Classification de variables et classification croisée utilisées préalablement à la recherche de règles d'association

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

* CNAM Laboratoire CEDRIC 292 Rue St Martin Case 441 Paris Cedex 03
niang@cnam.fr
saporta@cnam.fr
** PSA Peugeot Citroën, Zone d'Activité Louis Bréguet, 78943 Vélizy Villacoublay
marie.plasse@mpsa.com
alexandre.villeminot@mpsa.com
laurent.leblond1@mpsa.com

Résumé. La recherche de règles d'association conduit souvent à l'obtention d'un très grand nombre de règles, alors inexploitables. De plus, il est parfois difficile de faire varier les paramètres d'extraction des règles : le support et la confiance. En effet, dans le cas où les variables sont des événements rares, il est nécessaire de choisir des seuils de support très faibles. Nous avons proposé d'utiliser de manière conjointe la classification de variables et la recherche de règles d'association. La classification préalable des variables permet de construire des groupes homogènes où les variables sont liées. La recherche de règles à l'intérieur de chaque groupe conduit à réduire le nombre de règles à analyser. Les techniques de classification croisée permettent un double partitionnement sur les variables et sur les individus. Nous souhaitons étudier les apports d'une telle classification à notre approche. Cet article présente une comparaison des deux types de classification utilisés préalablement à la recherche de règles d'association. Nous présentons les résultats obtenus sur plusieurs échantillons de données issues de l'industrie automobile.

1 Introduction

Ce travail a pour objectif la découverte d'éventuels liens entre variables ou groupes de variables binaires représentant des événements rares au sein d'une base de données industrielle de taille importante. Plusieurs dizaines de milliers d'individus sont décrits par la présence ou l'absence de plusieurs milliers d'attributs. Les données pouvant se présenter sous la forme d'un tableau de données de transactions, une idée naturelle consiste à utiliser la méthode de recherche de règles d'association. Cependant, le nombre élevé de variables conjugué à la rareté des occurrences conduit à un très grand nombre de règles dont les supports sont très faibles et les confiances très élevées.

Nous avons proposé de réaliser une classification préalable des variables afin de construire des groupes homogènes d'attributs à l'intérieur desquels la recherche de règles d'association est plus pertinente. Cette approche appliquée à nos données nous a permis de diminuer de manière très significative le nombre et la complexité des règles obtenues.

Les techniques de classification croisée suscitent de plus en plus d'intérêt, notamment en bioinformatique où le nombre de variables est souvent très important. Ces méthodes conduisent, par permutation des lignes et des colonnes de la matrice initiale, à des blocs homogènes de données où les individus ont des profils semblables au regard des variables qui les décri-

vent. Dans cet article nous étudions, sur plusieurs échantillons de données, les apports d'une classification croisée préalable à la recherche de règles d'association.

Tout d'abord, nous présentons quelques éléments théoriques sur les méthodes utilisées. Ensuite, nous illustrons la recherche de règles d'association suite à une classification préalable des variables sur un jeu de données clairsemées, de taille importante, issu de l'industrie automobile. Enfin, nous comparons cette approche avec l'utilisation d'une classification croisée à la place de la classification de variables, sur plusieurs échantillons de taille réduite.

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, avec $A \cap C = \emptyset$. Une règle repose sur les notions de support et de confiance. Le support est le nombre ou le pourcentage de transactions qui contiennent tous les items de la règle. La confiance est le pourcentage de transactions qui contiennent les items du conséquent parmi celles qui contiennent l'antécédent.

En 1993 Agrawal et al. proposent les tous premiers algorithmes de recherche d'association : *AIS* et *SETM*. Ces derniers sont remis en question l'année suivante par les mêmes auteurs qui présentent *Apriori* (Agrawal et Srikant, 1994), considéré comme l'algorithme fondateur de la recherche d'association.

Les algorithmes de recherche de règles d'association, tel que *Apriori*, 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 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 $P(A \cap C)/P(A).P(C)$ proposé par Brin et al. (1997) 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 $P(A \cap C)/P(A \cup C)$ discrimine le mieux les règles qui nous intéressent (Plasse et al., 2006). Il nous est donc possible de sélectionner les règles les plus pertinentes grâce à cet indice.

