# Méthodes de classification pour l'extraction de règles

**Marie Plasse[* **], Ndeye Niang[*], Gilbert Saporta[*], Alexandre Villeminot[**], Laurent Leblond[**]**

** CNAM Laboratoire CEDRIC*
*292 Rue St Martin Case 441*
*75141 Paris Cedex 03, France*

*** PSA Peugeot Citroën*
*45 rue Jean-Pierre Timbaud*
*78307 Poissy Cedex, France*

*RÉSUMÉ. Cette communication présente une comparaison de la classification de variables et de la classification croisée utilisées préalablement à la recherche de règles d'association sur un jeu de données industrielles.*

*MOTS-CLÉS : Classification de variables, classification croisée, règles d'association, événements rares*

## 1 Introduction

Les données que nous analysons sont issues de la fabrication automobile où plusieurs dizaines de milliers de véhicules sont décrits par plusieurs milliers de variables binaires rares. La recherche de règles d'association entre ces attributs conduit à une profusion de règles. Une classification des variables préalable nous a permis de regrouper les attributs, puis d'orienter la recherche de règles afin d'en diminuer le nombre. Dans cette communication nous étudions, sur un petit échantillon de nos données, les apports d'une classification croisée à cette approche. Après avoir présenté quelques éléments théoriques sur les méthodes utilisées, nous comparons les règles obtenues après classification de variables d'une part et classification croisée d'autre part.

## 2 Quelques éléments théoriques sur les méthodes utilisées

### *2.1 La recherche de règles d'association*

La méthode de recherche de règles d'association est née pour analyser les articles fréquemment achetés ensemble dans les supermarchés. Chaque sortie de caisse correspond à une transaction où plusieurs items ont été achetés simultanément. Une règle d'association est une implication $A \rightarrow C$ où l'antécédent $A$ et le conséquent $C$ sont des ensembles d'items, où $A \cap C = \varnothing$. Une règle repose sur les notions de support et de confiance. Le support est le nombre ou le pourcentage de transactions qui contiennent les items de la règle. La confiance est le pourcentage de transactions qui contiennent les items du conséquent parmi celles qui contiennent l'antécédent. Les algorithmes de recherche de règles d'association, tel que l'algorithme fondateur *Apriori* [AGR 94], procèdent en deux étapes. La première est la recherche des ensembles d'items fréquents dont le support est supérieur à un seuil fixé par l'utilisateur. A partir de ces ensembles, la

Avec tous nos remerciements à M. Nadif pour son aide précieuse sur la classification croisée.

seconde étape est l'extraction des règles dont la confiance est jugée suffisante par l'utilisateur. Le nombre de règles extraites étant souvent important, pour sélectionner les plus intéressantes, il est utile de les classer par ordre décroissant de leur intérêt statistique au sens d'un indice de pertinence. De nombreux indices ont été proposés tels que le lift $\left(P(A \cap C) / P(A).P(C)\right)$ [BRI 97] qui est facilement interprétable. Le choix d'un indice plutôt qu'un autre dépend du contexte ; aussi, dans le cadre de notre application, l'indice de Jaccard $\left(P(A \cap C) / P(A \cup C)\right)$ discrimine le mieux les règles qui nous intéressent [PLA 06]. Il nous est donc possible de sélectionner les règles les plus pertinentes grâce à cet indice.

### *2.2 Classification de variables*

Comme pour la classification d'individus, il existe deux grandes familles de méthodes de classification de variables : des méthodes de partitionnement (telles que *Varcha* [VIG 03]) et des méthodes hiérarchiques. Dans cette seconde famille, la méthode descendante (procédure *Varclus* de SAS) recherche des classes unidimensionnelles décrites par une seule composante principale et les méthodes ascendantes conduisent à une hiérarchie de partitions emboîtées de l'ensemble des variables. Ces dernières reposent sur le choix d'une stratégie d'agrégation et d'un indice de similarité entre les variables. Le $\Phi^2$ de Pearson, l'indice de Jaccard ou encore celui de Russel-Rao sont des indices adaptés au cas binaire. L'utilisation conjointe de la classification de variables et des règles d'association [PLA 05] permet de faire face à la profusion de règles obtenue avec une recherche classique des règles. La classification de variables permet de construire des classes homogènes d'attributs. La recherche de règles d'association à l'intérieur de chacune de ces classes est pertinente car il est facilement possible d'identifier les classes où les attributs sont très corrélés et produisent donc de nombreuses règles. L'ensemble d'associations, plus restreint, est plus simple à analyser.

