RÈGLES ORDINALES : UNE GÉNÉRALISATION DES RÈGLES D'ASSOCIATION

SYLVIE GUILLAUME[1] — ALI KHENCHAF[2]

RÉSUMÉ*: La plupart des mesures des règles concerne les variables binaires et nécessite pour les autres types de variables une phase de codage disjonctif complet. Comme la complexité des algorithmes automatiques d'extraction de règles d'association est essentiellement fonction du nombre de variables, une telle transformation peut se heurter rapidement à une explosion combinatoire ou engendrer un nombre prohibitif de règles dont la plupart sont redondantes. De plus, les mesures de liaison statistiques entre les variables quantitatives ne sont pas sélectives dans le cas de données volumineuses. Dans cette communication, nous proposons une mesure pour variables ordinales (variables quantitatives et qualitatives ordinales), l'intensité d'implication ordinale, mesure sélective pour les données volumineuses. L'étude se termine par une évaluation sur une base de données bancaire.*

MOTS-CLÉS*: Extraction de Connaissances à partir des Données (ECD), règles d'association, données volumineuses, mesures statistiques, analyse implicative, mesures quantitatives.*

SUMMARY : *Most of rule-interest measures are suitable for binary variables and require a transformation (a complete disjonctive coding) for numeric and categorical variables. Given that the complexity of unsupervised usual algorithms for the discovery of association rules increases exponentially with the number of variables, this transformation can lead us, on the hand to a combinatorial explosion, and on the other hand to a prohibitive number of weakly significant rules with many redundancies. Moreover, measures suitable for numeric variables are not selective for large databases. In this paper, we propose a measure suitable for ordinal variables, the ordinal intensity of implication, selective measure for large databases. An evaluation to some banking data ends up the study.*

KEY WORDS *: Knowledge Discovery in Databases (KDD), data mining, association rules, large databases, statistical measures, implicative analysis, numerical measures.*

[1] *IRIN, Université de Nantes*
École Polytechnique de l'Université de Nantes
2, Rue de la Houssinière – BP 92208
44322 Nantes Cedex 3 - France
sguillau@ireste.fr

[2] *Laboratoire IRCCyN, UMR 6597 CNRS, Division SETRA*
Rue C. Pauc, La Chantrerie – BP 50609
44306 Nantes Cedex 3 – France
akhencha@ireste.fr

1. INTRODUCTION

La recherche de mesures pertinentes [1] [4] pour les règles d'association est un important problème en extraction de connaissances à partir des données. Cependant, la plupart des mesures objectives [10] [12] [14] concernent les variables binaires et nécessite une phase de codage disjonctif complet des variables pour utiliser tout algorithme automatique d'extraction de règles d'association [1] [11] [15]. Comme la complexité de ces algorithmes est essentiellement fonction du nombre de variables, une telle transformation peut se heurter rapidement à une explosion combinatoire ou engendrer un nombre prohibitif de règles dont la plupart sont redondantes ou faiblement significatives. De plus, la structure d'ordre des variables ordinales (*variables qualitatives ordinales et variables quantitatives*) est perdue par cette transformation. Afin de palier à ce problème, nous nous sommes intéressés aux mesures concernant les variables quantitatives. Or, les mesures de liaison entre ces variables, comme par exemple le coefficient de corrélation significatif ou l'indice de vraisemblance du lien local [9], ne sont pas sélectives dans le cas de données volumineuses [7] et nécessitent une autre mesure pour trouver le sens de l'implication ($X \rightarrow Y$ *ou* $Y \rightarrow X$). Dans [8], une mesure implicative et sélective pour les données volumineuses, l'intensité de propension, est utilisée pour des variables quantitatives à valeurs dans l'intervalle [*0..1*]. Nous pourrions utiliser cette mesure en adaptant les variables ordinales par la transformation suivante $(X - x_{min}) / (x_{max} - x_{min})$ où x_{min} et x_{max} sont respectivement les valeurs minimale et maximale de la variable X, mais une telle transformation est pénalisante pour des données volumineuses et surtout, cette mesure privilégie certains individus [7] ce qui a pour conséquence d'éliminer des règles pertinentes. C'est pourquoi nous avons adapté et modifié cette mesure aux variables quantitatives à valeurs dans tout intervalle de l'ensemble des réels.

