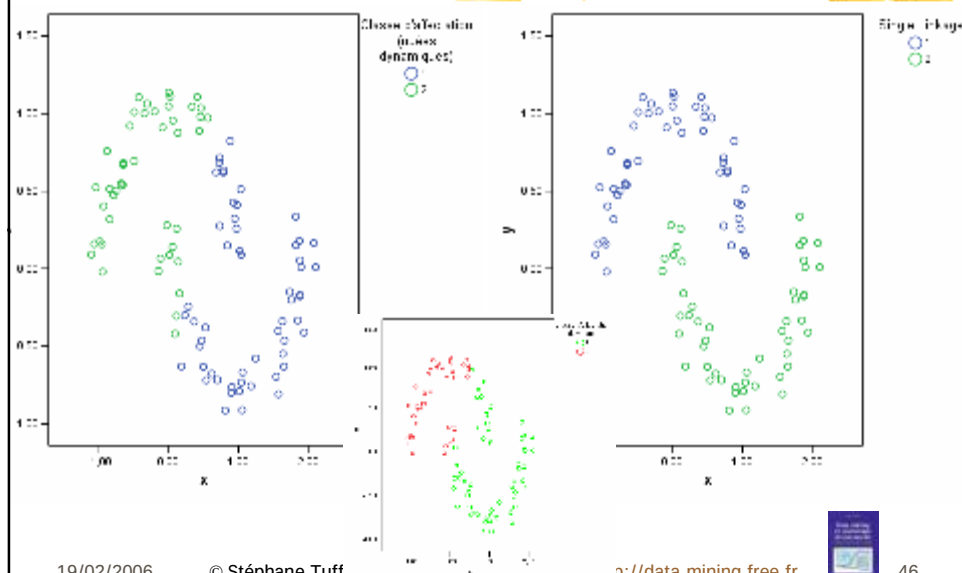


19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

45

CAH versus nuées dynamiques 1/2



19/02/2006

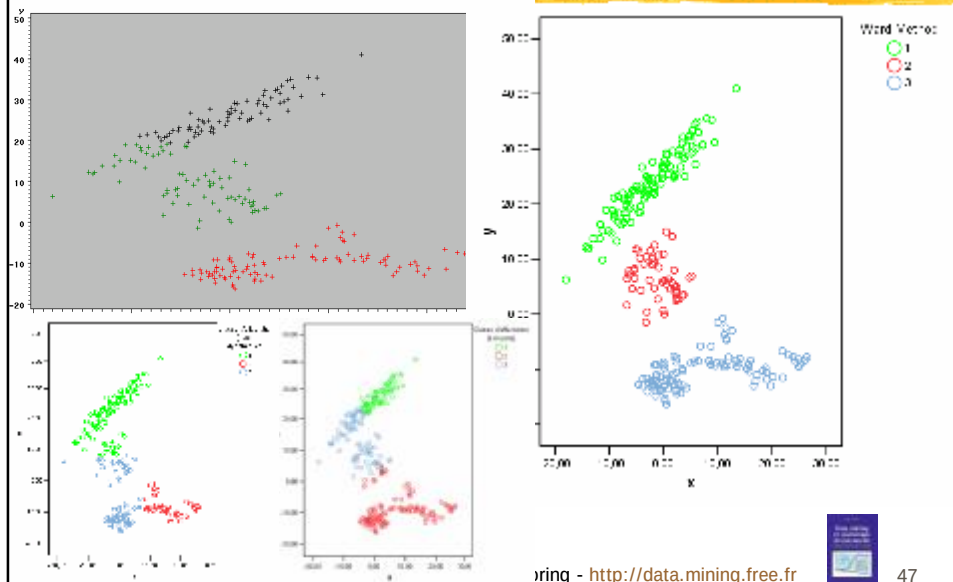
© Stéphane Tuff

<http://data.mining.free.fr>

46

CAH versus nuées dynamiques 2/2

Proc Fastclus de SAS



47

Méthodes mixtes

- Combiner les points forts de la CAH et des k-means
 - effectuer une 1^{ère} classification sur les n observations par les centres mobiles ou les k-means, en fixant le nombre de classes entre 10 et 100 (ou limite de Wong : $n^{0.3}$)
 - valeur assez grande pour limiter le risque de fusion de classes naturelles
 - puis effectuer la CAH sur les centres de ces pré-classes
- Exemples
 - la méthode par estimation de densité de Wong est obligatoirement précédée d'une procédure k-means
 - les autres méthodes par estim. de densité ne sont pas mixtes
 - les CAH « Ward » ou « average linkage » ou « centroid » peuvent être précédées d'une procédure k-means
 - algorithme BIRCH (= 2-step cluster component dans SPSS)

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

48

Méthodes mixtes avec SAS 1/12 : première phase de standardisation

- Standardiser les variables :
 - `proc standard data=dbscan out=test mean=0 std=1;`
 - `var &var;`
 - `run;`
- Cette étape est vivement recommandée si les variables ont des variances sensiblement inégales
- Utiliser une ACM si les variables sont qualitatives et non quantitatives

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

49

Méthodes mixtes avec SAS 2/12

- Lancer une procédure des centres mobiles pour obtenir :
 - (au +) maxc (ici 10) classes, dont le no est dans « cluster »
 - en (au +) maxiter (ici 50) itérations (sauf si convergence avant)
 - valeur par défaut = 1
 - valeur conseillée ≥ 10
 - valeur = 0 $\Rightarrow$ chaque individu est directement affecté à l'un des centres initiaux et on obtient aussitôt les classes finales
 - « converge=0.02 » $\Rightarrow$ les itérations s'arrêtent quand plus aucun centre ne se déplace d'une distance > 0.02
 - option sans objet si maxiter ≤ 1
 - chaque individu est affecté dans le fichier OUT à une classe
 - ces (ici 10) classes sont décrites dans le fichier MEAN
 - « delete=2 » supprime les centres avec ≤ 2 individus rattachés (pour éviter d'avoir pour centre un « outlier »)

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

50

Méthodes mixtes avec SAS 3/12 : requête pour les centres mobiles

- Code SAS :

- ```
proc fastclus data=test summary maxc=10
maxiter=50 converge=0.02 mean=centres out=presegm
cluster=presegm delete=2;
var &var; run;
```

limite les statistiques affichées  
(autre option : « short »)

- Autres options :

- « drift » : pour remplacer la méthode des centres mobiles par celle des k-means (convergence plus rapide)
- « radius=d » : spécifie la distance minimale entre 2 centres initiaux (= 0 par défaut)
- « replace=none » : accélère la phase de sélection des centres initiaux en empêchant les substitutions des centres les plus proches (choisir alors « radius » assez grand) – mieux vaut laisser l'option par défaut « replace=full »

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

51

## Méthodes mixtes avec SAS 4/12 : données écrites en sortie

- Le fichier MEAN contient des infos sur les classes :

| Obs | presegm | _FREQ_ | _RMSSTD_ | _RADIUS_ | _NEAR_ | _GAP_   | X        | Y        |
|-----|---------|--------|----------|----------|--------|---------|----------|----------|
| 1   | 1       | 21     | 0.26371  | 0.62286  | 7      | 0.90908 | -0.03354 | -0.78624 |
| 2   | 2       | 31     | 0.22412  | 0.45198  | 6      | 1.23883 | 1.25483  | 1.60977  |