### *2.3 Classification croisée*

L'objectif de la classification croisée est de trouver une paire de partitions (**z**,**w**), où **z** est une partition de l'ensemble *I* des *n* individus en *K* classes et **w** est une partition de l'ensemble *J* des *m* variables en *H* classes, *K* et *H* étant connus. Ce problème est résolu de manière itérative par une optimisation alternée de la partition des individus en bloquant celle des variables puis de la partition des variables en fixant celle des individus. Plusieurs algorithmes ont été proposés selon le type de données, dont l'algorithme *Crobin* dans le cas binaire [GOV 83] qui propose de maximiser un critère de type inertie. Cet algorithme est rapide et donne de bons résultats lorsque les blocs ont les mêmes proportions et des degrés d'homogénéité semblables. Lorsque ce n'est pas le cas, le problème de la classification croisée peut être traité par l'approche modèle de mélange, où les données sont supposées provenir d'un mélange de plusieurs distributions de probabilité, où chaque composant du mélange correspond à une classe ([GOV 03], [GOV 05]). Le problème consiste alors à retrouver pour chaque objet sa population d'origine la plus probable en fonction du vecteur d'observations qui le caractérise. Les données observées **x** enrichies par les informations manquantes (ici les classes) constituent les données complètes. Ainsi, les données manquantes sont, d'une part le vecteur $\mathbf{z}=(\mathbf{z}_1,...,\mathbf{z}_i,...,\mathbf{z}_n)$ où $\mathbf{z}_i = k$ (avec $k=1...K$) est le numéro $k$ de la classe de l'individu $i$, et le vecteur $\mathbf{w}=(\mathbf{w}_1,...,\mathbf{w}_j,...,\mathbf{w}_m)$ où $\mathbf{w}_j = h$ (avec $h=1...H$) est le numéro $h$ de la classe de la variable $j$. Le modèle de mélange croisé s'écrit $f(\mathrm{x};\theta) = \sum_{(z,w)\in Z\times W} \prod p_{z_i} \prod q_{w_j} \prod \varphi_{z_i w_j}(x_i^j;\alpha_{z_i}^{w_j})$ où les densités $\varphi_{kh}$ appartiennent à la même famille, les paramètres $p_k$ et $q_h$ sont les probabilités qu'une ligne et une colonne appartiennent respectivement aux $k^{ème}$ et $h^{ème}$ composants du mélange. L'estimation du vecteur $\theta$ des paramètres $(p_1,...,p_K,q_1,...,q_H,\alpha_{11},...,\alpha_{KH})$ de ce modèle est réalisée par la méthode du maximum de vraisemblance grâce à des extensions de l'algorithme *Estimation-Maximisation*. Ainsi, l'algorithme *Bloc-CEM* [GOV 03] propose de maximiser la log-vraisemblance des données complètes. Cette approche fournit des résultats rapidement mais présente certains inconvénients, elle conduit notamment à une estimation biaisée. Plus lent mais plus fiable, l'algorithme *Bloc-EM* [GOV 05] permet de maximiser l'espérance de la log-

