

Combined use of association rules mining and clustering methods

Marie Plasse, Ndeye Niang-Keita, Gilbert Saporta, Damien Gauthier

marie.plasse@mpsa.com

niang@cnam.fr

saporta@cnam.fr

damien.gauthier@mpsa.com

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

*3rd IASC world conference on
Computational Statistics & Data Analysis
Limassol, Cyprus, 28-31 October, 2005*








MOTIVATION

- **Industrial data :**

- A set of vehicles described by a large set of binary flags

- **Motivation : decision-making aid**

- Always searching for a greater quality level, the car manufacturer can take advantage of knowledge of associations between attributes.

Vehicles	A1	A2	A2	A2	A3	...	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	0		1
	0	1	0	0	0		0
	0	0	1	0	0		0

- **Our work :**

- We are looking for patterns in data : Associations discovery

Overview

Motivation

→ Association rules mining

Clustering techniques

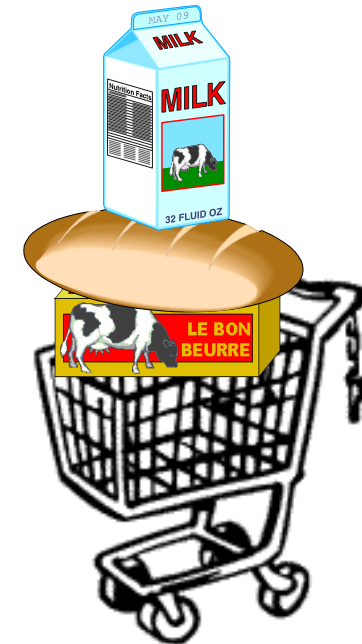
Combined use of the two methods & application

Conclusion & future work

ASSOCIATION RULES MINING

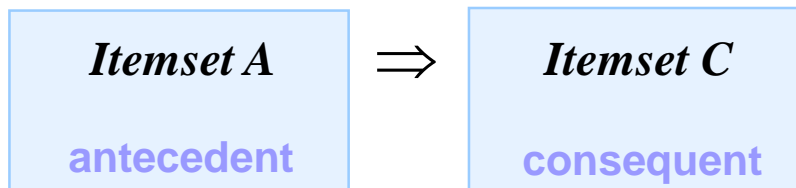
- Marketing target : basket data analysis

Basket	Purchases
1	{bread, butter, milk}
2	{bread, meat}
...	
n	{fruit juice, fish, strawberries, bread}



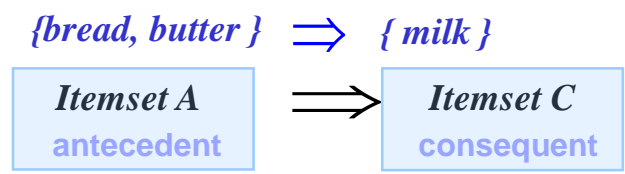
"90% of transactions that purchase bread and butter also purchase milk" (Agrawal et al., 1993)

$$\{ \textit{bread, butter} \} \Rightarrow \{ \textit{milk} \}$$



where $A \cap C = \emptyset$

THE BASICS






- **Reliability : Support : % of transactions that contain all items of A and C**

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

- **Supp = 30 %** → 30% of transactions contain  +  + 

- **Strength : Confidence : % de transactions that contain C among the ones that contain C**

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

- **Conf = 90 %** → 90% of transactions that contain  +  , contain also 

ALGORITHM'S RUDIMENTS

- **First algorithm in 1993 (Agrawal et Srikant) → *APriori* algorithm in 1994**
 - Find frequent itemsets (support)
 - Find association rules (confidence)