- \_FREQ\_ : effectif de la classe
  - \_RMSSTD\_ : inertie intra-classe
  - \_NEAR\_ : no de la classe la + proche
  - \_GAP\_ : distance entre le centre et l'autre centre le + proche
  - \_RADIUS\_ : distance max entre 1 individu et son centre
  - X, Y : coordonnées du centre de la classe
- données utilisées par la CAH ultérieure

- Le fichier OUT contient les var. initiales + les variables :

- CLUSTER = presegm : no de la classe
- DISTANCE : séparant l'individu du centre de sa classe

| Obs | X        | Y        | N | presegm | DISTANCE |
|-----|----------|----------|---|---------|----------|
| 1   | -1.92073 | -1.49419 | 1 | 9       | 0.38436  |
| 2   | -1.88339 | -1.36798 | 1 | 9       | 0.33055  |

19/02/2006

<http://data.mining.free.fr>

52

## Méthodes mixtes avec SAS 5/12 : qualité de la classification

- Vérifier que  $R^2 \gg 0$  et  $CCC > 2$

- Exemple :

Pseudo F Statistic = 401.74

Observed Over-All R-Squared = 0.94535

Approximate Expected Over-All R-Squared = 0.90858

- calculé en supposant les variables non corrélées
- non calculé si le nb de classes > 20 % du nb d'observations

Cubic Clustering Criterion = 8.533

- calculé en supposant les variables non corrélées
- Maximiser ces valeurs en testant différentes valeurs du nb de classes

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

53

## Méthodes mixtes avec SAS 6/12 : FASTCLUS en 2 temps pour éliminer des outliers

- Première FASTCLUS pour créer un fichier de centres
  - `proc fastclus data=test summary maxc=10 maxiter=50 converge=0.02 outseed=centres delete=2;`
  - `var &var; run;`
- Suppression des centres n'ayant au plus que 2 individus
  - `data centres;set centres;where _freq_ > 2;run;`
- Nouvelle FASTCLUS sur le fichier des centres épurés avec option « `strict=0.6` » pour écarter les individus éloignés d'une distance > 0.6 de toute classe
  - `proc fastclus data=test seed=centres summary maxc=10 maxiter=50 converge=0.02 strict=0.6`
  - `mean=centres2 out=presegm cluster=presegm ;`
  - `var &var;run;`

|     |          |          |   |    |         |
|-----|----------|----------|---|----|---------|
| 102 | -0.59774 | -1.05012 | 4 | -1 | 0.65400 |
| 103 | -0.80046 | -0.97533 | 4 | -3 | 0.64009 |

19/02/2006

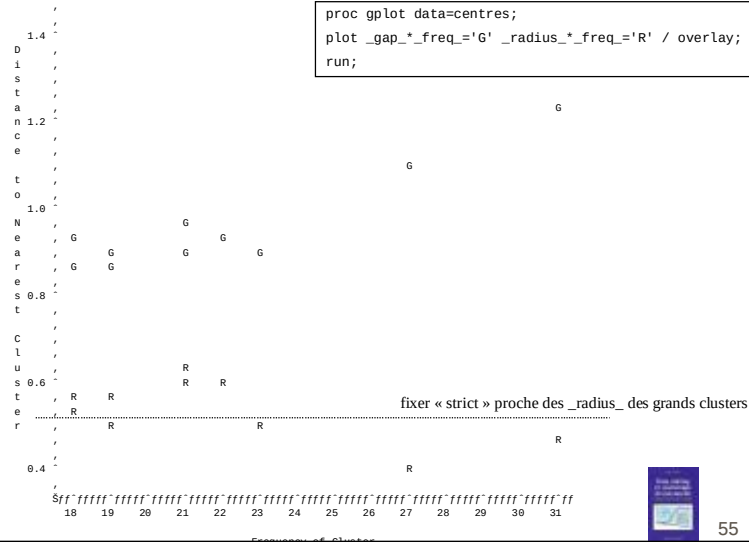
© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

54

## Méthodes mixtes avec SAS 7/12 : visualisation des classes

Plot of \_GAP\_.\_FREQ\_. Symbol used is 'G'.  
Plot of \_RADIUS\_.\_FREQ\_. Symbol used is 'R'.

```
proc gplot data=centres;
plot _gap_._freq_='G' _radius_._freq_='R' / overlay;
run;
```



19/02/2006

55

## Méthodes mixtes avec SAS 8/12 : classification hiérarchique ascendante

- CAH par la méthode de Ward appliquée aux centres en sortie de la procédure FASTCLUS :
  - `proc cluster data=centres outtree=tree method=ward ccc pseudo print=10;`
  - `var &var;`
  - `copy presegm;` ← pour recopier la variable « presegm » dans le fichier OUTTREE
  - `run;`
- CAH par la méthode de Wong appliquée aux centres en sortie de la procédure FASTCLUS :
  - `proc cluster data=centres outtree=tree method=density hybrid ccc pseudo print=10...`
  - `proc cluster data=centres outtree=tree method=twostage hybrid ccc pseudo print=10...`
- Structure hiérarchique décrite dans le fichier OUTTREE

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

56

## Méthodes mixtes avec SAS 9/12 : historique des dernières fusions de classes

| Cluster History |                 |      |      |        |      |      |      |     |      |             |
|-----------------|-----------------|------|------|--------|------|------|------|-----|------|-------------|
| NCL             | Clusters Joined |      | FREQ | SPRSQ  | RSQ  | ERSQ | CCC  | PSF | PST2 | T<br>i<br>e |
| 9               | OB6             | OB8  | 37   | 0.0160 | .931 | .906 | 5.04 | 350 | 51.2 |             |
| 8               | CL9             | OB7  | 56   | 0.0181 | .913 | .893 | 3.38 | 312 | 31.2 |             |
| 7               | OB1             | OB5  | 43   | 0.0198 | .893 | .877 | 2.41 | 292 | 69.4 |             |
| 6               | OB3             | OB10 | 38   | 0.0199 | .873 | .855 | 2.37 | 290 | 62.0 |             |
| 5               | CL7             | OB4  | 65   | 0.0315 | .842 | .824 | 1.95 | 281 | 56.3 |             |
| 4               | CL6             | CL8  | 94   | 0.0739 | .768 | .778 | -.88 | 235 | 84.1 |             |
| 3               | CL5             | CL4  | 159  | 0.1980 | .570 | .700 | -8.2 | 142 | 140  |             |
| 2               | CL3             | OB9  | 186  | 0.2255 | .344 | .530 | -7.4 | 113 | 98.0 |             |
| 1               | CL2             | OB2  | 217  | 0.3442 | .000 | .000 | 0.00 | .   | 113  |             |

ex  
aequo

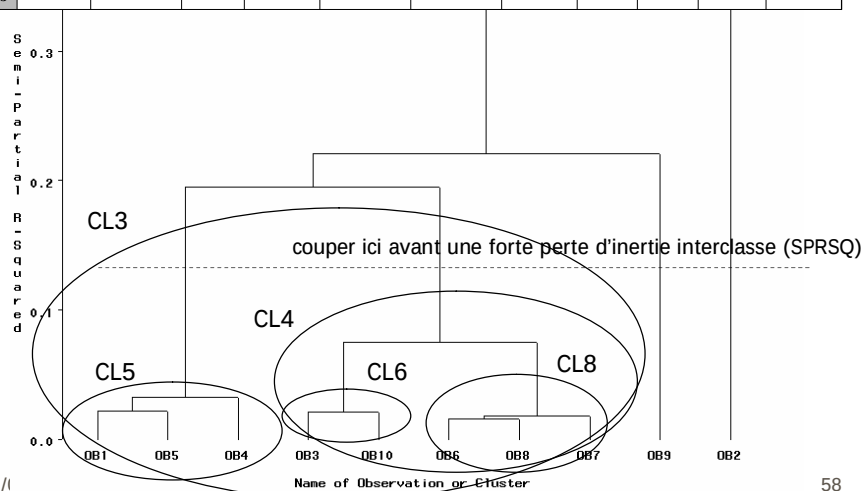
19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