vraisemblance des données complètes, conditionnellement aux données observées **x** et à l'estimation courante de $\theta$. Dans le cas des données binaires, la distribution de probabilités est la distribution de Bernoulli $\varphi_{kh}\left(x_i^j;\alpha_k^h\right)=\left(\alpha_k^h\right)^{x_i^j}\left(1-\alpha_k^h\right)^{\left(1-x_i^j\right)}$. Après initialisation, une première étape, où la partition sur les colonnes est fixée, est constituée d'une phase Estimation où sont calculées les probabilités a posteriori qu'un individu *i* appartiennent à une classe *k*. Vient ensuite la phase Maximisation où sont déduites les proportions $p_k$ des composants du mélange et les probabilités $\alpha_k^h$ de prendre la valeur "1" dans le bloc (*k,h*). Une seconde étape, où la partition en ligne est bloquée, estime les probabilités a posteriori qu'une variable *j* soit dans la classe *h*. La phase de maximisation attribue ensuite les proportions $q_h$ de chaque classe *h* ainsi que de nouvelles probabilités $\alpha_k^h$. Ces deux étapes sont répétées jusqu'à la convergence. La recherche de règles d'association dans des blocs homogènes où la plupart des véhicules présentent les mêmes attributs permet en outre de diminuer l'espace de recherche. En effet, les blocs de "0" sont ignorés et l'interprétation des blocs entiers de "1" est triviale et elle ne nécessite pas d'effectuer une recherche d'associations.

## 3 Classification de variables vs classification croisée sur une application

### *3.1 Recherche de règles d'association sans classification préalable*

Les données constituent un échantillon de 727 véhicules décrits par la présence ou l'absence de 109 attributs. La matrice de données binaires est clairsemée puisqu'elle ne comprend que 2,9% de "1". Un véhicule possède en moyenne 3,2 attributs et l'attribut le plus fréquent apparaît sur environ 10% des véhicules mais 80% des attributs apparaissent sur moins de 1% des véhicules.
Les premières règles trouvées ont un support de 50 véhicules mais elles ont des confiances faibles. Les attributs étant rares, il est préférable de fixer un seuil très bas pour le support et d'être plus sévère au niveau de la confiance. De plus, pour sélectionner le plus de règles pertinentes, nous pouvons fixer un seuil minimum pour l'indice de Jaccard. Avec un tel paramétrage nous espérons obtenir des règles fiables sur des événements rares. Le Tableau 1 montre les résultats obtenus avec une confiance de 90% et des seuils différents pour le support et l'indice de Jaccard. Le nombre de règles à analyser est trop important.

| Support minimum | Confiance minimum | Jaccard minimum | Nombre d'ensembles fréquents | Nombre de Règles |
|---|---|---|---|---|
| 30 véhicules | 90% | 0 | 1 230 | 39 867 |
| 30 véhicules | 90% | 0,9 | 1 230 | 21 254 |
| 10 véhicules | 90% | 0 | 65 583 | 26 210 753 |
| 10 véhicules | 90% | 0,6 | 65 583 | 11 839 141 |
| 10 véhicules | 90% | 0,9 | 65 583 | 10 127 600 |

**Tableau 1 : Règles obtenues sans classification préalable**

### *3.2 Classification de variables préalable à la recherche de règles*

Le dendrogramme résultant d'une classification ascendante hiérarchique avec la stratégie de Ward et l'indice de Russel-Rao préconise une partition des variables en 2 ou 5 classes. Nous présentons les résultats obtenus sur deux classes, sachant que la première classe des partitions en 2 et 5 classes est identique et que les autres classes sont regroupées en une seule. Les résultats de la recherche de règles d'association à l'intérieur des deux classes avec un support minimum de 10 véhicules et une confiance minimum de 90% sont présentés dans le Tableau 2.

| Classe | Nombre de variables | Pourcentage de "1" | Nombre d'ensembles fréquents | Nombre de règles | | |
|---|---|---|---|---|---|---|
| | | | | *Jaccard ≥ 0,9* | *Jaccard ≥ 0,6* | *Jaccard ≥ 0* |
| 1 | 16 | 13 | 65535 | 10160318 | 11839140 | 26210797 |
| 2 | 93 | 1 | 36 | 0 | 1 | 2 *(jac ≥ 0,55)* |

**Tableau 2 : Composition des classes et règles produites**

Les règles de la classe 1 concernent 11 attributs très corrélés : il sont présents simultanément sur 16 véhicules, ce qui explique le nombre élevé de règles. Les deux règles isolées dans la classe 2 sont assez

intéressantes du point de vue de l'indice de Jaccard. La classification préalable permet de découvrir des associations sur des items plus rares. En effet, sans classification, les items les plus fréquents créent une profusion des règles qui noie les résultats et empêche de voir les associations intéressantes.

