Une comparaison de certains indices de pertinence des règles d'association

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

* CNAM Laboratoire CEDRIC 292 Rue St Martin Case 441 Paris Cedex 03
niang@cnam.fr, saporta@cnam.fr
** PSA Peugeot Citroën 45 rue Jean-Pierre Timbaud 78307 Poissy Cedex
marie.plasse@mpsa.com, laurent.leblond1@mpsa.com

Résumé. Cet article propose une comparaison graphique de certains indices de pertinence pour évaluer l'intérêt des règles d'association. Nous nous sommes appuyés sur une étude existante pour sélectionner quelques indices auxquels nous avons ajouté l'indice de Jaccard et l'indice d'accords désaccords (IAD). Ces deux derniers nous semblent plus adaptés pour discriminer les règles intéressantes dans le cas où les items sont des événements peu fréquents. Une application est réalisée sur des données réelles issues du secteur automobile.

1 Introduction

Notre étude a été motivée par le problème suivant : nous disposons de données concernant plusieurs dizaines de milliers d'individus décrits par quelques milliers d'attributs binaires assez rares et nous recherchons les éventuels liens entre certains attributs ou groupes d'attributs. La similitude de nos données avec des données de transactions nous a naturellement amenés à utiliser un algorithme de recherche de règles d'association. Cependant, le nombre élevé d'attributs conjugué à leur rareté conduit à un très grand nombre de règles dont les supports sont très faibles et les confiances très élevées. C'est pourquoi nous avons cherché à compléter l'approche support-confiance pour extraire les règles les plus pertinentes. De nombreux indices ont été proposés dans la littérature pour évaluer l'intérêt des règles d'association. Quelques uns font l'objet d'une analyse graphique à l'aide de courbes de niveaux. Nous exposons ensuite une application sur données industrielles.

2 Contexte

Ce travail est issu d'un projet industriel où l'objectif est d'exploiter une partie de l'informationnel d'un grand constructeur automobile afin d'extraire de nouvelles connaissances. Les données, issues du process de fabrication des véhicules, sont sous la forme d'une matrice où chaque véhicule est décrit par la présence ou l'absence d'attributs binaires. La connaissance d'éventuelles corrélations entre certains attributs ou groupes d'attributs représente un avantage non négligeable pour le constructeur automobile qui met un point d'honneur à améliorer

son niveau de qualité de façon continuelle. Pour répondre à cette problématique, nous utilisons la méthode de recherche de règles d'association.

Soit la règle d'association $A \rightarrow C$ où l'ensemble A, la partie antécédent ou prémisse, implique l'ensemble C, la partie conséquent ou conclusion. A et C sont des ensembles disjoints d'attributs binaires. Dans le contexte particulier de notre application, il est nécessaire de préciser que le sens de l'implication n'a pas d'importance. Une règle d'association est entièrement caractérisée par son tableau de contingence (TAB.1).

	C	$\overline{C}$	Profils lignes
A	$P(AC)$	$P(A\overline{C})$	$P(A)$
$\overline{A}$	$P(\overline{A}C)$	$P(\overline{A}\overline{C})$	$P(\overline{A})$
Profils colonnes	$P(C)$	$P(\overline{C})$	1

TAB. 1 – *Tableau de contingence de la règle d'association $A \rightarrow C$.*

Plusieurs algorithmes permettent de rechercher les règles d'association de façon déterministe à partir d'une base de données contenant n cas décrits par des variables binaires. Parmi eux, on peut citer Apriori (Agrawal et al., 1993, 1994), l'algorithme fondateur de la recherche de règles d'association, ou l'algorithme Eclat (Zaki, 2000), qui est plus rapide. Tous les algorithmes procèdent en deux étapes. Tout d'abord, ils recherchent les sous-ensembles fréquents, c'est-à-dire les conjonctions d'attributs qui apparaissent avec un support ($P(AC)$) supérieur à un seuil fixé par l'utilisateur. Puis, la seconde étape consiste à construire les règles à partir des sous-ensembles fréquents trouvés lors de la première étape. Seules les règles dotées d'une confiance ($P(C/A)$) supérieure à un seuil minimum défini par l'utilisateur seront conservées.