57

## Méthodes mixtes avec SAS 10/12 : fichier OUTTREE et dendrogramme

| Obs | _NAME_ | _PARENT_ | _NCL_ | _FREQ_ | _HEIGHT_ | _RMSSTD_ | _SPRSQ_ | _RSQ_   | _PSF_   | _PST2_ | _ERSQ_  |
|-----|--------|----------|-------|--------|----------|----------|---------|---------|---------|--------|---------|
| 1   | OB6    | CL9      | 10    | 19     | 0.00000  | 0.26769  | 0.00000 | 0.94681 | .       | .      | .       |
| 2   | OB8    | CL9      | 10    | 18     | 0.00000  | 0.25155  | 0.00000 | 0.94681 | .       | .      | .       |
| 3   | CL9    | CL8      | 9     | 37     | 0.01599  | 0.40234  | 0.01599 | 0.93081 | 349.798 | 51.224 | 0.90621 |



19/1

58

## Méthodes mixtes avec SAS 11/12 : choix du nombre de classes

- Les classes sont notées CLn sauf les singletons notés OBn
- FREQ : nb d'individus dans la nouvelle classe
- SPRSQ ( $R^2$  semi-partiel = perte d'inertie interclasse)
- PST2 (pseudo  $t^2$ )
  - un pic pour k classes et un creux pour k+1 classes indique que la classification en k+1 classes est bonne
- RSQ (proport. de la variance expliquée par les classes)
  - le  $R^2$  doit être le + proche possible de 1
- CCC (cubic clustering criterion) ← les petites classes isolées expliquent ici un CCC légèrement négatif
  - éviter CCC << 0 (présence d'outliers)
- PSF (pseudo F)
  - rechercher une grande valeur

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

59

## Méthodes mixtes avec SAS 12/12 : fin de la requête

- La proc TREE sert à :
  - `proc tree data=tree ncl=4 out=segmhier;`
  - `copy presegm; run;` horizontal
  - afficher le dendrogramme (dans le sens – ou |)
  - créer un fichier OUT contenant le no (CLUSTER) et le nom (CLUSNAME) de la classe de chaque individu (noté OBn)
    - pour la classification ayant le nb de classes spécifié par NCL
- Il ne reste plus qu'à croiser :
  - le fichier PRESEGM contenant la pré-classe PRESEGM de chaque individu
  - le fichier SEGMHIER contenant la classe définitive CLUSTER de chaque pré-classe
  - `data segm;`
  - `merge presegm segmhier; by presegm; run;`

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

60

## Méthodes par estimation de densité

- Principe : une classe est une région entourant un maximum local de la fonction de densité de probabilité
- Détecte bien tous types de classes, en particulier les classes de formes irrégulières et de tailles et de variances inégales
- Mieux vaut standardiser les variables continues et écarter les outliers
- Réclame des effectifs suffisamment importants
- On ne fixe pas le nb de classes, mais un paramètre de lissage (« smoothing parameter ») qui est :
  - le nombre k de voisins de chaque point x
  - ou le rayon r de la sphère entourant x
  - ou le nb de classes des k-means préliminaires (Wong)

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

61

## Estimation de densité avec SAS

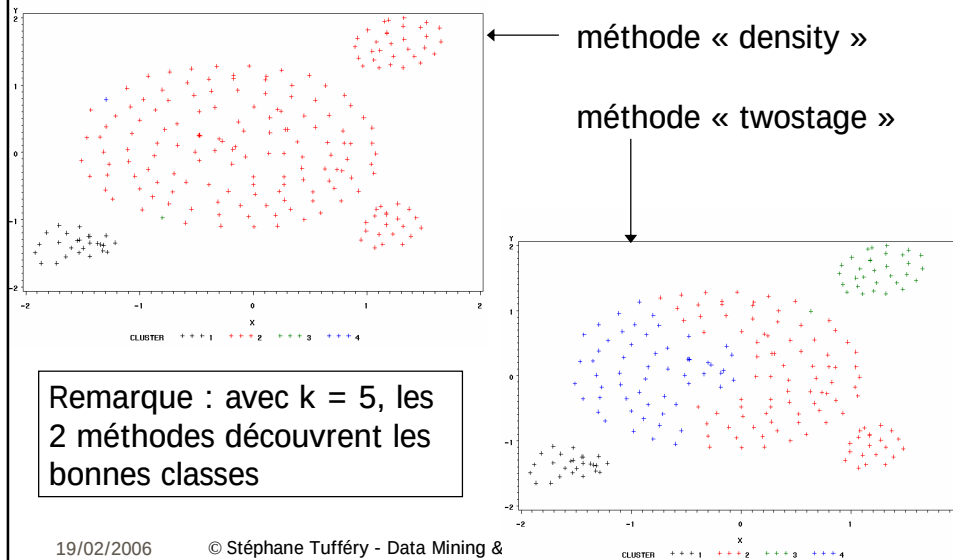
- Hormis la méthode de Wong, aucune méthode par estimation de densité n'est précédée de FASTCLUS
- Exemple de classification par les 5 plus proches voisins :
  - `proc cluster data=test outtree=tree  
method=density k=5 ccc pseudo print=10;`
  - `var &var;run;`
  - `proc tree data=tree ncl=4 out=segmhier;run;`
- Variante empêchant la fusion des grosses classes avant que les petites classes n'aient fusionné avec les grosses
  - appelée algorithme « two stage »
  - `proc cluster data=test outtree=tree  
method=two k=5 ccc pseudo print=10;`

19/02/2006

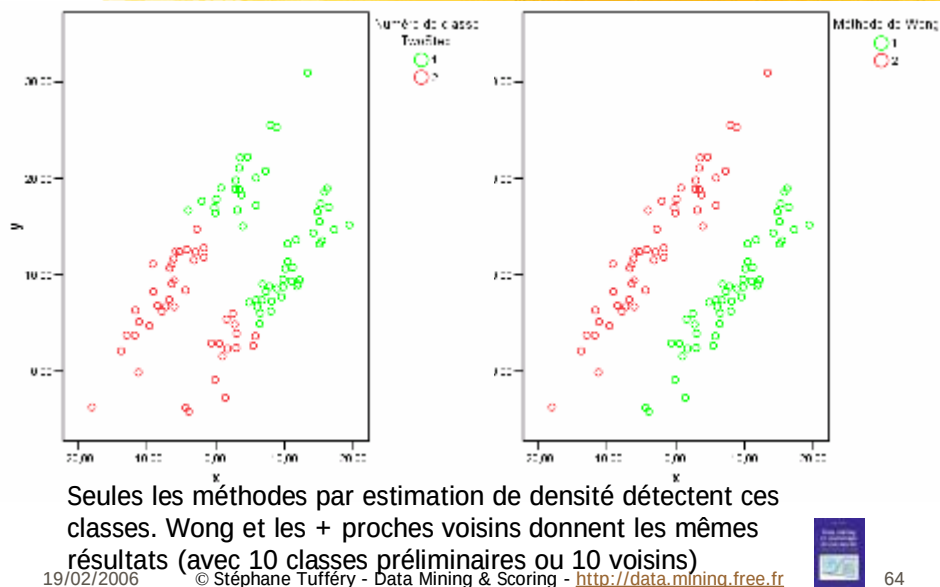
© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

