Utilisation conjointe des règles d'association et de la classification de variables

Marie Plasse, Ndeye Niang Keita, Gilbert Saporta

UTILISATION CONJOINTE DES RÈGLES D'ASSOCIATION ET DE LA CLASSIFICATION DE VARIABLES

Marie Plasse[*] [**], Ndeye Niang-Keita[**] & Gilbert Saporta[**]

[*] PSA Peugeot Citroën
DINQ/DSIN/SIFA/APST
45 avenue Jean-Pierre Timbaud, 78 307 Poissy Cedex, France
marie.plasse@mpsa.com

[**] Chaire de Statistique Appliquée et CEDRIC
Conservatoire National des Arts et Métiers
292, rue Saint Martin, 75141 Paris Cedex 03, France
niang@cnam.fr, saporta@cnam.fr

RESUME
Dans cette communication, nous nous intéressons à l'application de la recherche de règles d'association dans le domaine de l'industrie automobile. Nous mettons en évidence quelques limitations liées au grand nombre et à la complexité des règles produites. Ces limitations sont essentiellement dues au nombre important de variables combiné à la rareté des événements. Nous proposons une utilisation des méthodes de classification de variables qui permettent de construire des groupes de variables plus homogènes. L'utilisation des règles d'association à l'intérieur de ces groupes où les variables sont moins nombreuses est alors plus pertinente.

MOTS-CLES
Recherche de règles d'association, Classification de variables, fouille de données.

SUMMARY
In this paper we deal with association rules mining applied to car manufacturing. We point out some limitations due to the great number and complexity of producted rules. Those limitations are related to the fact that we have a sparse data matrix with a very large number of variables. We propose to use variables clustering methods in order to build homogeneous groups of variables. Then, the association rules discovery inside each of these groups is more relevant.

KEYWORDS
Association rules discovery, variables clustering, data mining.

Ce travail a été réalisé sous la tutelle de Damien Gauthier chez PSA Peugeot Citroën, au sein du service DINQ/DSIN/SIFA/APST dirigé par Henri Lem.

1. Introduction

Dans cette communication, nous nous intéressons à la découverte de liens entre un très grand nombre de variables représentant des événements rares. Les données se présentant 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 l'utilisation de cette technique suppose de fixer certains paramètres comme le support, ce qui se révèle difficile dans le cas d'un très grand nombre d'individus où la fréquence d'un événement est rare. Nous proposons de réaliser au préalable une classification de variables afin de travailler sur des groupes homogènes de variables. Cette approche est appliquée à des données industrielles.
Tout d'abord, nous rappelons la méthode de recherche de règles d'association. Ensuite, nous exposons quelques méthodes de classification de variables. Enfin nous décrivons notre approche qui sera illustrée sur un exemple.

2. La recherche de règles d'association

Introduite par Agrawal et al. (1993), la méthode de recherche de règles d'association a été proposée pour permettre l'analyse des ventes des supermarchés afin d'extraire des règles du type : *"lorsqu'un client achète du pain et du beurre, il achète aussi du lait 9 fois sur 10"*. Comme le souligne Hébrail (2003), même si elle a été développée dans un objectif marketing, elle peut être utilisée dans tout autre domaine pour la recherche de cooccurrences fréquentes, si la structure des données s'y prête.
Une règle d'association est une implication de la forme $x \Rightarrow y$ où $x \subset I$, $y \subset I$ et $x \cap y = \varnothing$ avec l'ensemble $I = \{it_1, it_2, ..., it_i, ..., it_p\}$ de p éléments appelés items, et l'ensemble $T = \{t_1, t_2, ... t_j, ..., t_n\}$ de n éléments appelés transactions. Une règle comporte donc une partie prémisse (ou antécédent) composée d'un ensemble d'items x et une partie conclusion (ou conséquent) composée d'un ensemble d'items y disjoint de x.
La pertinence d'une règle est mesurée par son support s et sa confiance c. Il existe deux visions pour exprimer le support et la confiance. Du point de vue ensembliste, le support est le nombre de transactions contenant à la fois tous les items de x et tous les items de y, par rapport au nombre total de transactions dans T :

$$s = sup(x \Rightarrow y) = \frac{card\{t \in T / x \cup y \subseteq t\}}{card(T)}.$$