Bien souvent, l'approche support-confiance précédemment décrite conduit à l'obtention de règles en trop grand nombre. Par conséquent, il est impossible de les faire valider par un expert. Dès lors, il est utile de les trier par ordre décroissant de leur intérêt au sens d'un indice de pertinence, tel que le lift (Brin et al., 1997), pour citer un des plus connus.

3 Choix de quelques indices de pertinence

3.1 Une typologie existante

Il existe tellement d'indices de pertinence des règles d'association qu'il est très compliqué pour l'utilisateur de savoir lequel choisir. Nous trouvant dans cette situation, pour aider à orienter notre choix, nous nous sommes appuyés sur une suite de travaux réalisés sur ce sujet. Afin d'évaluer les indices de pertinence, Lenca et al. (2004) définissent huit propriétés formelles telles que la décroissance en fonction du nombre d'occurrences du conséquent ou la facilité à fixer un seuil d'acceptation de l'indice. Ces propriétés permettent d'évaluer les indices de pertinence et de leur attribuer des notes. Les auteurs proposent ainsi un classement d'une vingtaine d'indices. Cette étude formelle a ensuite été complétée par une étude expérimentale (Vaillant et al., 2004) où les auteurs illustrent le fait que les indices ont un comportement différent en fonction des données traitées : une classification ascendante hiérarchique de 18 indices de pertinence est réalisée à partir d'une matrice de distance déduite de la matrice de décision issue de l'évaluation formelle ; elle aboutit à la partition suivante :

Classe 1 : *{Piatetsky-Shapiro, Indice de Qualité de Cohen, Gain informationnel, Confiance centrée, Lift, Coefficient de corrélation de Pearson, Indice d'Implication, Indice probabiliste discriminant}*

Classe 2 : *{Support, Confiance, Surprise, Laplace, Sebag et Schoenauer, Taux d'exemples et de contre-exemples}*

Classe 3 : *{Multiplicateur de Cotes, Zhang, Conviction , Loevinger}*

3.2 Sélection des meilleurs indices

Afin de mieux compléter l'approche support-confiance, nous nous intéressons aux indices qui ont obtenu les notes les plus élevées selon Lenca et al. (2004) parmi ceux des classes 1 et 3 de la typologie précédente : la confiance centrée dans la classe 1, à laquelle nous rajoutons le lift en raison de son utilisation très répandue et de son interprétation facile ; le multiplicateur de cotes et le Loevinger dans la classe 3. Le tableau 2 rappelle les propriétés de ces indices dans les cas extrêmes :

Indices de pertinence	Définition	Incompatibilité $P(AC)=0$	Indépendance $P(AC)=P(A)P(C)$	Règle logique $P(C/A) = 1$
Confiance centrée	$P(C/A)-P(C)$	$-P(C)$	0	$P(\overline{C})$
Lift	$\frac{P(AC)}{P(A).P(C)}$	0	1	$\frac{1}{P(C)}$
Multiplicateur de cotes	$\frac{P(AC)P(\overline{C})}{P(A\overline{C})P(C)}$	0	1	$+\infty$
Loevinger	$\frac{P(C/A)-P(C)}{P(\overline{C})}$	$\frac{-P(C)}{P(\overline{C})}$	0	1

TAB. 2 – *Indices de pertinence retenus et leurs valeurs de référence.*

3.3 Proposition de deux indices de pertinence supplémentaires

Dans notre cas, les quatre indices de pertinence détaillés ci-dessus prennent des valeurs extrêmement élevées car de nombreuses règles ont un conséquent très fréquent par rapport au support de la règle. Dans le cas des données de transaction, cela équivaut à une règle du type *{dictionnaire→lait}*. Le lait est un achat tellement commun que de nombreux caddies en contiennent en sortie de caisse. L'achat d'un dictionnaire est moins fréquent mais toutes les transactions contenant un dictionnaire risquent de contenir aussi du lait. La règle *{dictionnaire→lait}* aura un faible support étant donnée la rareté de dictionnaire, mais sa confiance sera proche de 100%. A titre d'exemple, considérons 100 consommateurs : 8 ont acheté un dictionnaire, 40 ont acheté du lait et 7 ont acheté les deux en même temps. Cette règle qui n'a, en réalité, aucun intérêt va tout de même présenter des indices de pertinence élevés (*TAB. 3 ET 4*). Selon le lift, le nombre d'exemples de *{dictionnaire→lait}* est deux fois plus grand que sous l'indépendance de *{dictionnaire}* et *{lait}*.