62

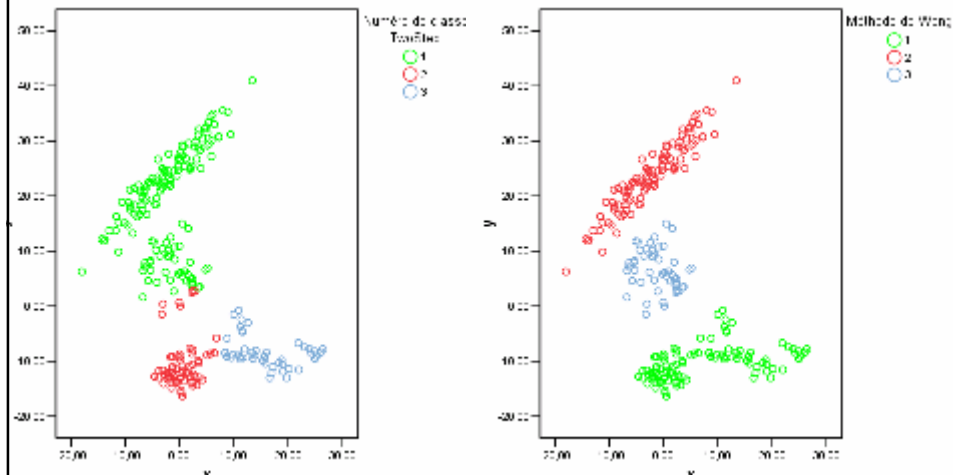
## Comparaison des méthodes avec $k = 10$



## Efficacité des méthodes par estimation de densité 1/3



## Efficacité des méthodes par estimation de densité 2/3



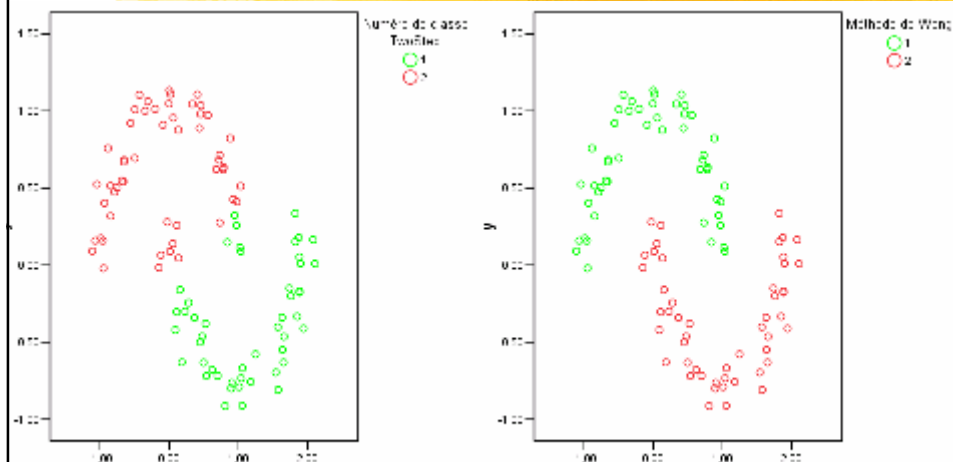
Seules les méthodes Ward et par estimation de densité détectent ces classes. Ward, Wong et les + proches voisins donnent les mêmes résultats (avec 10 classes préliminaires ou 10 voisins)

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

65

## Efficacité des méthodes par estimation de densité 3/3



Seules les méthodes single linkage et par estimation de densité détectent ces classes. Single linkage, Wong et les + proches voisins donnent les mêmes résultats (avec 10 classes prélim. ou 10 voisins)

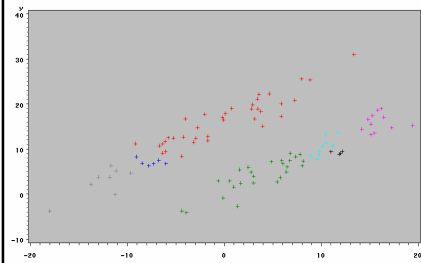
19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

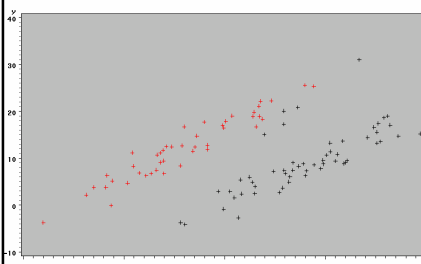
66

# Importance du paramètre de lissage

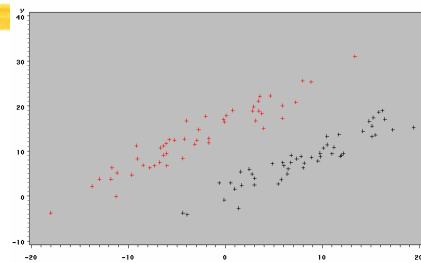
Estimation de densité par les 3 plus proches voisins



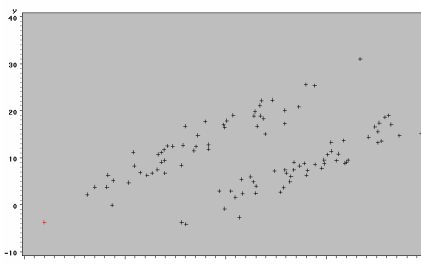
Estimation de densité par les 20 plus proches voisins



Estimation de densité par les 10 plus proches voisins



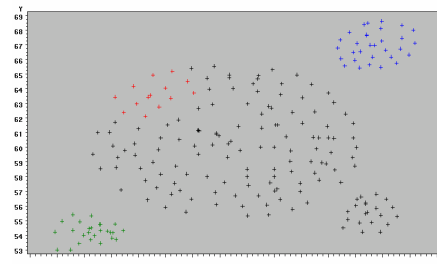
Estimation de densité par les 25 plus proches voisins



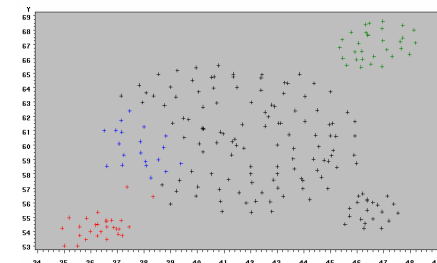
mining &

# Comparaison de méthodes par densité (single linkage est la seule autre méthode efficace ici)

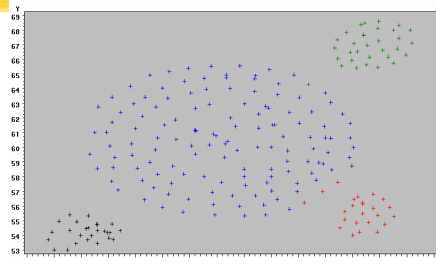
Estimation de densité par Wong (10 classes par k-means)