2.2 La 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* de Vigneau et Qannari (2003), et des méthodes hiérarchiques. Dans cette seconde famille, la méthode descendante (procédure *Varclus* de SAS, 2003) 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 Φ^2 de Pearson, l'indice de Jaccard ou encore celui de Russel-Rao sont des indices adaptés au cas binaire (Nakache et Confais, 2005).

Nous avons montré que l'utilisation conjointe de la classification de variables et des règles d'association permet de faire face à la profusion de règles obtenues avec une recherche classique des règles (Plasse et al., 2005). 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* développé pour le cas binaire par Govaert (1983), 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, Govaert et Nadif (2003, 2005) proposent de traiter le problème de la classification croisée par l'approche modèle de mélange, où les données sont supposées provenir d'un mélange de plusieurs distributions de probabilités, où chaque composant du mélange correspond à une classe.

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 $\mathbf{z}=(\mathbf{z}_1,...,\mathbf{z}_i,\ldots,\mathbf{z}_n)$ où $\mathbf{z}_i=k$ (avec $k=1...K$) est le numéro k de la classe de l'individu i, et le vecteur $\mathbf{w}=(\mathbf{w}_1,...,\mathbf{w}_j,\ldots,\mathbf{w}_m)$ 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 φ_{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 θ 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 *Espérance-Maximisation (EM)*.

Ainsi, l'algorithme *Bloc-CEM* (Govaert et Nadif, 2003) 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* (Govaert et Nadif, 2005) permet de maximiser l'espérance de la log-vraisemblance des données complètes, conditionnellement aux données observées **x** et à l'estimation courante de θ. Dans le cas des données binaires, $x_i^j=1$ si l'individu i possède l'attribut j et $x_i^j=0$ sinon. La distribution de probabilités utilisée est la distribution de Bernoulli $\varphi_{kh}(x_i^j;\alpha_k^h)=(\alpha_k^h)^{x_i^j}(1-\alpha_k^h)^{(1-x_i^j)}$.

Après initialisation, une première étape, où la partition sur les colonnes est fixée, est constituée d'une phase *Espérance* où sont calculées les probabilités a posteriori qu'un individu i appartienne à une classe k. Vient ensuite la phase *Maximisation* où sont déduites les proportions p_k des composants du mélange et les probabilités α_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 α_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 préalable à la recherche de règles d'association

Nos données concernent un ensemble de plus de 80000 véhicules décrits par plus de 3000 attributs binaires rares, représentant des événements de fabrication relevés sur une période de quatre mois. Ces événements sont relatifs aux trois ateliers d'une usine de production : le ferrage, la peinture et le montage. La matrice contient seulement 0,13% de "1". L'attribut le plus fréquent apparaît sur seulement 12% des véhicules mais 97% des attributs apparaissent sur moins de 1% des véhicules. Enfin, un véhicule présente en moyenne 4 attributs.

Les attributs étant rares, il est préférable de fixer un seuil très bas pour le support. Comme le montre le tableau 1, une faible modification à la baisse du support implique une très forte augmentation du nombre de règles produites ainsi que de leur complexité.

Support minimum	Confiance minimum	Nombre d'ensembles fréquents	Nombre de règles	Taille maximum des règles obtenues
500 OF	50 %	188	18	3
400 OF	50 %	256	31	3
300 OF	50 %	389	213	5
200 OF	50 %	2398	86836	9
100 OF	50 %	7704	600632	11

TAB. 1 – Recherche de règles avec différents seuils pour le support minimum.

Un premier moyen de réduire le nombre de règles est d'augmenter le seuil de la confiance minimum. Le tableau 2 montre que cela est insuffisant car une confiance de 99% conduit encore à des dizaines de milliers de règles.

Confiance minimum	Nombre de règles
50 %	600632
80 %	416312
90 %	240362
99 %	60841

TAB. 2 – Recherche de règles avec différents seuils pour la confiance minimum.

Nous avons montré (Plasse et al., 2005) que la recherche de règles d'association à l'intérieur des classes obtenues suite à une classification préalable des variables conduit à une sélection efficace des règles produites. En effet, la classification permet d'identifier des groupes d'attributs très liés dont les combinaisons multiples produisent de nombreuses règles. Une fois ces groupes isolés, il est possible de découvrir des liens moins évidents entre des attributs plus rares. Ces liens sont indétectables lorsque la recherche est menée sur l'ensemble de la base. Le tableau 3 montre la réduction du nombre de règles selon l'utilisation de plusieurs méthodes de classification.