	C	$\overline{C}$	Profils lignes
A	0,07	0,01	0,08
$\overline{A}$	0,33	0,59	0,92
Profils colonnes	0,4	0,6	1

Indices	Valeur
Confiance	0,88
Confiance centrée	0,48
Lift	2,19
Multiplicateur de cotes	10,5
Loevinger	0,79

TAB. 3 ET 4 – *Valeurs des indices sur un exemple numérique.*

Cela nous amène à proposer un autre indice de pertinence qui pénalise les règles où le conséquent est fréquent par rapport à l'antécédent, l'indice d'accords désaccords (IAD), qui correspond à un indice proposé par Kulczynski (1927) :

$$IAD = \frac{P(AC)}{P(\overline{A}C) + P(A\overline{C})} = \frac{accords\ positifs}{désaccords}$$

Plus cet indice est grand, plus l'antécédent et le conséquent sont présents simultanément. IAD peut également s'exprimer de la manière suivante :

$$IAD = \frac{P(AC)}{P(A)+P(C)-2P(AC)} = \frac{P(A \cap C)}{P(A \Delta C)}$$

L'indice d'accords désaccords est proche de l'indice de Jaccard :

$$Jaccard = \frac{P(AC)}{P(A)+P(C)-P(AC)} = \frac{P(A \cap C)}{P(A \cup C)}$$

La différence entre les deux indices se situe au niveau du dénominateur : pour l'indice IAD, c'est un "ou" exclusif (différence symétrique) alors que c'est un "ou" inclusif pour celui de Jaccard (union). Malgré cette différence, les deux indices sont parfaitement équivalents : ils conduisent à un classement identique des règles d'association car :

$$\frac{1}{Jaccard} = \frac{1}{MAD} + 1$$

L'indice de Jaccard présente l'intérêt d'être borné entre 0 et 1 (*TAB. 5*).

Indices de pertinence	Définition	Incompatibilité $P(AC)=0$	Indépendance $P(AC)=P(A)P(C)$	Règle logique $P(C/A) = 1$
IAD	$\frac{P(A \cap C)}{P(A \Delta C)}$	0	$\frac{P(A)P(C)}{P(A)+P(C)-2P(A)P(C)}$	$\frac{P(A)}{P(C)-P(A)}$
Jaccard	$\frac{P(A \cap C)}{P(A \cup C)}$	0	$\frac{P(A)P(C)}{P(A)+P(C)-P(A)P(C)}$	$\frac{P(A)}{P(C)}$

TAB. 5 – *Valeurs de référence pour les indices d'accords désaccords et de Jaccard.*

De manière empirique, IAD et Jaccard permettent une meilleure sélection des règles issues de nos données. Sur un exemple typique, comme celui du tableau 3, ils se montrent plus sévères que les autres indices retenus tels que le lift, en effet : IAD=0,21 et Jaccard =0,17. Ce résultat est généralisable à l'ensemble de nos données mais en aucun cas à toutes les applications.

4 Comparaison graphique des indices de pertinences

Afin de comparer graphiquement les indices, nous les avons exprimés en fonction des probabilités conditionnelles, notées $\lambda_A = P(C/A)$ et $\lambda_C = P(A/C)$, et de $P(C)$ (*TAB. 6*).

