CONSERVATOIRE NATIONAL DES ARTS ET MÉTIERS
CEDRIC

PSA PEUGEOT CITROËN

Centre de Recherche en Informatique du CNAM

Conservatoire National des Arts et Métiers
Chaire de Statistique Appliquée - Case 41
292 rue Saint Martin
75 141 Paris Cedex 03

Classification préalable à la recherche de règles d'association

Marie Plasse
marie.plasse@mpsa.com

Ndeye Niang
niang@cnam.fr

Gilbert Saporta
saporta@cnam.fr

RIAs'2006

2èmes Rencontres Inter-Associations

"la classification et ses applications"

20-21 mars 2006, Lyon

PSA PEUGEOT CITROËN

SOMMAIRE

CONSERVATOIRE NATIONAL DES ARTS ET METIERS

CEDRIC

PROBLEMATIQUE INDUSTRIELLE

Les données

- Plus de 80000 véhicules décrits par plus de 3000 attributs binaires

Véhicules	A1	A2	A3	A4	A5	…	Ap
	1	0	0	1	0		0
	0	0	1	1	0		0
	0	1	0	0	1		0
	1	0	0	0	1		0
	0	1	0	0	1		1
	0	1	0	0	1		0
	0	0	1	0	0		0

Matrice de données binaires

=

Véhicules	Attributs présents
	{A1, A4}
	{A3, A4}
	{A2, A5}
	{A1, A5}
	{A2, A5, Ap}
	{A2, A5}
	{A3}

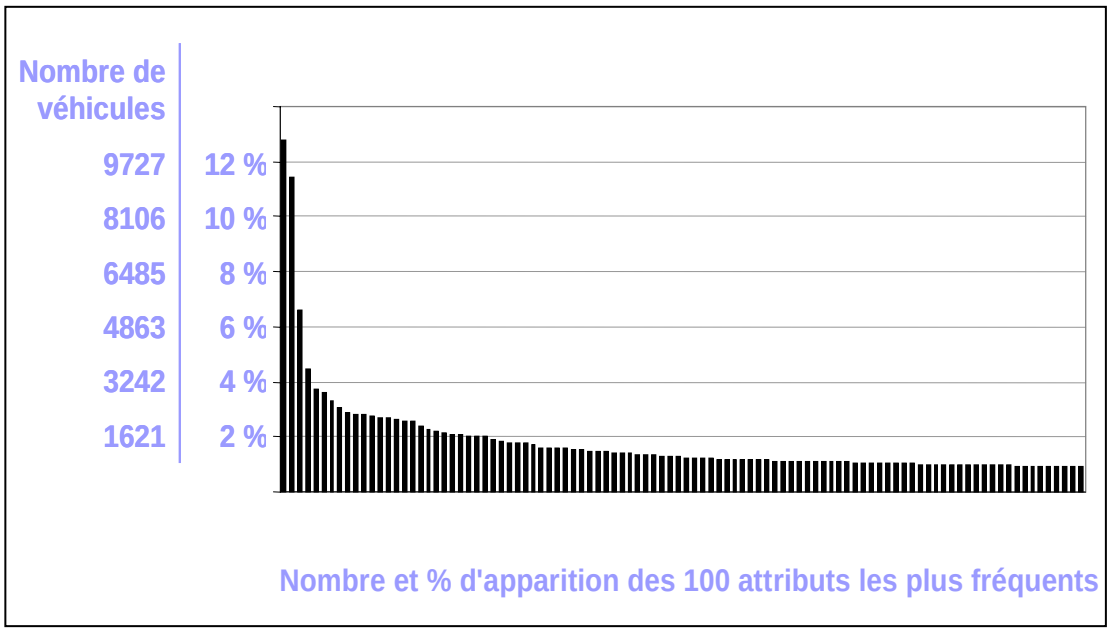
Données de transaction

- Trouver des corrélations entre les attributs…
- … grâce à la recherche de règles d'association