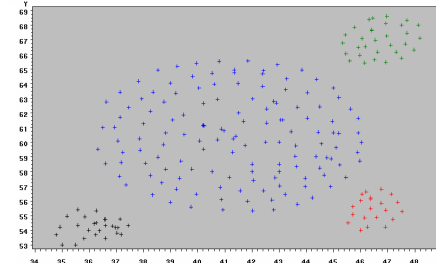
Estimation de densité par Wong (20 classes par k-means)



Estimation de densité par les 10 plus proches voisins



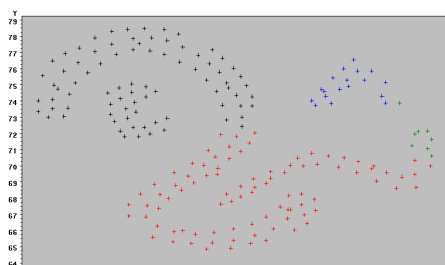
Estimation de densité par les 5 plus proches voisins



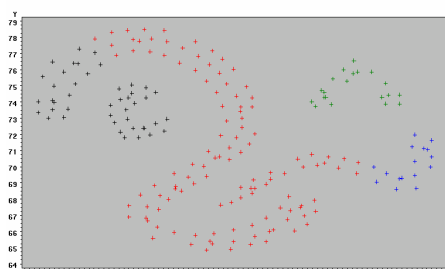
ing &

## Comparaison de méthodes par densité (single linkage est la seule autre méthode efficace ici)

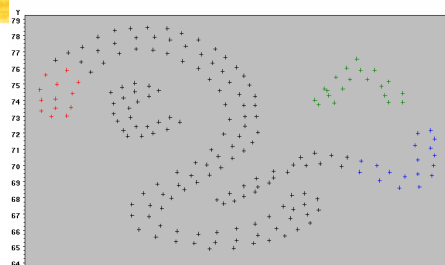
Estimation de densité par Wong (10 classes par k-means)



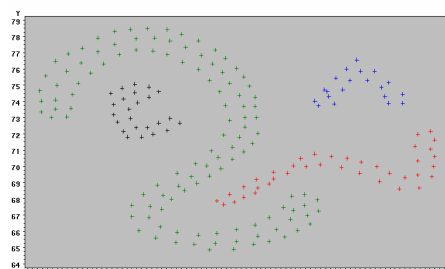
Estimation de densité par Wong (20 classes par k-means)



Estimation de densité par les 10 plus proches voisins



Estimation de densité par les 5 plus proches voisins

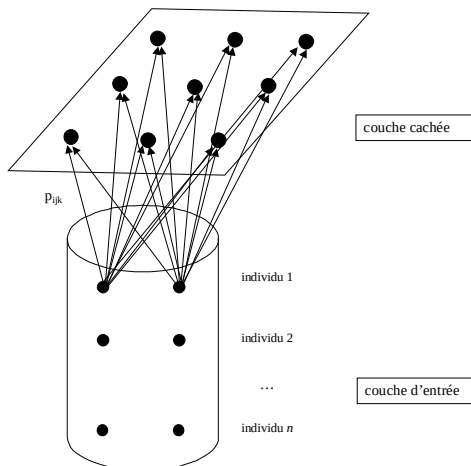


CLUSTER ++ + 1 - - - 2 - - - 3 - - - 4

ig &

CLUSTER ++ + 1 - - - 2 - - - 3 - - - 4

## Le réseau de Kohonen



Le + utilisé des réseaux de neurones  
à apprentissage non supervisé

- Pour tout individu  $(x_k)_{k \in [1, n]}$  sont calculées les distances euclidiennes le séparant des  $l \times m$  nœuds :

$$d_{ij}(x) = \sum_{k=1}^n (x_k - p_{ijk})^2$$

- Le nœud retenu pour représenter  $(x_k)$  est le nœud  $(i, j)$  pour lequel  $d_{ij}(x)$  est minimum

- Ce nœud et tous les nœuds voisins voient leurs poids ajustés  $p_{ijk} + \Theta \cdot (x_k - p_{ijk})$  pour les rapprocher de  $(x_k)$

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

70

## Application d'un réseau de Kohonen

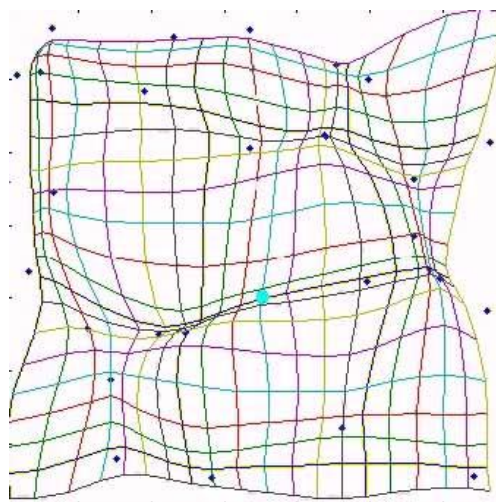
- Tout se passe comme si la grille du réseau était en caoutchouc et si on la déformait pour lui faire traverser le nuage des individus en s'approchant au plus près des individus.
  - $\neq$  avec un plan factoriel : c'est une projection non-linéaire
  - $\neq$  avec les autres méthodes : réduction de la dimension
- Une fois que tous les individus de l'échantillonnage d'apprentissage ont été présentés au réseau et que tous les poids ont été ajustés, l'apprentissage est achevé.
- En phase d'application, le réseau de Kohonen fonctionne en représentant chaque individu en entrée par le nœud du réseau qui lui est le plus proche au sens de la distance définie ci-dessus. Ce nœud sera la classe de l'individu.

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

71

## Représentation d'une carte de Kohonen



19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

72

## Utilisation des réseaux de Kohonen

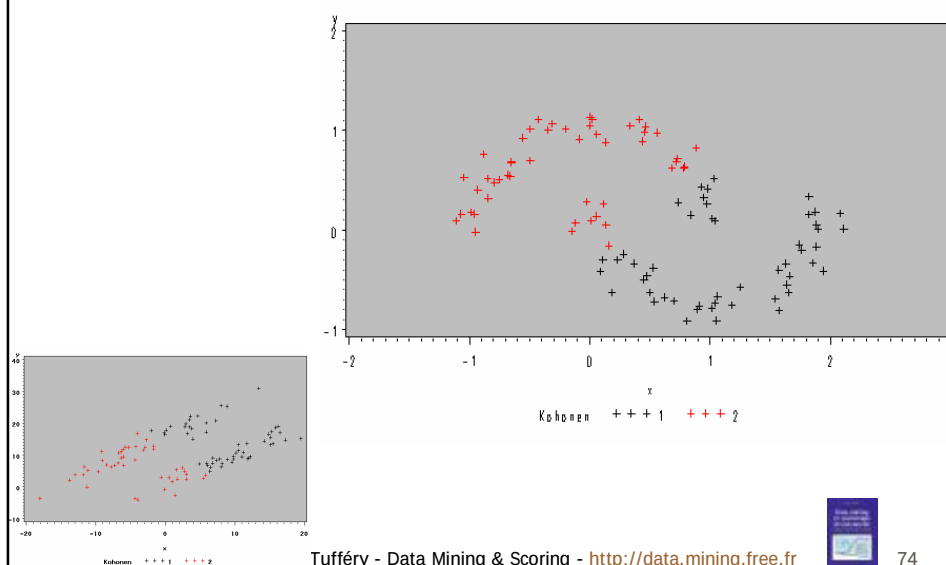
- Synonymes : 1) carte de Kohonen 2) SOM (Self Organizing Map)
- Utilisation comme une « ACP » non linéaire
  - pour représenter sur une carte les groupes d'individus et comparer les groupes s'opposant sur la carte
- Utilisation comme pré-classification avant une CAH (voir la classification mixte)
  - on construit une carte de taille au moins 10 x 10 nœuds
  - on regroupe ensuite les 100 nœuds en un nb plus petit de classes qui doivent être connexes dans le plan
- Utilisation pour placer les prototypes d'un réseau de neurones RBF
- Éviter d'utiliser directement pour obtenir qq classes
- voir les exemples suivants