Indices	Définition	Indices	Définition
Confiance centrée	$\lambda_A = CONFCEN - P(C)$	Loevinger	$\lambda_A = LOE\left(1-P(C)\right)+P(C)$
Lift	$\lambda_A = LIFT.P(C)$	IAD	$\lambda_A = \frac{1}{\frac{1}{IAD}-\frac{1}{\lambda_C}+2}$
Multiplicateur de cotes	$\lambda_A = \frac{MC.P(C)}{1-P(C)+MC.P(C)}$	Jaccard	$\lambda_A = \frac{1}{\frac{1}{JAC}-\frac{1}{\lambda_C}+1}$

TAB. 6 – *Equations des différentes courbes de niveaux.*

Ensuite, nous avons tracé des courbes de niveaux pour chaque indice (*FIG. 1 ET 2*).

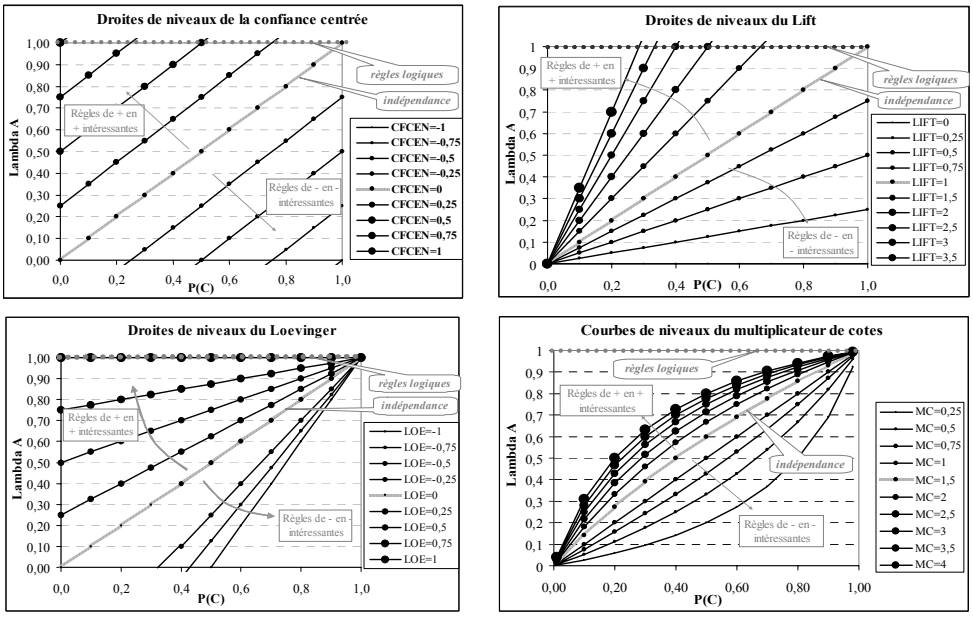

FIG. 1 – *Courbes de niveaux en fonction de* λ_A *et de P(C).*

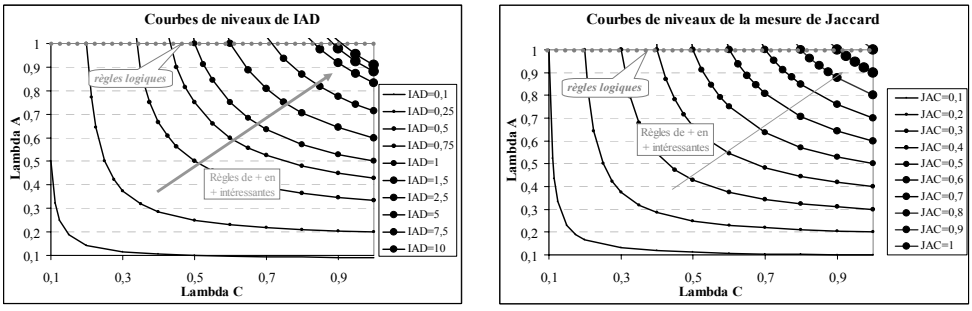

FIG. 2 – Courbes de niveaux en fonction de λ_A *et de* λ_C.

Le lift, la confiance centrée, le Loevinger et le multiplicateur de cotes ont l'avantage de prendre des valeurs fixes à l'indépendance entre l'antécédent et le conséquent, contrairement aux indices de Jaccard et IAD. L'avantage principal de ces deux derniers est la prise en compte simultanée des probabilités conditionnelles λ_A et λ_C.