La confiance est le nombre de transactions contenant à la fois tous les items de x et tous les items de y, par rapport au nombre de transactions contenant les items de x :

$$c = conf(x \Rightarrow y) = \frac{card\{t \in T / x \cup y \subseteq t\}}{card(t \in T / x \subseteq t)}.$$

Du point de vue probabiliste, chaque sous-ensemble d'items se voit associé l'événement selon lequel la transaction contient les items de ce sous-ensemble. Le support s'exprime donc par la probabilité de réaliser simultanément les événements X et Y :

$$s = sup(X \Rightarrow Y) = P(X \cap Y) = P(Y / X) \cdot P(X)$$

où X est l'événement "la transaction contient tous les items de l'ensemble x" et Y est l'événement "la transaction contient tous les items de l'ensemble y". La confiance est égale à la probabilité de réalisation de l'événement Y sachant que l'événement X est réalisé :

$$c = conf(X \Rightarrow Y) = P(Y / X) = \frac{P(X \cap Y)}{P(X)} = \frac{sup(X \Rightarrow Y)}{sup(X)}$$

Plus le support est élevé, plus la règle est fréquente. Plus la confiance est élevée, moins il y a de contre-exemples de la règle.
L'algorithme fondateur pour la recherche de règles d'association, Apriori, a été proposé par Agrawal et al. (1994). Il a ensuite connu plusieurs extensions de ses fonctionnalités et performances. L'algorithme procède en deux étapes. Il commence par sélectionner tous les sous-ensembles fréquents dont le support est supérieur au seuil minimum, noté s_0, fixé a priori. Cette

procédure est basée sur la propriété selon laquelle un sous-ensemble d'un ensemble fréquent est forcément fréquent. L'algorithme recherche donc de manière itérative les ensembles fréquents de cardinal $k+1$ à partir des ensembles fréquents de cardinal k déterminés lors de l'itération précédente.
La deuxième étape est la recherche des règles d'association dont la confiance est supérieure à un seuil minimum, noté c_0, également fixé a priori. Pour chaque ensemble fréquent x trouvé lors de la première étape, l'algorithme génère tous les sous-ensembles non vides possibles x_a et x_b tels que $x_a \cap x_b = x$. Il construit ensuite toutes les associations possibles entre x_a et x_b. La règle $x_a \Rightarrow x_b$ est retenue si et seulement si : $\frac{support(x)}{support(x_b)} > c_0$.
Dans la mesure où le nombre de règles produites croît exponentiellement en fonction du nombre d'items, il est indispensable de pouvoir se limiter aux règles les plus intéressantes. C'est pourquoi de nombreux critères pour mesurer la qualité des règles ont été présentés dans la littérature dont le lift, l'indicateur de Piatetsky-Shapiro, l'indice de Gini, le gain informationnel, la conviction, ou encore la surprise. Cependant, le nombre élevé de telles mesures est une difficulté supplémentaire pour l'expert qui doit choisir celle qui est la plus appropriée à ses besoins. Plusieurs articles se proposent donc d'évaluer ces critères de choix, notamment Tan et al. (2002) et Lallich et Taytaud (2004).
Dans notre cas, contrairement aux produits de grande consommation, les items sont rarement présents dans les transactions et le paramétrage du support s'avère particulièrement délicat. Le support minimum est nécessairement très faible étant donnée la nature clairsemée des données, et pour différents seuils s_0 assez rapprochés, le nombre de règles passe de 0 à plus d'un million et le nombre d'items par règle, de 2 à 13. Nous proposons donc de réaliser une classification préalable des items en groupes homogènes pour optimiser l'utilisation des règles d'association.