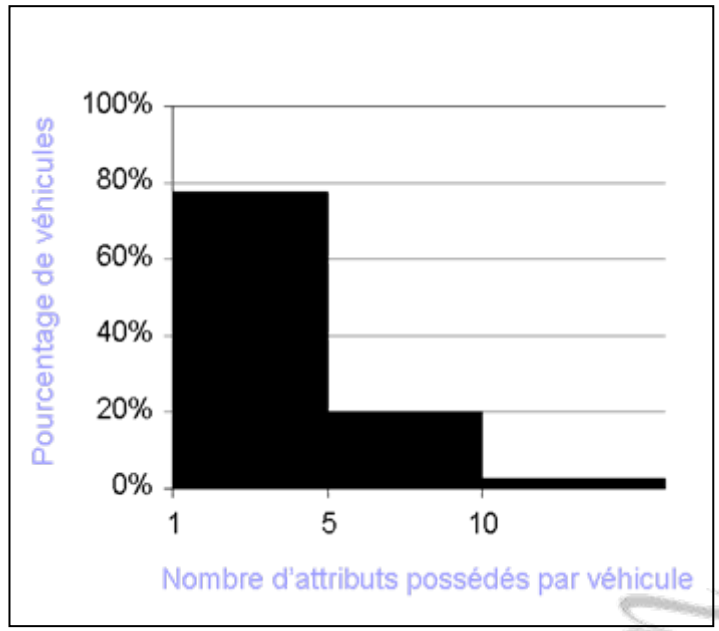
PROBLEMATIQUE INDUSTRIELLE

Spécificités des données

- Des données clairsemées :
- Répartition des 100 attributs les plus fréquents :



- Nombre d'attributs présents par véhicule :



- 4 attributs en moyenne

LA RECHERCHE DE REGLES D'ASSOCIATION

Rappel de la méthode

- Origine marketing : analyser les ventes des supermarchés

  "lorsqu'un client achète du pain et du beurre, il achète 9 fois sur 10 du lait en même temps"

- Formalisation : $A \rightarrow C$ **où** $A \cap C = \emptyset$

- Fiabilité : **Support** : % de transactions contenant A et C

$$sup(A \rightarrow C) = P(A \cap C) = P(C / A) \cdot P(A)$$

s(A → C) = 30 %
➲ 30% des transactions contiennent à la fois + +

- Précision : **Confiance** : % de transactions contenant C sachant qu'elles ont A

$$conf(A \rightarrow C) = P(C / A) = \frac{P(A \cap C)}{P(A)} = \frac{sup(A \rightarrow C)}{sup(A)}$$

c(A → C) = 90 %
➲ 90% des transactions qui contiennent + contiennent aussi

- Algorithmes :
  - Recherche des **sous-ensembles fréquents** (avec minsup)
  - Extraction des **règles d'association** (avec minconf)

- **Apriori** *(Agrawal & Srikant, 1994)*
- **Partition** *(Saverese et al., 1995)*
- **Sampling** *(Brin & Motwani, 1997)*
- **Eclat** *(Zaki, 2000)*
- **FP-Growth** *(Han & Pei, 2003)*

LA RECHERCHE DE REGLES D'ASSOCIATION

Extraction des règles

Support minimum (nb de véhicules vérifiant la règle)	Confiance minimum	Nombre de règles	Taille maximum des règles obtenues
500	50 %	16	3
400	50 %	29	3
300	50 %	194	5
250	50 %	1299	6
200	50 %	102 981	10
100	50 %	1 623 555	13

Objectifs :

- **Réduire le nombre de règles**
- **Simplifier les règles**

Première réduction en regroupant les attributs de manière experte :

Support minimum	Confiance minimum	Nombre de règles	Taille maximum des règles obtenues
100	50 %	600636	12