Lorsque λ_A est élevée, le lift et la confiance centrée favorisent les règles où le conséquent est peu fréquent. Le Loevinger dépend surtout de λ_A ; *P(C)* n'est en fait qu'une faible pondération : en effet, plus *P(C)* est élevée, plus l'indice est élevé, mais avec *P(C)=0*, le Loevinger est égal à λ_A. Le multiplicateur de cotes ne pénalise pas les règles dont le conséquent est relativement fréquent du moment que λ_A est élevée. En effet, si *P(C)=0,8* et λ_A *=0,9*, le multiplicateur de cotes est égal à *2,25*. Or λ_A *=0,9* peut aussi bien correspondre à un cas où *P(A) = 0,8* qu'à un cas où *P(A) = 0,1* ; tout dépend en fait de la valeur du support.

Les indices de Jaccard et IAD conviennent mieux à notre application dans le sens où ils permettent d'éviter ce genre de problèmes. En effet, ils favorisent les règles où il y a forte co-présence de l'antécédent et du conséquent et où leurs fréquences sont du même ordre de grandeur.

Cependant ces indices ne conviennent pas nécessairement à tout type de problème. Le tableau suivant résume leur comportement en fonction de λ_A ou *P(A)* et de λ_C ou *P(C)*, en se référant à des achats en supermarché :

	λ_C *faible ou P(C) relativement élevé*	λ_C *élevé ou P(C) relativement faible*
λ_A *faible ou P(A) relativement élevé*	**Indices faibles - Co-absence élevée** Cas où il est rare d'observer l'antécédent et le conséquent en même temps tellement ils sont peu corrélés. *{pommes de terre → substitut de repas}*	**Indices faibles** Règle à antécédent très fréquent par rapport au conséquent. *{Vodka → Caviar}*
λ_A *élevé ou P(A) relativement faible*	**Indices faibles** Règle à conséquent très fréquent par rapport à l'antécédent. *{Dictionnaire → Lait}*	**Indices élevés - Co-présence élevée** Cas idéal dans notre application, les fréquences de l'antécédent et du conséquent sont assez proches. *{Beurre → Lait}*

TAB. 7 – Comportement des indices de Jaccard et IAD.

Note : D'un point de vue marketing, la règle inverse *{Caviar→Vodka}* est aussi intéressante car le caviar est un produit de luxe rarement acheté. Aussi, il est intéressant de savoir que les consommateurs de caviar achètent systématiquement de la Vodka en même temps. Cette règle présente un conséquent beaucoup plus fréquent que son antécédent. Elle est donc de la même famille que *{dictionnaire→lait}*, qui est le type de règles que nous cherchons à sanctionner avec les indices de Jaccard et IAD, et qui ne sera donc pas retenu.

5 Une application à un ensemble de règles

Nous disposons d'un ensemble de plus de 80000 véhicules décrits par plus de 3000 attributs binaires rares. La recherche de règles d'association sur ce jeu de données, avec un support minimum de 100 véhicules et une confiance minimum de 75%, conduit à plus de 1,5 millions de règles. Une classification de variables préalable (Plasse et al., 2005) a permis de réduire considérablement le nombre de règles candidates. Après avoir obtenu une partition en

10 classes, nous avons recherché les règles d'association à l'intérieur de chaque groupe. Nous avons ensuite établi un classement des règles obtenues selon les indices de pertinence décrits ci-dessus.

5.1 Classements des règles

Les graphiques suivants montrent les différences de classement selon les indices de pertinence, des règles contenues dans deux des dix classes obtenues : les classes A et B dans lesquelles se trouvent respectivement 19 et 29 règles candidates.