3. Quelques méthodes de classification de variables

La classification de variables est un sujet important peu abordé dans les ouvrages classiques. En effet, dans divers domaines d'application comme le data mining, les variables mesurées sont souvent très nombreuses et il est indispensable de les réduire ou de mieux les structurer en les transformant en dimensions indépendantes ou en recherchant des groupes fortement corrélés. Les méthodes de classification de variables permettent d'atteindre ces différents objectifs.
Comme pour la classification d'individus, il existe deux grandes familles de méthodes de classification de variables. D'une part, les méthodes hiérarchiques permettent d'obtenir un arbre de classification ou une succession de partitions emboîtées de l'ensemble des variables en groupes homogènes. Elles sont elles-mêmes divisées en deux groupes : les méthodes ascendantes basées sur un algorithme agglomératif type classification ascendante hiérarchique et les méthodes descendantes reposant sur un algorithme divisif (proc VARCLUS de SAS par exemple). D'autre part, il existe des méthodes de partitionnement direct dans lesquelles le nombre de classes doit être défini à l'avance telles que VARHCA, une variante de VARCLUS proposée par Vigneau et Qannari (2003).
Les techniques de classification ascendante hiérarchique d'un ensemble de variables reposent sur le choix d'un indice de dissimilarité entre variables et d'une stratégie d'agrégation qui permet de construire un système de classes de variables de moins en moins fines par regroupement successifs. Plusieurs indices de similarité ont été proposés dans la littérature selon la nature des variables (Nakache et Confais, 2005). Pour les données binaires, un des indices usuels est le Φ^2 de Pearson obtenu à partir du Khi^2 de contingence calculé sur le tableau de contingence croisant les deux variables X_j et $X_{j'}$:

$$\Phi^2_{jj'} = \frac{\chi^2_{jj'}}{n} = \frac{n_{11}n_{22} - n_{21}n_{12}}{n_{1.}n_{2.}n_{.1}n_{.2}}.$$

On a $\Phi^2_{jj'} = r^2_{jj'}$, carré du coefficient de corrélation linéaire entre les variables indicatrices de la première modalité. On peut aussi utiliser le coefficient de corrélation linéaire. D'autres mesures de

similarité entre variables binaires peuvent également être employées : l'indice de Qannari et Vigneau (1998) basé sur l'opérateur d'Escoufier qui dans le cas de deux variables X_j et $X_{j'}$ à m et q modalités est obtenu à partir du coefficient de Tschuprow T :

$D^2(X_j, X_{j'}) = 2(1 - T^2(X_j, X_{j'}))$ avec $T^2(X_j, X_{j'}) = \dfrac{\chi^2_{jj'}}{n\{(m-1)(q-1)\}^{1/2}} = \Phi^2_{jj'}$ dans le cas binaire,

le coefficient d'affinité de Matusita utilisé par Bacelar-Nicolau (2003) : $a_{jj'} = \sum_{i=1}^{n} \dfrac{n_{ij}}{n_{.j}} \dfrac{n_{ij'}}{n_{.j'}}$ où n_{ij} est le nombre d'occurrences du couple (*individu i, variable j*) et $n_{.j}$ la marge de la colonne j, ou encre les indices de similarité de type Jaccard, Russel-Rao (Nakache et Confais, 2005)...
La classification hiérarchique nécessite la transformation de ces indices de similarité *s(i,i')* en indices de dissimilarité *d(i,i')* = $\max_{i,i'}$ *(s(i,i'))* - *s(i,i')*. Ensuite il suffit d'appliquer les mêmes stratégies d'agrégation que pour la classification d'individus (Nakache et Confais, 2005) : critère de Ward, critère du saut minimal, du diamètre, ou de la moyenne.

Les méthodes de classification descendantes fournissent des arbres hiérarchiques dont les segments terminaux constituent une partition des éléments à classer. La partition obtenue est telle que les variables d'une même classe sont le plus corrélées possible et deux variables appartenant à des classes différentes sont le moins corrélées possible. La procédure VARCLUS de SAS permet une telle méthode de classification. Elle recherche des classes unidimensionnelles, c'est-à-dire décrites par une seule composante principale. L'algorithme consite à réaliser une analyse factorielle particulière sur l'ensemble des variables et à retenir les composantes principales correspondant aux deux plus grandes valeurs propres si la seconde est supérieure à 1. Chaque variable est alors affectée à la composante principale dont elle est la plus proche au sens du carré du coefficient de corrélation linéaire, formant ainsi deux groupes de variables. Ceux-ci sont, à leur tour, divisés selon la même méthode. Ces méthodes divisives peuvent être utilisées pour réduire le nombre de variables : une fois les classes construites, il suffit de ne garder qu'un représentant par la classe, par exemple une combinaison linéaire des variables de la classe. Dans notre cas, nous les utilisons principalement dans le but de construire des groupes homogènes de variables à l'intérieur desquels nous recherchons des règles d'association.