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

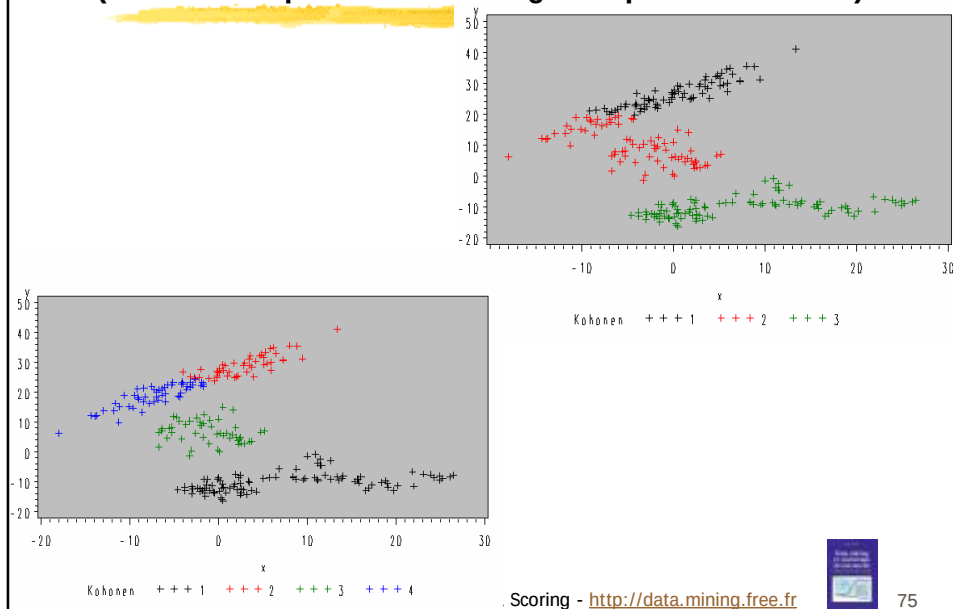
73

## Kohonen à peine mieux que k-means (moins bien que single-linkage ou Wong ou + proches voisins)



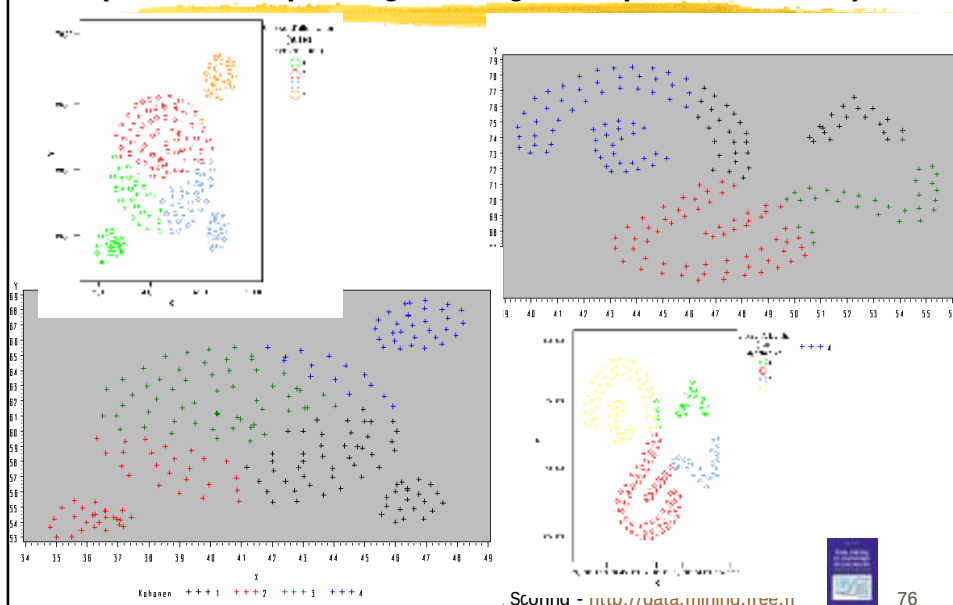
74

## Kohonen pas mieux que k-means (moins bien que Ward ou Wong ou + proches voisins)



75

## Kohonen pas mieux que k-means (moins bien que single-linkage ou + proches voisins)



76

## Analyse relationnelle

- Définition d'un critère de similarité : le **critère de Condorcet**
- Critère de Condorcet de deux individus A et B
  - $c(A,B) := (\text{nb de variables ayant même valeur pour A et B}) - (\text{nb de variables ayant des valeurs différentes pour A et B})$
  - on considère que deux variables continues ont la « même valeur » si elles sont dans le même décile
- Critère de Condorcet d'un individu A et d'une classe S
  - $c(A,S) = \sum c(A,B_i)$ , la somme étant sur tous les  $B_i \in S$ .

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

77



## La classification relationnelle

- On commence la constitution des classes en plaçant chaque individu A dans la classe S telle que  $c(A,S)$  soit maximum et  $\geq 0$ . Si  $c(A,S) < 0$  pour tout S existant, A constitue le 1<sup>er</sup> élément d'une nouvelle classe. Cette étape constitue la 1<sup>ère</sup> itération.
- On peut réaliser une 2<sup>de</sup> itération, en reprenant chaque individu et en le réaffectant éventuellement à une autre classe, parmi ceux définis lors de la 1<sup>ère</sup> itération.
- On réalise ainsi plusieurs itérations, jusqu'à ce que soit atteint le nb max d'itérations ou de classes spécifié.

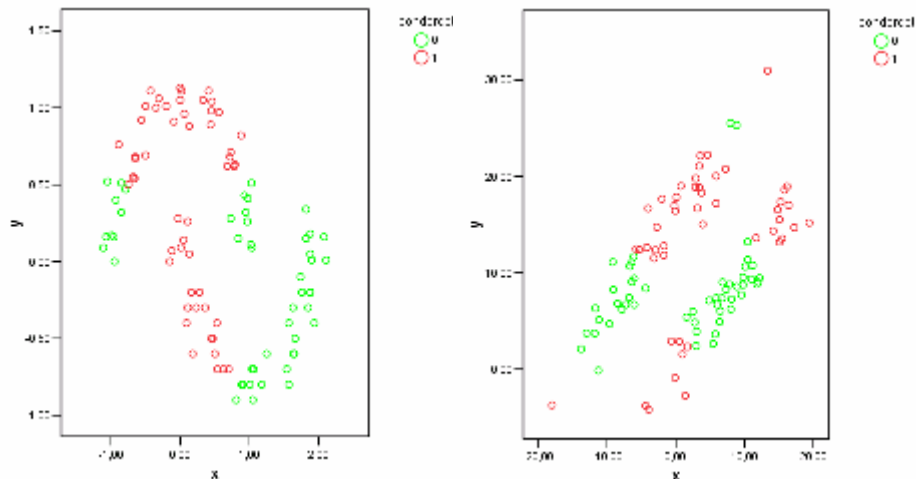
19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