		Nombre de règles	Complexité maximum	Réduction du nombre de règles
Sans classification préalable		*600636*	*12*	-
Classifications ascendantes Méthode de Ward	Indice R^2	43	4	+ de 99%
	Indice de Jaccard	479	5	
	Indice de Russel Rao	218	4	
	Indice de Ochiai	459	5	
	Indice de Dice	478	5	
Classification descendante	Procédure Varclus	165	4	

TAB. 3 - Réduction du nombre et de la complexité des règles grâce à la classification préalable des variables.

Suite à une classification ascendante hiérarchique avec la stratégie de Ward et l'indice de Russel-Rao par exemple, il reste 218 règles d'association à analyser. Il est alors possible de sélectionner les règles les plus intéressantes grâce à un indice de pertinence, puis de les faire analyser et éventuellement valider par un expert du terrain.

4 Comparaison de la classification de variables et de la classification croisée pour la recherche de règles d'association

Dans le cadre de notre application où la recherche de règles d'association conduit à un nombre trop élevé de résultats, une classification de variables préalable est une approche intéressante car elle permet de partager de manière adroite l'espace de recherche. Aussi l'idée d'utiliser une classification croisée des individus et des variables, qui mène à un double partitionnement des données, nous a semblé prometteuse. Cette partie est donc consacrée à la comparaison des résultats obtenus grâce à la classification de variables d'une part et grâce à la classification croisée d'autre part.

4.1 Recherche de règles d'association sans classification préalable

Dans cette partie, l'analyse porte sur un échantillon de 727 véhicules décrits par 109 attributs relatifs à un secteur restreint de l'usine : la partie Habillage Caisse de l'atelier Montage. Cet échantillon est représentatif de l'ensemble de la base entière car il présente les mêmes caractéristiques en termes de fréquences. En effet, un véhicule possède en moyenne 3,2 attributs. L'attribut le plus fréquent apparaît sur environ 10% des véhicules mais 80% des attributs apparaissent sur moins de 1% des véhicules (FIG. 1). Par contre, la matrice est un peu moins clairsemée car elle contient 2,9% de "1".

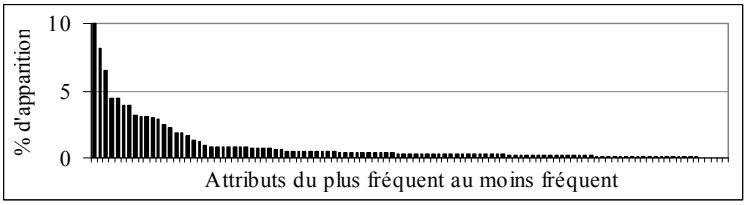

FIG. 1 – Répartition des attributs.

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 qui permet d'évaluer le degré de pertinence des règles. Avec un tel paramétrage nous espérons obtenir des règles fiables sur des événements rares. Le tableau 4 montre les résultats obtenus avec une confiance de 90% et des seuils différents pour le support et l'indice de Jaccard. Dans tous les cas, 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

TAB. 4 - Règles obtenues sans classification préalable.

4.2 Classification de variables préalable à la recherche de règles

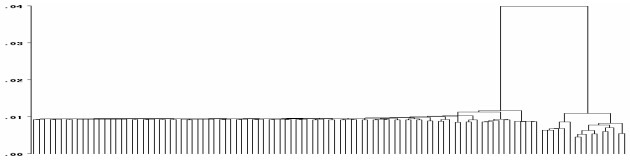

FIG. 2 – Dendrogramme des variables.

Le dendrogramme (FIG. 2) résultant d'une classification ascendante hiérarchique avec la stratégie de Ward et l'indice de Russel-Rao suggère une partition des variables en 2 classes. La figure 3 montre comment l'espace initial de recherche des règles d'association est transformé en deux espaces distincts et plus homogènes, après permutation et regroupement des colonnes selon la partition engendrée par la classification des variables.