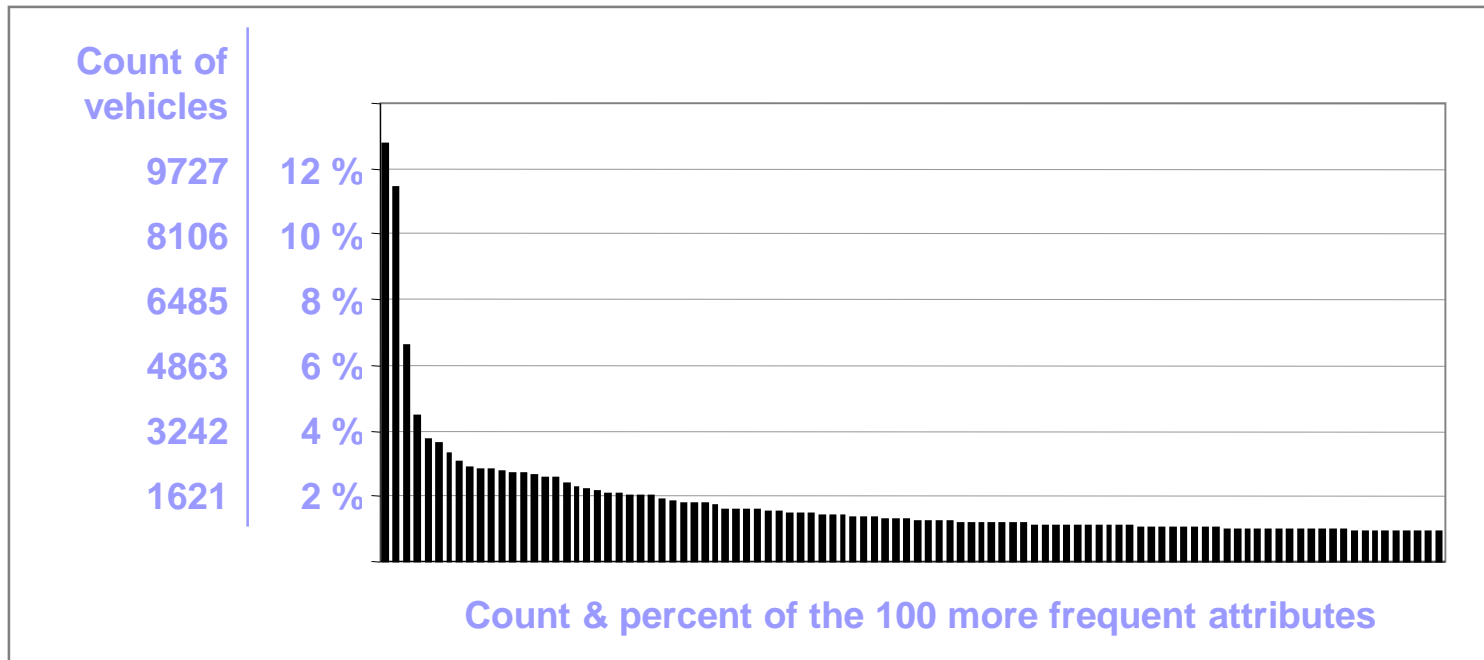
- **A measure of rules interest : "lift"**

$$lift(A \Rightarrow C) = \frac{P(A \cap C)}{P(A).P(C)}$$

- ***lift* = 2** → transactions that contain  +  +  are twice more than if the purchase of  +  and the purchase of  were independent purchases.