78



## Illustration de l'analyse relationnelle 1/2

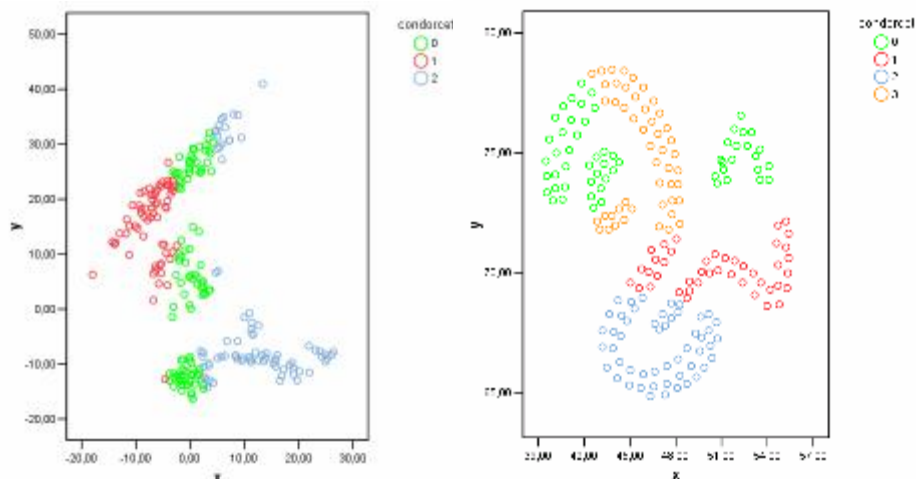


19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

79

## Illustration de l'analyse relationnelle 2/2



19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

80

## Utilisation de l'analyse relationnelle

- Faiblesse de l'analyse relationnelle sur les variables continues
  - la discrétisation fait perdre la relation d'ordre dans **R**, que ne gère pas l'analyse relationnelle
  - le principe de la méthode conduit à rassembler les individus qui sont proches sur une majorité de variables
    - dans le cas de 2 variables, on verra par exemple que si :
      - deux individus x et y sont dans le même décile de la 1<sup>ère</sup> variable, et très différents sur la 2<sup>de</sup> variable
      - x et un autre individu z sont « assez proches », car dans le même quintile (mais pas le même décile) pour les 2 variables
      - alors x sera agrégé avec y et non avec z !
    - sensibilité aux var binaires et aux variables redondantes
- Conseil : réserver l'analyse relationnelle aux variables catégorielles

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

81

## Techniques de classification : La classification de variables

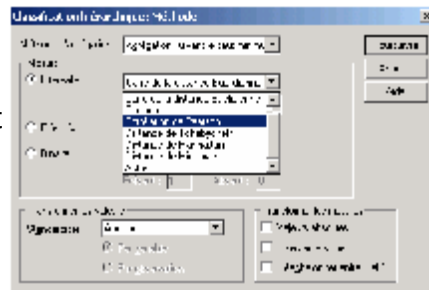
19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

82

## Classification de variables numériques

- Classification hiérarchique ascendante sur variables
  - remplacer la distance euclidienne par le coefficient de corrélation de Pearson
  - utiliser la méthode single linkage car il suffit que 2 groupes V et W de variables contiennent  $v \in V$  et  $w \in W$  suffisamment corrélées pour que les groupes soient proches
- Classification hiérarchique descendante (combinée avec ACP oblique) : procédure VARCLUS de SAS
- Ces méthodes s'appliquent à des variables numériques
  - éventuellement binaires



19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

83

## Algorithme VARCLUS de SAS 1/2 (option « maxeigen » par défaut)

- 1) On part de l'ensemble des variables
- 2) On fait une ACP et on retient les 2 premiers axes, de valeurs propres  $\lambda_1$  et  $\lambda_2$ 
  - si  $\lambda_2 > 1$  (critère de Kaiser – on peut remplacer la valeur « 1 » par une autre), on effectue une rotation quartimax oblique et on répartit les var. dans le groupe  $V_1$  des var plus corrélées avec le 1<sup>er</sup> axe, et le groupe  $V_2$  des var plus corrélées avec le 2<sup>d</sup> axe
    - ensuite on réaffecte chaque var dans  $V_1$  ou  $V_2$  pour tenter de maximiser encore plus la variance expliquée par la 1<sup>ère</sup> composante principale de chacun de ces groupes (non orthogonales entre elles)
  - si  $\lambda_2 \leq 1$ , l'ensemble des variables n'est pas divisé

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

84

## Procédure VARCLUS de SAS 2/2

- 3) On reprend en 2) avec chacune des classes  $V_1$  et  $V_2$
- 4) On s'arrête quand on n'a plus de classes avec  $\lambda_2 > 1$ 
  - ou quand un nb de classes spécifié a été atteint
  - ou quand les scissions produisent des classes vides après réaffectation

## Remarques sur VARCLUS

- Pour chaque classe, la « variable-classe » est (par défaut) la 1<sup>ère</sup> composante principale de la classe
- Le but de VARCLUS est de maximiser la somme (sur l'ensemble des classes) des variances de ces variables-classe
- VARCLUS exprime la variable-classe comme combinaison linéaire des variables contenues dans la classe
- Du fait des rotations obliques de VARCLUS, les différentes variables-classe sont (faiblement) corrélées
- On peut limiter la réaffectation de variables à l'intérieur d'un découpage de classes, pour ne pas affecter une variable à une classe anciennement formée, ce qui détruirait la structure d'arbre de la classification
  - option « hierarchy » - permet de limiter le temps de calcul

## Différence entre VARCLUS et une ACP standard

- Dans une ACP, toutes les composantes principales sont calculées à partir des mêmes variables (variables initiales)
- Dans VARCLUS
  - les variables initiales sont séparées itérativement en sous-groupes (par quartimax)
  - et les composantes principales sont calculées sur ces sous-groupes et non l'ensemble des variables
- Si on limite à k le nb de classes de variables, de sorte que l'on ne retient que k composantes par VARCLUS, ces k composantes expliquent peut-être moins de variance que les k premières composantes de l'ACP, mais elles sont plus faciles à interpréter
- VARCLUS est une méthode performante de classification
  - mais plus consommatrice en temps de calcul que l'ACP : attention si plus de 30 variables

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

87

## Techniques de data mining : Recherche d'associations (analyse du ticket de caisse)

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

88

## Les recherches d'associations

- Rechercher les **associations** consiste à rechercher les règles du type :
  - « Si pour un individu, la variable  $A = x_A$ , la variable  $B = x_B$ , etc., alors, dans 80 % des cas, la variable  $Z = x_Z$ , cette configuration se rencontrant pour 20 % des individus. »
- La valeur de 80 % est appelée **indice de confiance** et la valeur de 20 % est appelée **indice de support**.
- Par exemple, dans l'ensemble de transactions suivant :

|       |   |   |   |   |   |
|-------|---|---|---|---|---|
| T26   | A | B | C | D | E |
| T163  | B | C | E | F |   |
| T1728 | B | E |   |   |   |
| T2718 | A | B | D |   |   |
| T3141 | C | D |   |   |   |

l'indice de confiance de «  $B \Rightarrow E$  » = 3/4 et son indice de support = 3/5.