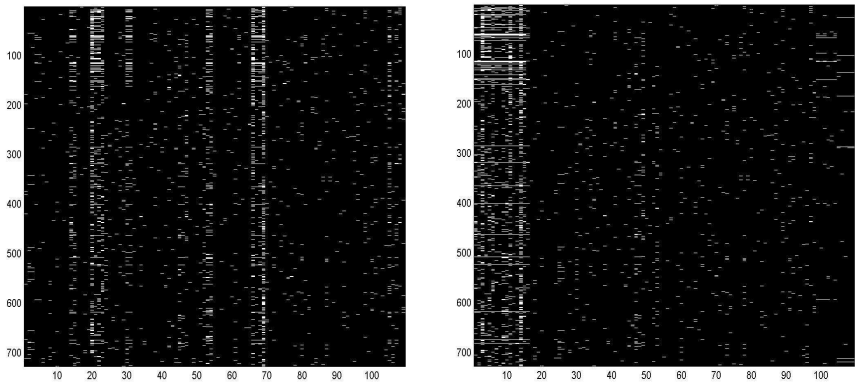

FIG. 3 – Matrice des données originales (à gauche) et matrice des données réorganisées en fonction de la classification de variables (à droite). Un "1"(présence) est symbolisé par une marque blanche et un "0" (absence) par une marque noire.

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 5.

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)*

TAB. 5 - 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.

4.3 Classification croisée préalable à la recherche de règles

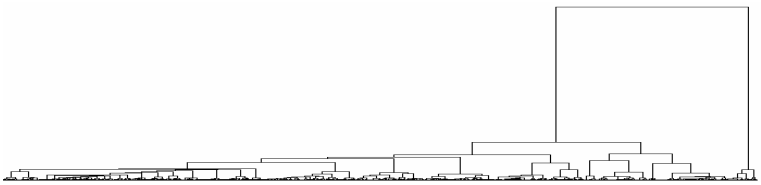

FIG. 4 – Dendrogramme des individus.

Une classification hiérarchique sur les individus permet d'avoir une idée du nombre de classes à fixer en ligne (FIG. 4). La classification croisée cherche les plus homogènes possibles de "1" et de "0". Certains sont visibles sur la figure 5 grâce au regroupement des lignes en 3 classes et des colonnes en 2 classes.

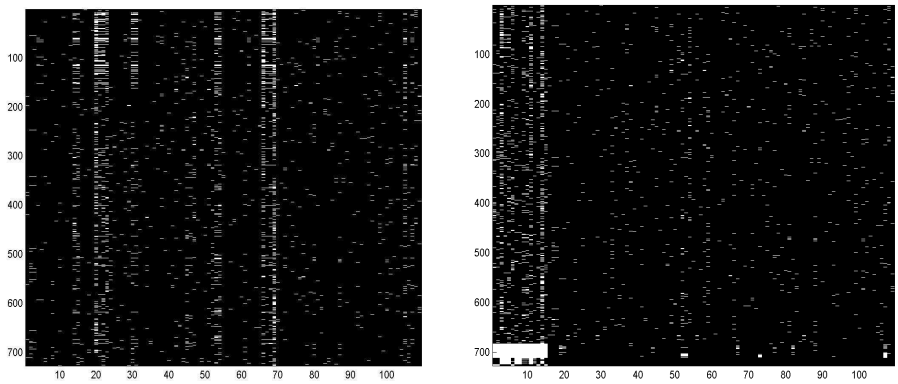

FIG. 5 - Matrice des données originales (à gauche) et matrice des données réorganisées en fonction de la classification croisée (à droite).

Le tableau 6 montre les résultats obtenus avec 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

TAB. 6 - 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 donc 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.

Même si les tableaux de cet exemple ne traduisent pas une réduction du nombre de règles, la classification préalable induit une réelle simplification de l'analyse des résultats. Les deux techniques de classification permettent d'identifier des formes atypiques à isoler pour découvrir ce qu'elles masquent.

4.4 Comparaisons sur d'autres échantillons réels et simulés

Nous avons étudié les résultats de la recherche de règles d'association sans classification préalable, après classification des variables et après classification croisée. Pour comparer ces trois méthodes sur d'autres jeux de données, nous avons construit des échantillons aléatoires à partir des données relatives aux ateliers ferrage et peinture d'une part ("FP"), et à l'atelier montage d'autre part. Comme la base originale, ces données sont clairsemées. Par ailleurs, nous avons simulé des données moins clairsemées à partir de modèles de mélange. Le tableau 7 décrit les caractéristiques des échantillons dont nous présentons les résultats dans ce paragraphe.