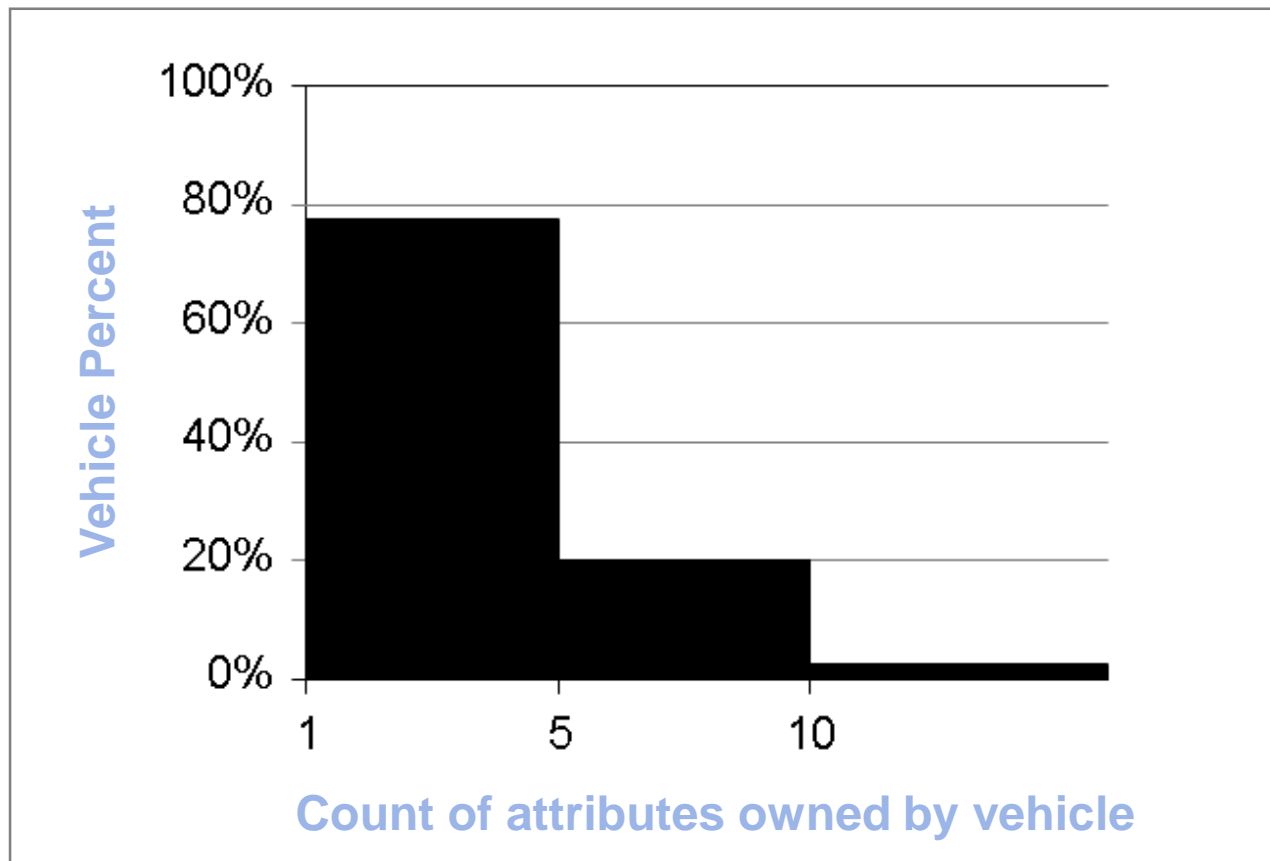
DATA FEATURE

- **Data size :**
 - More than 80 000 vehicles (\approx transactions) \rightarrow 4 months of manufacturing
 - More than 3000 attributes (\approx items)
- **Sparse data :**



DATA FEATURE

- **Count of co-occurrences per vehicle :**



OUTPUT : ASSOCIATION RULES

Minimum support (minimum count of vehicles that support the rule)	Minimum confidence	Count of rules	Maximum size of rules
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

- **Aims :**
 - Reduce count of rules
 - Reduce size of rules

- **A first reduction is obtained by manual grouping :**

Minimum support	Minimum confidence	Count of rules	Maximum size of rules
100	50 %	600636	12