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

89

## Les associations : définitions

- Une règle est donc une expression de la forme :
  - > Si *Condition* alors *Résultat*.
- Exemple :
  - > Si *couches* et *samedi*, alors *bière*.
- L'indice de support est la probabilité :
  - >  $p(\text{condition et résultat})$ .
- Il ne doit pas être trop petit.
- L'indice de confiance est la probabilité :
  - >  $p(\text{condition et résultat}) / p(\text{condition})$ .

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

90

## Intérêt d'une règle d'association

- Dans l'exemple précédent, on a :
  - indice de confiance de l'association  $C \Rightarrow B$  est  $2/3$
  - indice de support =  $2/5$ .
- Or, probabilité (B) =  $0,8$ 
  - B est présent dans presque tous les tickets de caisse.
- Cette probabilité est supérieure à l'indice de confiance de  $C \Rightarrow B$ , ce qui fait que l'on ne gagne rien à utiliser la règle  $C \Rightarrow B$  pour prédire B.
- Si l'on suppose aléatoirement qu'un ticket de caisse contient B, on n'a qu'1 chance / 5 de se tromper, contre 1 chance / 3 en appliquant la règle  $C \Rightarrow B$ .

## Lift d'une règle : mesure son intérêt

- L'amélioration apportée par une règle, par rapport à une réponse au hasard est appelée « **lift** » et vaut :
  - $\text{lift (règle)} = \text{confiance (règle)} / p(\text{résultat})$
  - $= p(\text{condition et résultat}) / [p(\text{condition}) \times p(\text{résultat})]$ .
- Quand le lift est  $< 1$ , la règle n'apporte rien.
- Exemples :
  - $\text{lift}(C \Rightarrow B) = 5/6$  (règle inutile)
  - $\text{lift}(B \Rightarrow E) = 5/4$  (règle utile).

## Lift de la règle inverse

- Il faut noter que si le lift de la règle
  - Si *Condition* alors *Résultat*
- est  $< 1$ , alors le lift de la règle inverse, c.a.d. de :
  - Si *Condition* alors *NON Résultat*
- est  $> 1$ , puisque :
  - confiance (règle inverse) =  $1 - \text{confiance (règle)}$
- et
  - $p(\text{NON résultat}) = 1 - p(\text{résultat})$ .
- Si une règle n'est pas utile, on peut donc essayer la règle inverse... en espérant que cette dernière soit intéressante en termes de métier ou de marketing.

## Taxinomie : définition

- Les produits peuvent être définies avec un niveau plus ou moins fin de détail.
- On peut par exemple considérer :
  - les produits d'épargne bancaire, financière...
  - parmi les produits d'épargne bancaire, les comptes de chèques, les livrets...
  - parmi les livrets, les livrets A, les Codevi, les LEP...
- La **taxinomie** des produits est l'ensemble de ces niveaux.

## Taxinomie : utilisation

- Le **niveau le plus fin** permet d'entreprendre des actions commerciales plus précises.
  - Mais travailler au niveau le plus fin multiplie les règles, parmi lesquelles un grand nombre n'auront qu'un faible support et seront peut-être éliminées.
  - Travailler au **niveau le plus général** permet d'obtenir des règles plus fortes.
- > Les 2 points de vue ont leurs avantages et leurs inconvénients.
- > Il faut adapter le niveau de généralité à chaque produit, en fonction notamment de sa rareté.

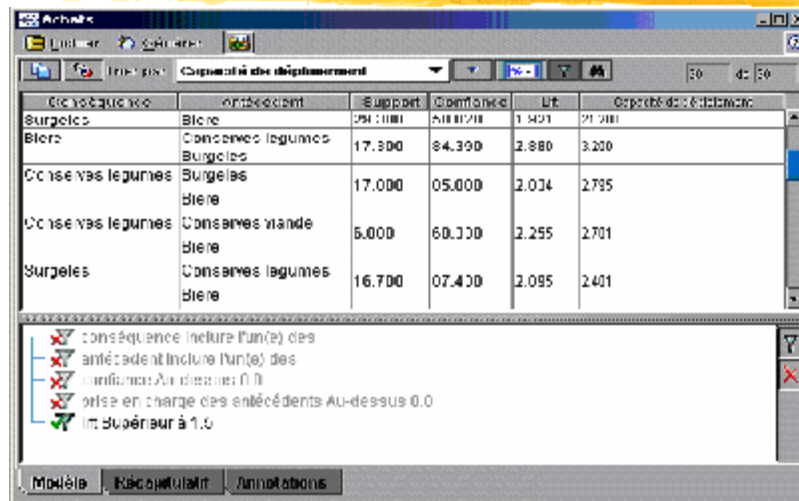


## Taxinomie : intérêt

- Les articles les plus rares et les plus chers (exemple : micro-informatique ou HIFI dans un grand magasin) seront codifiés au niveau le plus fin
- Les articles les plus courants (exemple : produits alimentaires) seront codifiés à un niveau plus général.
- On regroupera par exemple tous les yaourts, fromages blancs, flancs... en « produits laitiers », tout en distinguant un téléviseur d'un magnétoscope ou d'un caméscope.
- L'intérêt de cette façon de procéder est d'obtenir des règles plus pertinentes, dans lesquelles les articles les plus courants ne dissimulent pas, par leur fréquence, les articles les moins courants.



## Illustration avec <sup>®</sup>Clementine



| Catégorie         | Antécédent        | Support  | Confiance | Ut    | Confiance de l'association |
|-------------------|-------------------|----------|-----------|-------|----------------------------|
| Surgeles          | Bière             | 294.1111 | 818.1218  | 1.921 | 21.2111                    |
| Bières            | Conserves légumes | 17.300   | 84.390    | 2.880 | 3.200                      |
|                   | Surgeles          |          |           |       |                            |
| Conserves légumes | Surgeles          | 17.000   | 05.000    | 2.014 | 2.795                      |
|                   | Bière             |          |           |       |                            |
| Conserves légumes | Conserves viande  | 5.000    | 60.100    | 2.255 | 2.701                      |
|                   | Bière             |          |           |       |                            |
| Surgeles          | Conserves légumes | 16.700   | 07.400    | 2.095 | 2.401                      |
|                   | Bière             |          |           |       |                            |

☒ conséquence inclure l'un(e) des  
☒ antécédent inclure l'un(e) des  
☒ confiance Au-dessus de 0.0  
☒ mise en charge des antécédents Au-dessus de 0.0  
☒ Ut Supérieur à 1.0

19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

97

## Le panier de la ménagère

- Cette technique est très utilisée dans la *grande distribution* pour analyser les tickets de caisse
  - > d'où le terme de « **panier de la ménagère** » pour désigner les associations.
- Autre usage :
  - associations d'options retenues dans les produits packagés (banque, téléphonie, assurance...)



19/02/2006

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

98

## Utilisation de variables supplémentaires

- En ajoutant des variables temporelles (jour et heure de la transaction), on pourra rechercher l'ensemble des événements qui débouchent sur l'acquisition d'un nouveau produit, sur le départ du client...
- En ajoutant le nom d'un fabricant présent dans plusieurs types de produits, on pourra détecter des phénomènes d'attachement à une marque
- Autres variables supplémentaires :
  - canal de distribution
  - mode de paiement
  - ...