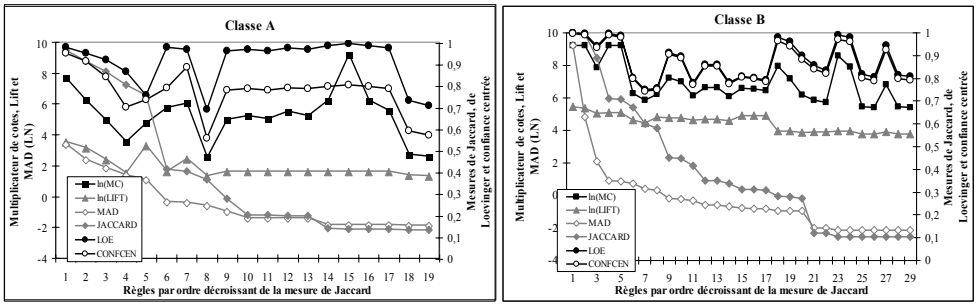

FIG. 3 – Comparaison des classements des règles selon les différents indices de pertinence.

Ces graphiques montrent deux groupes d'indices qui fournissent des classements proches : d'une part le Loevinger, la confiance centrée et le multiplicateur de cotes et d'autre part, les indices de Jaccard et IAD. Ces deux derniers conduisent bien sûr à un classement identique. Le lift ne discrimine pas les règles car ses valeurs varient très peu.

5.2 Analyse factorielle des classements

Une analyse en composantes principales des rangs attribués à chaque règle par les différents indices confirme ce qui précède. Les deux premiers facteurs expliquent 95% de l'inertie. Les cercles de corrélation des classes A et B sont identiques et aboutissent à une typologie des indices légèrement différente de celle de Vaillant et al. (2004). Les indices de Loevinger et du multiplicateur de cotes sont très corrélés sans être équivalents. Les indices de Jaccard et IAD sont confondus et orthogonaux aux deux précédents.

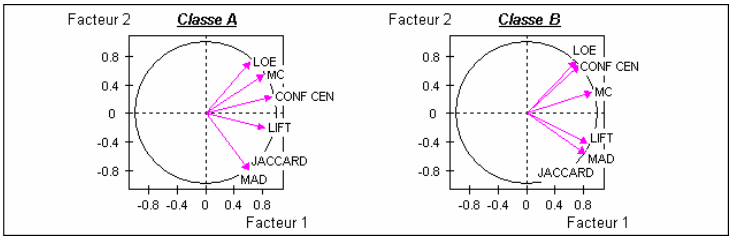

FIG. 4 – ACP sur les rangs attribués à chaque règle par les différents indices de pertinence.

6 Conclusion

Afin de visualiser et de comparer les comportements des indices qui mesurent la pertinence des règles d'association, nous avons proposé une représentation graphique originale basée sur des courbes de niveaux. De plus, dans le cadre de notre application, nous avons montré l'intérêt d'utiliser l'indice de Jaccard ou l'indice d'accords désaccords (IAD). Bien qu'ils soient symétriques, ces deux indices discriminent mieux les règles qui nous intéressent, favorisant les fortes co-présences. Cependant cette propriété peut constituer un inconvénient dans certaines applications, notamment en marketing.

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.

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.

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.

Kulczynski S. (1927) Classe des Sciences mathématiques et Naturelles. *Bulletin International de l'Académie Polonaise des Sciences et des Lettres*. Série B (Sciences Naturelles) Supplément II pages 57-203

Lenca P., Meyer P., Vaillant B., Picouet P., Lallich S. (2004). Evaluation et analyse multicritère des mesures de qualité des règles d'association. *Mesures de qualité pour la fouille de données, n° spécial RNTI Revue des Nouvelles Technologies de l'Information*, Cepaduès.

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.

Vaillant B., Lenca P., Lallich S. (2004). Etude expérimentale de mesures de qualité de règles d'association. *Revue des Nouvelles Technologies de l'Information, Actes 4e Conférence Exctraction et Gestion des Connaissances, EGC'04*, Série E, n°2, Vol.2, pp.341-352, Clermont-Ferrand.

Zaki M.J. (2000) Scalable algorithms for association mining. IEEE Transactions on Knowledge and Data Engineering, 12(3):372–390

Summary

This paper deals with a graphical comparison of some measures of association rules interest. Various measures have been studied in many papers. We propose a new measure (MAD) and compare it with the others using level curves. In an application to automobile industry, we illustrate that our measure is more relevant.