### *3.3 Classification croisée préalable à la recherche de règles*

Une classification hiérarchique sur les individus permet d'avoir une idée du nombre de classes à fixer en ligne. Le Tableau 3 montre les résultats obtenus avec 3 classes en ligne, 2 classes en colonne et le même paramétrage qu'avec la classification de variables (le bloc 6 ne figure pas car il ne contient que des "0").

| Bloc | Pourcentage de "1" | Nombre d'individus | Nombre de variables | Nombre d'ensembles fréquents | Nombre de règles | | |
|---|---|---|---|---|---|---|---|
| | | | | | *Jaccard ≥ 0,9* | *Jaccard ≥ 0,6* | *Jaccard ≥ 0* |
| 1 | 8,8 | 682 | 15 | 29 | 0 | 0 | 0 |
| 2 | 1,2 | 682 | 94 | 30 | 0 | 0 | 1 *(jac=0,55)* |
| 3 | 100 | 29 | 15 | 32767 | 14283372 | 142283372 | 142283372 |
| 4 | 2,86 | 29 | 94 | 1 | 0 | 0 | 0 |
| 5 | 48,3 | 16 | 15 | 63 | 602 | 602 | 602 |

**Tableau 3 : Composition des blocs et règles produites**

Le bloc 3 est intégralement constitué de "1" : les 15 attributs sont présents simultanément sur les 29 véhicules du bloc. Les 14 millions de règles issues de ce bloc sont porteuses d'une seule et même information. Dans le bloc 5, la plupart des règles sont provoquées par la présence de 6 attributs sur 13 des 16 véhicules. Enfin, la règle du bloc 2 avait été détectée grâce à la classification de variables également.

## 4 Conclusion

Les deux approches, simple et croisée, conduisent à une réduction du nombre de règles, une fois les groupes analysés. Elles permettent d'identifier puis d'isoler les groupes d'attributs fortement liés, et enfin d'orienter la recherche de règles vers des groupes moins homogènes où des associations moins évidentes seront découvertes. La classification croisée fournit une partition des données plus fine et intéressante. De plus, elle présente l'avantage de pouvoir exclure les blocs totalement homogènes de la recherche de règles, ce qui peut se révéler très utile sur des données de taille importante. Cette approche, prometteuse sur un échantillon restreint, va donc être utilisée sur la base entière comportant plusieurs dizaines de milliers de véhicules et des milliers de variables.

## 5 Bibliographie


[AGR 94] AGRAWAL R., SRIKANT R. *Fast Algorithms for Mining Association Rules*. Proceedings of the 20th Conference on Very Large Databases, Santiago, Chile, pp. 487-499, 1994.

[BRI 97] BRIN S., MOTWANI R., SILVERSEIN C. *Beyond market baskets: generalizing association rules to correlations*. Proceedings of the ACM-SIGMOD Conference on Management of Data, Tucson, Arizona, USA, 1997.

[GOV 83] GOVAERT G. Classification croisée, Thèse d'Etat, Université Paris 6, France, Juin 1983.

[GOV 03] GOVAERT G., NADIF M. Clustering with block mixture models. Pattern Recognation, 36(2) : pp. 463-473, 2003.

[GOV 05] GOVAERT G., NADIF M. *An EM Algorithm for the Block Mixture Model.* IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 643-647, 2005.

[PLA 05] PLASSE M., NIANG N., SAPORTA G. Utilisation conjointe des règles d'association et de la classification de variables. Journées Françaises de Statistique, Pau, France, 2005.

[PLA 06] PLASSE M., NIANG N., SAPORTA G., LEBLOND L. *Une comparaison de certains indices de pertinence des règles d'association.* Revue des Nouvelles Technologies de l'Information, Actes 6e Conférence Extraction et Gestion des Connaissances, EGC'06, Série E, n°6, Vol.II, pp.561-568, Lille, janvier 2006.

[VIG 03] VIGNEAU E., QANNARI E.M. *Clustering of variables around latent component - application to sensory analysis*. Communications in Statistics , Simulation and Computation, 32(4) : pp. 1131-1150, 2003.