CLASSIFICATION DE VARIABLES

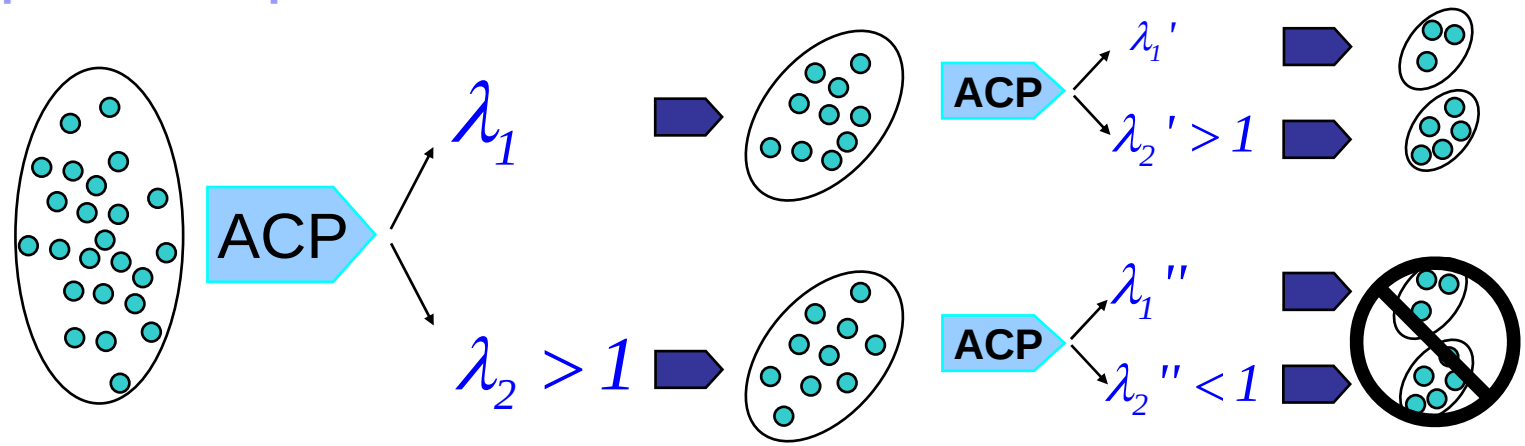
Méthodes utilisées

Classification ascendante hiérarchique

- **Hiérarchie de partitions emboîtées**
- **Stratégie d'agrégation de Ward**

Classification descendante (Proc Varclus)

- **Recherche d'une partition en classes unidimensionnelles**
- **Regroupement autour de la composante principale avec laquelle la variable présente le plus fort coefficient de corrélation**



Partitionnement : méthode de Qannari et Vigneau (2003)

- **Regroupement des variables autour de variables latentes**

CLASSIFCIATION DE VARIABLES

Indices de similarité utilisés

Véhicule	j	j'
1	1	0
2	0	1
3	0	0
…		
n	1	1

Variable j \ Variable j'	1	0	
1	n_{11}	n_{10}	$n_{1.}$
0	n_{01}	n_{00}	$n_{0.}$
	$n_{.1}$	$n_{.0}$	Σ = n

$r^2_{jj'} = \Phi^2_{jj'} = \frac{\chi^2_{jj'}}{n} = \frac{n_{11}n_{00} - n_{01}n_{10}}{n_{1.}n_{0.}n_{.1}n_{.0}}$

Ochiai $s_O(jj') = \frac{n_{11}}{\sqrt{(n_{11}+n_{10})(n_{11}+n_{01})}} = \frac{n_{11}}{\sqrt{n_{1.}n_{.1}}}$

Jaccard $s_J(jj') = \frac{n_{11}}{n_{11}+n_{10}+n_{01}} = \frac{n_{11}}{n-n_{00}} = \frac{n_{11}}{n_{1.}+n_{01}}$

Dice $s_D(jj') = \frac{2n_{11}}{2n_{11}+n_{10}+n_{01}}$

Russel & Rao $s_{RR}(jj') = \frac{n_{11}}{n_{11}+n_{10}+n_{01}+n_{00}} = \frac{n_{11}}{n}$

$$S_{RusselRao}(jj') = \frac{n_{11}}{n} = \frac{\text{nb de cooccurrences}}{\text{nb de cas}} = support(\{j, j'\})$$

CLASSIFCIATION DE VARIABLES

Composition des classes

Nombre de variables par classe :

		Classes									
Indices	R^2	3058	10	6	5	4	4	4	4	3	3
	Ochiai	2762	201	84	13	11	8	6	6	5	5
	Jaccard	2973	72	12	10	8	6	6	5	5	4
	Dice	2690	298	61	12	11	7	6	6	5	5
	Russel & Rao	2928	117	16	12	10	5	5	4	2	2
Varclus		1282	1001	349	156	111	61	60	41	28	12

Remarque sur le choix du nombre de classes

CLASSIFCIATION DE VARIABLES

Comparaison des partitions obtenues

Comparaison des partitions 2 à 2 grâce à l'indice de Rand :

$$R = \frac{2\sum_u \sum_v n_{uv}^2 - \sum_u n_{u.}^2 - \sum_v n_{.v}^2 + n^2}{n^2}$$

	Ward-R²	Ward-Ochiai	Ward-Jaccard	Ward-Dice	Ward-Russel Rao	Varclus
Ward-R²						
Ward-Ochiai	0,82					
Ward-Jaccard	0,94	0,87				
Ward-Dice	0,78	0,79	0,82			
Ward-Russel Rao	0,87	0,80	0,84	0,86		
Varclus	0,31	0,39	0,34	0,41	0,35	