4. Application

Les données concernent un ensemble de n=81057 véhicules décrits par p=3065 caractéristiques. Chacune est une variable booléenne X_j qui prend la valeur "1" si le véhicule possède l'attribut et "0" s'il ne le possède pas. Pour l'application des règles d'association, l'ensemble des véhicules est assimilé à l'ensemble T des transactions et les variables à l'ensemble I des items.
De simples calculs de fréquences permettent de mettre en évidence quelques particularités des données. La matrice de données est clairsemée : elle contient seulement environ 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. De plus, un véhicule possède en moyenne moins de 4 attributs. En réalité, quelques véhicules présentent plus de 50 attributs mais la majorité d'entre eux n'en possèdent que entre 1 et 5 (Figure 1).

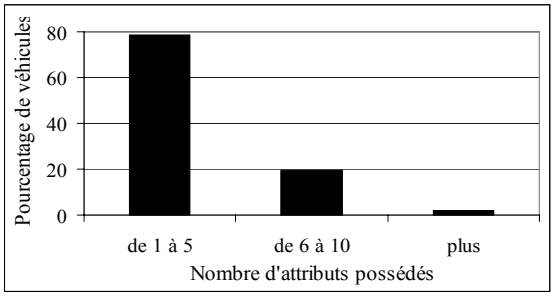

Figure 1 : Nombre d'attributs par véhicule

L'utilisation de l'outil d'extraction de règles d'association de SAS Entreprise Miner permet d'illustrer la difficulté à fixer les paramètres du support et de la confiance, en particulier dans le cas d'un grand

nombre d'événements rares. Une faible modification du support conduit à une forte augmentation du nombre de règles produites ainsi que de leur complexité (nombre d'items contenus dans une règle). Le Tableau 1 présente une synthèse des résultats obtenus à confiance minimum fixée.

Support minimum (nombre minimum de véhicules vérifiant la règle)	Confiance minimum	Nombre de règles	Nombre maximum d'items dans les règles obtenues
500	50%	16	3
400	50%	29	3
300	50%	194	5
250	50%	1 299	6
200	50%	103 981	10
100	50%	1 623 555	13
100	90%	601 873	13

Tableau 1 : Règles d'associations produites avec variation du support minimum

La règle suivante est un exemple d'association complexe entre 13 items :

$\{X835X, X813X, X810X, X802X, X800X\} \Rightarrow \{X812X, X811X, X809X, X808X, X806X, X805X, X804X, X803X\}$

L'utilisation de la classification de variables permet d'améliorer la pertinence des règles. Nous avons donc appliqué la procédure CLUSTER de SAS en utilisant l'indice de dissimilarité basé sur le coefficient de corrélation linéaire entre variables et la stratégie d'agrégation de Ward. On observe une sorte d'effet de chaîne moins prononcé cependant qu'avec la stratégie du saut minimum. Quelle que soit la coupure de l'arbre, les classes sont fortement deséquilibrées quant au nombre de variables qu'elles contiennent, comme le montre le Tableau 2.

Nombre de classes	**Nombre de variables dans les 5 plus grosses classes**				
	Classe 1	**Classe 2**	**Classe 3**	**Classe 4**	**Classe 5**
100	1575	826	35	27	27
50	1509	1177	40	34	30
30	2794	44	30	20	19

Tableau 2 : Composition des classes avec la proc CLUSTER de SAS (Ward)

Les stratégies d'agrégation selon le critère du diamètre ou, mieux encore, de la moyenne permettent de mieux équilibrer les classes.
La procédure VARCLUS de SAS, plus adaptée à la classification de variables, fournit des classes plus équilibrées (cf. Tableau 3).