Overview

Motivation

Association rules mining

→ Clustering techniques

Combined use of the two methods & application

Conclusion & future work

CLUSTERING OF VARIABLES

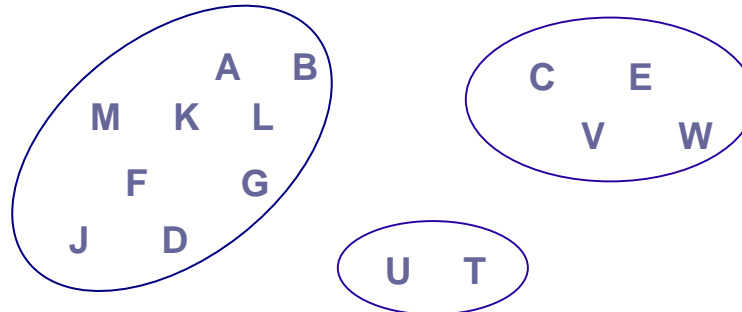
- **Input data : binary matrix**

More than 3000 columns

More than 80000 rows

Vehicles	Attribute 1	Attribute 2	...	Attribute p
1	1	0		1
2	0	0		1
...				
n	0	1		0

- **Clustering of variables aim at grouping attributes into a limited number of homogenous clusters**



- Motivation
- Association rules mining
- Clustering techniques
- Combined use of the two methods & application
- Conclusion & future work

PROXIMITY MEASURES

Vehicle	<i>j</i>	<i>j'</i>
1	1	0
2	0	1
3	0	0
...		
<i>n</i>	1	1



		Variable <i>j'</i>		
		1	0	
Variable <i>j</i>	1	n_{11}	n_{10}	$n_{1.}$
	0	n_{01}	n_{00}	$n_{0.}$
		$n_{.1}$	$n_{.0}$	$\Sigma = n$

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

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

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

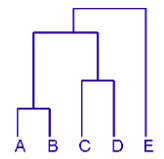
$$\text{Dice } s_D(jj') = \frac{2n_{11}}{2n_{11} + n_{10} + n_{01}}$$

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

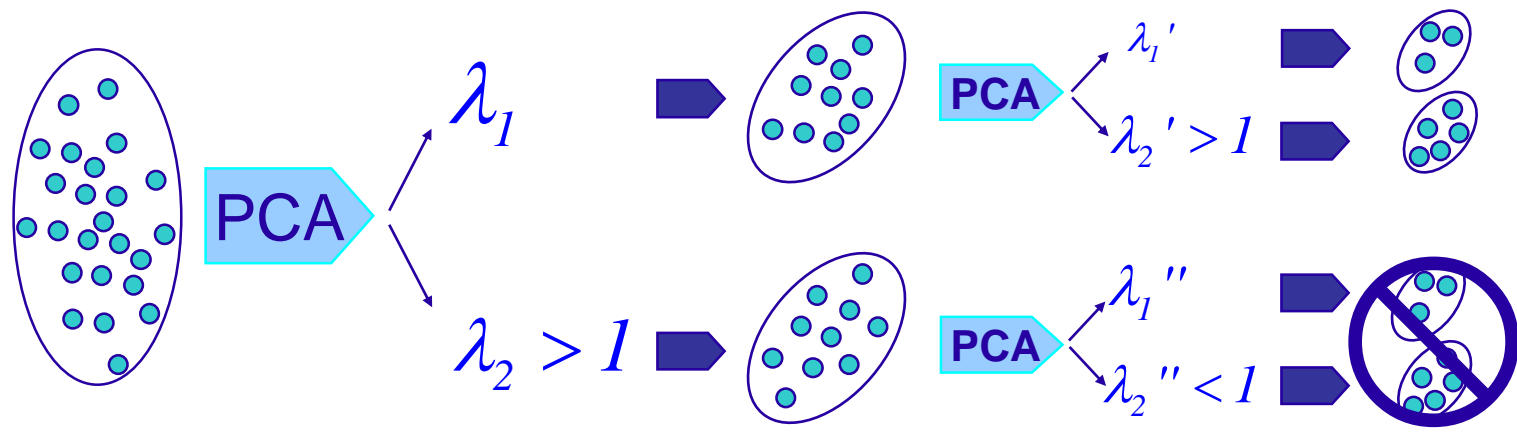
$$S_{RusselRao}(jj') = \frac{n_{11}}{n} = \frac{\text{co-occurences count}}{\text{total count}} = \text{support}(\{j, j'\})$$