Pourcentage de paires en accord :

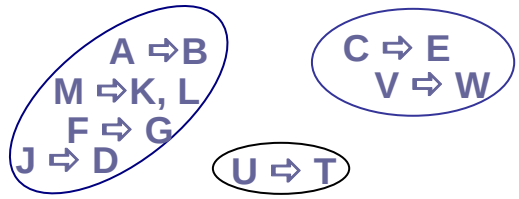
- Variables dans la même classe dans les deux partitions

- Variables séparées dans les deux partitions

UTILISATION CONJOINTE

Extraction des règles après classification

Recherche des règles d'association à l'intérieur de chaque classe :



Essai	Nombre de règles	Complexité maximum	Réduction du nombre de règles
Sans Classif.	600636	12	.
Ward - R^2	600637	12	0 %
Ward - Jaccard	481649	12	20 %
Ward - Russel & Rao	481388	12	
Ward - Ochiai	479474	12	
Ward - Dice	481648	12	
Varclus	5239	4	99 %

A première vue, pas de réduction du nombre de règles

UTILISATION CONJOINTE

Analyse du nombre de règles dans chaque classe

Partition en dix classes avec l'indice de Russel & Rao :

Classe	Nombre de variables dans la classe	Nombre de règles obtenues	Complexité maximum
1	2	0	0
2	12	481170	12
3	2	0	0
4	5	24	4
5	117	55	4
6	4	22	4
7	10	33	4
8	5	22	4
9	16	1	2
10	2928	61	4

Classe identique quelque soit la classification

➲ règles complexes

UTILISATION CONJOINTE

Réduction du nombre de règles

Recherche des règles d'association à l'intérieur de chaque classe sauf de la classe atypique :

	Nombre de règles	Complexité maximum	Réduction du nombre de règles
Sans Classif.	600636	12	.
Ward - R^2	43	4	+ de 99 %
Ward - Jaccard	479	5	
Ward - Russel & Rao	218	4	
Ward - Ochiai	459	5	
Ward - Dice	478	5	
Varclus	21	4	

Réduction du nombre de règles à analyser

Simplification de ces règles

Identification d'un groupe de 12 attributs à traiter séparément

CONCLUSION & PERSPECTIVES

La classification de variables préalable a permis de :

- **Mettre en évidence une classe atypique**
- **Diminuer le nombre et la complexité des règles produites**

Le choix de l'indice de Russel Rao est cohérent en raison de sa relation avec le support

Pour extraire les règles les plus intéressantes, nous avons utilisé des indices de pertinence tels que le lift, et dans notre cas l'indice de Jaccard :

$$Lift = \frac{P(AC)}{P(A).P(C)} \qquad Jac = \frac{P(AC)}{P(A)+P(C)-P(AC)}$$

Nous sommes en train d'étudier les apports de la classification croisée à notre approche…

REFERENCES BIBLIOGRAPHIQUES

Agrawal R., Srikant R. (1994) *Fast Algorithms for Mining Association Rules*. In : Proceedings of the 20th Int'l Conference on Very Large Databases (VLDB), Santiago, Chile.

Fichet B., Le Calve G. (1984) *Structure géométrique des principaux indices de dissimilarité sur signes de présence-absence*. Statistiques et Analyse des données, Vol. 9 n°3 pp11-44

Gower J.C., Legendre P. (1986) *Metric and euclidean properties of dissimilarity coefficients*. In : Journal of Classification Vol.3, pp 5-48

Hébrail G., Lechevallier Y. (2003) Data mining et analyse des données. In : Govaert G. Analyse des données. Ed. Lavoisier, Paris, pp 323-355

Nakache J.P., Confais J. (2005) *Approche pragmatique de la classification*, Ed. Technip, Paris

Plasse M., Niang N., Saporta G. (2005). Utilisation conjointe des règles d'association et de la classification de variables. Journées Françaises de Statistiques, Pau, France.

Vigneau E., Qannari E.M. (2003) *Clustering of variables around latent component - application to sensory analysis*. Communications in Statistics , Simulation and Computation, 32(4), pp 1131-1150

Youness G., Saporta G. (2004) *Some Measures of Agreement Between Close Partitions* - Student vol. 5(1), pp. 1-12.