Nombre de classes	**Nombre de variables dans les 5 plus grosses classes**				
	Classe 1	**Classe 2**	**Classe 3**	**Classe 4**	**Classe 5**
100	673	364	106	105	101
50	970	405	181	125	120
30	1029	591	181	169	168

Tableau 3 : Composition des classes avec la proc VARCLUS de SAS

Nous avons ensuite procédé à la recherche de règles d'association en considérant plusieurs partitions obtenues à l'issue soit de la proc VARCLUS, soit de la proc CLUSTER de SAS. Les différents essais réalisés montrent que la recherche de règles à l'intérieur de classes permet de diminuer le nombre de règles. Dans notre application, cette approche a permis de mettre en évidence une quizaine de variables liées dont les associations croisées produisent énormement de règles compliquées. Quelque soit la classification testée, ces variables sont systématiquement regroupées formant une classe atypique, certainement à traiter séparemment des autres classes. Sans la classification de variables, la détection de ce groupe de variables n'aurait pas été possible. En mettant de côté les règles relatives aux variables de cette classe, la réduction du nombre de règles à analyser par l'expert est importante : plus de 99%. De plus, les règles restantes sont plus simples. Le Tableau 4 est une synthèse des essais réalisés avec un support de 100 véhicules et une confiance de 50%.

Essai	Nombre de règles au total	Nombre maximum d'items dans toutes les règles obtenues	Réduction du nombre de règles	Nombre de règles hors de la classe atypique	Nombre maximum d'items hors de la classe atypique	Réduction du nombre de règles
Sans classification préalable	1 623 555	13	.	.	.	.
Sur 100 classes (VARCLUS)	97979	10	≈ 94 %	282	4	> 99,9 %
Sur 50 classes (VARCLUS)	1623247	13	< 1%	275	5	> 99,9 %
Sur 50 classes (CLUSTER)	1623268	13	< 1%	296	4	> 99,9 %

Tableau 4 : Synthèse des essais avec un support de 100 véhicules et une confiance de 50%

5. Conclusion et perspectives

Nous avons montré ici l'apport de la classification de variables à la recherche de règles d'association. En effet, rechercher des associations sur un grand ensemble d'événements rares conduit à une profusion de règles difficiles à interpréter de part leur nombre et leur complexité. En constituant des groupes de variables plus restreints et plus homogènes, les règles obtenues sont moins nombreuses et plus simples. De plus, la classification nous a permis de mettre en évidence le groupe de variables qui produisait les règles les plus complexes. Cependant, dans toutes les méthodes utilisées, nous avons considéré nos variables binaires comme des variables quantitatives. Dans la suite des travaux, nous envisageons d'utiliser les méthodes de classification de variables basées sur les coefficients spécifiques aux données binaires (tels que Jaccard, Ochiai...). De plus, il serait intéressant de comparer les différents arbres de classification en calculant le coefficient de corrélation linaire entre les différentes distances ultramétriques associées (Silva, 2002) ou de comparer les partitions obtenues grâce à l'indice de Rand par exemple (Youness et Saporta, 2004).

Références

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

[2] 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.

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

[4] Tan P-N., Kumar V., Srivastava J. (2002) *Selecting the Right Interestingness Measure for Association Rules.* In : Proceeding of the 8th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada

[5] Lallich S., Taytaud O. (2004) *Evaluation et validation des règles d'association.* In : Mesures de qualité pour la fouille des données, Numéro spécial Revue des Nouvelles Technologies de l'Information, RNTI, Cépaduès

[6] 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

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

[8] Qannari E.M., Vigneau E. (1998) *Une nouvelle distance entre variables. Application en classification.* Revue de Statistique Appliquée, XLVI (2), pp 21-32.

[9] Bacelar-Nicolau H., Nicolau F.C.(2003) *Teaching and learning hierachical clustering probabilistic models for categorical data*. In : Proceedings of the 54th ISI session, Berlin.August 2003

[10] Lorga da Silva A., Bacelar Nicolau H., Saporta G. (2002) *Missing Data in Hierarchical Classification of Variables, a simulation study*. In : Proceedings IFCS 2002. 8th Conference of International Federation of Classification Societies, Cracovie

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