Echantillons	Nombre d'individus	Nombre de variables	Pourcentage de "1"
"FP 1"	1392	235	1,3%
"FP 2"	1424	234	1,1%
"Montage 1"	1141	178	1,9%
"Montage 2"	1229	211	1,6%
"Simulé 1"	2500	100	11,5%
"Simulé 2"	2500	100	9,6%

TAB. 7 – Taille et pourcentage de "1" dans les échantillons traités.

Les tableaux 8 et 9 en annexe présentent, pour chaque échantillon, le nombre d'ensembles fréquents et de règles d'association obtenus avec une recherche globale sur tout l'échantillon, à l'intérieur des classes après une classification de variables et à l'intérieur des blocs après une classification croisée.

En ce qui concerne les deux échantillons issus des ateliers ferrage et peinture ("FP"), les résultats d'une recherche de règles sans classification préalable sont tout à fait satisfaisants car il n'y a que quelques dizaines de règles à analyser. Cependant, la classification, qu'elle soit simple sur les variables ou croisée, permet quand même de réduire le nombre de règles en sortie.

A l'opposé, les échantillons issus du montage génèrent beaucoup trop de règles, il est donc pertinent d'utiliser l'approche d'une classification préalable. Comme dans l'application

sur l'échantillon "habillage caisse" détaillée dans les paragraphes précédents, quelle que soit la classification utilisée, elle permet d'identifier à chaque fois le groupe d'attributs très liés qui produit des règles en grand nombre. Ces règles se résument à une même information et ne nécessitent ainsi qu'une seule interprétation. Les classifications préalables permettent alors d'isoler les règles cachées.

Enfin, les échantillons simulés conduisent aussi à un nombre trop important de règles. La classification préalable des variables ne parvient pas à en réduire le nombre car les mêmes règles se retrouvent dans une seule classe. En revanche, sur le premier échantillon, la classification croisée conduit à une réduction très importante du nombre de règles (+ de 99%). Sur le deuxième échantillon, elle permet d'écarter un bloc à très forte concentration de "1", nettement visible sur la figure 6.

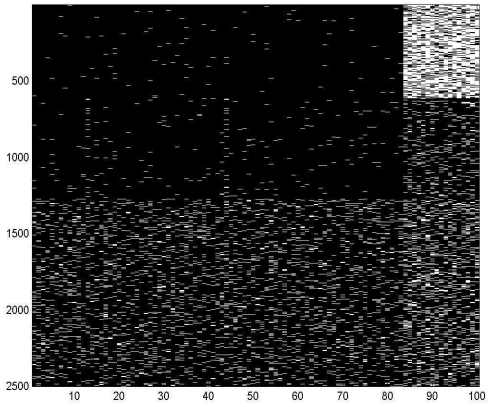

FIG. 6 - Echantillon simulé n°2 - Matrice des données réorganisées en fonction de la classification croisée (2 classes en colonnes, 3 classes en lignes).

Sur ces six échantillons, la classification croisée se montre plus appropriée que la classification de variables. Elle réalise un découpage plus fin des données, et permet ainsi l'identification de formes invisibles dans le cas où le partitionnement ne porte que sur les variables.

5 Conclusion

Dans cet article, nous avons proposé d'utiliser des méthodes de classification pour créer des groupes homogènes de données, à l'intérieur desquels la recherche de règles d'association est plus appropriée. Nous avons comparé les règles obtenues sans classification préalable, après une classification de variables et après une classification croisée. Ces comparaisons ont été menées sur plusieurs échantillons de données industrielles où des véhicules sont décrits par de nombreuses variables binaires représentant des événements rares.

Dans ce cadre, nous avons montré que les deux approches de classification, 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. L'utilisation de la classification croisée, qui fournit une double partition des données, semble la plus pertinente car elle détecte des groupes ignorés par une simple partition sur les variables. De plus, elle permet d'exclure les blocs intégralement homogènes

de la recherche de règles, ce qui peut se révéler très utile sur des données de taille importante.

Cette approche utilisant la classification croisée a été testée sur plusieurs échantillons de taille restreinte et paraît prometteuse. Nous développons un outil informatique qui nous permettra de traiter la base entière comportant plusieurs dizaines de milliers de véhicules et des milliers de variables.

Enfin, dans ces travaux, nous avons privilégié l'étude des liens entre variables. Les règles obtenues après une classification croisée préalable sont valables pour le groupe d'individus présents dans un bloc donné. Cependant, si des informations supplémentaires étaient disponibles sur les individus (modèle du véhicule, type de moteur…), il serait intéressant de trouver quelles sont les caractéristiques dominantes des individus de chaque bloc en réalisant une analyse factorielle des correspondances par exemple.