CLUSTERING TECHNIQUES USED

- **Agglomerative hierarchical clustering procedure**
 - Provides hierarchical clusters : dendrogram
 - Ward's method



- **Divisive clustering procedure (Proc Varclus)**
 - Provides one-dimensional clusters
 - Assigns each variable to the component with which it has the higher squared correlation



NUMBER OF VARIABLES IN CLUSTERS

- **Number of variables per clusters :**

		Clusters									
Coef.	R²	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

- **Remark on the choice of the clusters number**

COMPARISON OF PARTITIONS

- **Paired comparison of partitions thanks to Rand's coefficient :**

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

	Ward-R ²	Ward-Ochiai	Ward-Jaccard	Ward-Dice	Ward-Russel Rao	Varclus
Ward-R ²						
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	

Percent of pairs in agreement :

- 2 variables that are clustered together in the two partitions

- 2 variables that are clustered in different clusters in the two partitions

Overview

Motivation

Association rules mining

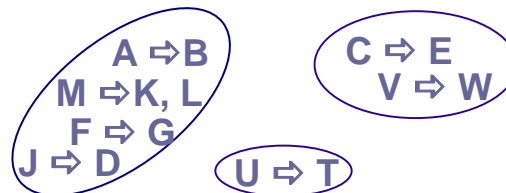
Clustering techniques

→ Combined use of the two methods & application

Conclusion & future work

COMBINED USE AND APPLICATION

- **Mining association rules inside each clusters :**



Test	Count of rules	Maximum size of rules	Reduction of the count of rules
without Clustering	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 %

- **At first sight, there's no improvement**

DETECTION OF ATYPICAL CLUSTER

- **10-clusters partition with agglomerative clustering and Russel Rao coefficient**

Cluster	Number of variables in the cluster	Number of rules found in the cluster	Maximum size of rules
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

- **Cluster 2 is the same whatever the clustering procedure**
 - **it produces many complex rules**

RESULTS

- **Mining association rules inside each clusters except atypical cluster :**

	Count of rules	Maximum size of rules	Reduction of the count of rules
Without clustering	600636	12	.
Ward - R²	43	4	+ de 99 %
Ward - Jaccard	479	5	
Ward - Russel & Rao	218	4	
Ward - Ochiai	459	5	
Ward - Dice	478	5	
Varclus	21	4	

- **The number of rules to analyse has significantly decreased**
- **The output rules are more simple to analyse**
- **Clustering has detected an atypical cluster of attributes to treat separately**

Overview

Motivation

Association rules mining

Clustering techniques

Combined use of the two methods & application

→ Conclusion & future work

CONCLUSION & FUTURE WORK

- **Previous clustering of variables provides to :**
 - Point an atypical cluster of attributes to analyse separately
 - Decrease the number of generated rules
 - Decrease their complexity
- **The choice of Russel Rao is coherent because of his link with the support**
- **Current and future works :**
 - Adapt of Qannari & Vigneau method to binary data
 - Study of the different measures of rules relevancy
 - Apply simultaneous clustering of rows and columns

REFERENCES

- 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.
- Hébrail G., Lechevallier Y. (2003) Data mining et analyse des données. In : Govaert G. Analyse des données. Ed. Lavoisier, Paris, pp 323-355
- 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
- Nakache J.P., Confais J. (2005) *Approche pragmatique de la classification*, Ed. Technip, Paris
- Gower J.C., Legendre P. (1986) *Metric and euclidean properties of dissimilarity coefficients*. In : Journal of Classification Vol.3, pp 5-48
- Youness G., Saporta G. (2004) *Some Measures of Agreement Between Close Partitions* - Student vol. 5(1), pp. 1-12.