Références

Agrawal R., Imielinski T., Swani A. (1993) Mining Association rules between sets of items in large databases. *Proceedings of the ACM- SIGMOD Conference on Management of Data*, Washington DC, USA, pp.207-216

Agrawal R., Srikant R. (1994) Fast Algorithms for Mining Association Rules. *Proceedings of the 20th Int'l Conference on Very Large Databases (VLDB)*, Santiago, Chile, pp.487-499.

Brin S., Motwani R., Silversein C. (1997) Beyond market baskets: generalizing association rules to correlations. *Proceedings of the ACM-SIGMOD Conference on Management of Data*, Tucson, Arizona, USA.

Govaert G. (1983) Classification croisée, Thèse d'Etat, Université Paris 6, France

Govaert G., Nadif M. (2003) Clustering with block mixture models. *Journal of Pattern Recognition* 36, pp. 463-473

Govaert G., Nadif M. (2005) An EM Algorithm for the Block Mixture Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.27, n°4, pp.1-5

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

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

Plasse M., Niang N., Saporta G., Leblond L. (2006) 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

SAS Institute Inc (2003), *SAS/STAT User's Guide*, Cary, NC: SAS Institute Inc

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

Annexe

Echantillon	Support minimum	Confiance Minimum	Classification préalable	Classe / Bloc	Nombre d'ensembles fréquents	Nombre de règles	Totaux
"FP 1"	10 véhicules	60%	Aucune	-	120	21	21
			Variables	1	3	2	12
				2	25	1	
				3	6	9	
			Croisée	1	35	0	12
				2	15	0	
				3	8	0	
				4	50	10	
				5	0	0	
				6	7	2	
"FP 2"	10 véhicules	60%	Aucune	-	139	29	29
			Variables	1	3	2	12
				2	24	0	
				3	70	10	
			Croisée	1	23	0	19
				2	28	0	
				3	79	18	
				4	3	0	
				5	3	1	
				6	0	0	
"Montage 1"	10 véhicules	90%	Aucune	-	1462	41102	41102
			Variables	1	1290	41080	41102 (dont 22 à analyser)
				2	75	22	
				3	95	0	
			Croisée	1	57	3	14547 (dont 5 à analyser)
				2	98	1	
				3	511	14542	
				4	3	1	
				5	0	0	
				6	0	0	
"Montage 2"	10 véhicules	90%	Aucune	-	1459	39233	39233
			Variables	1	1284	39215	39233 (dont 18 à analyser)
				2	156	18	
				3	18	0	
			Croisée	1	72	10	50737 (dont 13 à analyser)
				2	97	3	
				3	1279	50724	
				4	0	0	
				5	0	0	
				6	0	0	

TAB. 8 - Recherche de règles d'association sans classification, après classification de variables (partition en 3 classes) et après classification croisée (en 6 blocs) sur échantillons de données réelles.

Méthodes de classification utilisées préalablement à la recherche de règles d'association

Echantillon	Support minimum	Confiance Minimum	Classification préalable	Classe / Bloc	Nombre d'ensembles fréquents	Nombre de règles	Totaux
"Simulé 1"	50 véhicules	60%	Aucune	-	10292	29767	29767
			Variables	1	10292	29767	29767
				2	79	0	
			Croisée	1	79	0	49
				2	23	0	
				3	0	0	
				4	0	0	
				5	0	0	
				6	52	49	
"Simulé 2"	50 véhicules	60%	Aucune	-	30569	176618	176618
			Variables	1	30485	176618	176618
				2	84	0	
			Croisée	1	0	0	177792 (dont 2 à analyser)
				2	30470	177790	
				3	3	2	
				4	17	0	
				5	83	0	
				6	67	0	

TAB. 9 - Recherche de règles d'association sans classification, après classification de variables (partition en 2 classes) et après classification croisée (en 6 blocs) sur échantillons de données simulées.

Summary

This paper proposes a way to analyse links between binary attributes in a large sparse dataset. We use clustering methods to obtain homogeneous patterns and then, to mine association rules from each pattern. Here, we compare association rules obtained after a clustering of columns on one hand, and after a simultaneous clustering of rows and columns on the other hand. Our approach is illustrated by an industrial application from the automotive industry.