Ainsi, cet article s'organise de la façon suivante. Dans la section 2 nous présentons les mesures statistiques quantitatives et expliquons pourquoi celles-ci ne sont pas sélectives pour les données volumineuses. Dans la section 3 nous définissons notre mesure d'intérêt, *l'intensité d'implication ordinale*, et donnons dans la section 4 la signification des règles extraites avec cet indice, règles que nous appelons *ordinales*. La section 5 évalue cette mesure sur des données bancaires et une conclusion générale résume l'ensemble des points abordés et quelques perspectives sont envisagées pour la suite de ce travail.

2. MESURES QUANTITATIVES

Dans cette section nous présentons trois mesures statistiques quantitatives, à savoir le coefficient de corrélation linéaire significatif, l'indice de vraisemblance du lien et l'intensité de propension.

2.1. COEFFICIENT DE CORRELATION LINEAIRE SIGNIFICATIF

Pour une population dont la taille N est supérieure à *100*, la variable aléatoire R dont le coefficient de corrélation r est une valeur observée, suit approximativement la loi normale de moyenne *0* et d'écart-type $\frac{1}{\sqrt{N-1}}$ [SAP 90]. Plus la taille de la population est importante, plus l'écart-type diminue et tend vers zéro. Par conséquent, la fonction de répartition de R ne

possède que deux valeurs pour les grandes populations : *0* et *1*. Cette mesure n'est donc pas sélective pour les données volumineuses.

2.2. INDICE DE VRAISEMBLANCE DU LIEN LOCAL

Soient *X* et *Y* deux variables quantitatives, *N* la taille de la population Ω et *r* le coefficient de corrélation linéaire. L'indice de vraisemblance du lien local *V(X,Y)* [9] est défini de la façon suivante :

$$V(X,Y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\sqrt{N-1}\times r} e^{-\frac{1}{2}t^2} dt$$

Lorsque *r* est négatif (*respectivement positif*) et la taille *N* de la population élevée, alors la valeur de $\sqrt{N-1} \times r$ tend vers moins l'infini (*respectivement l'infini*), et la valeur de l'indice *V(X,Y)* tend vers *0* (*respectivement 1*). Pour finir, lorsque la valeur de *r* est égale à *0*, la valeur de *V(X,Y)* est égale à *0.5*.

Par conséquent, cet indice n'est également pas une mesure sélective pour les données volumineuses car il ne possède que trois valeurs : *0*, *0.5* et *1*.

Afin de corriger l'inconvénient de cette mesure, I.C. Lerman l'a modifiée par un centrage et une réduction de toutes les valeurs de cette mesure extraites à partir d'une base de données.

Soit μ la moyenne des valeurs de cette mesure et soit σ l'écart-type. La nouvelle mesure *V'(X,Y)*, appelée indice de vraisemblance du lien global, est définie par :

$$V'(X,Y) = \frac{V(X,Y) - \mu}{\sigma}$$

Cette nouvelle mesure, sélective pour des données volumineuses, nécessite l'utilisation d'une autre mesure afin de détecter le sens des règles ($X \rightarrow Y$ ou $Y \rightarrow X$).

2.3. INTENSITE DE PROPENSION

Soient *X* et *Y* deux variables quantitatives à valeurs x_i et y_i dans l'intervalle [*0..1*] et soit *n* la taille de l'échantillon *E* représentatif de la population Ω ($i \in \{1,..,n\}$). Soient m_X et m_Y les moyennes arithmétiques des variables *X* et *Y*, et soient v_X et v_Y les variances des variables *X* et *Y*.

L'intensité de propension *P(X→Y)* [8] généralise l'intensité d'implication [5], mesure utilisée dans des systèmes de découverte de connaissances comme FIABLE [3] et PEDRE [16], aux variables quantitatives à valeurs dans l'intervalle [*0..1*]. Elle est définie de la façon suivante :

$$P(X \rightarrow Y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{s} e^{-\frac{1}{2}t^2} dt \quad \text{avec} \quad s = \frac{\frac{\sum_{i=1}^{n} x_i(1-y_i)}{n} - m_X(1-m_Y)}{\sqrt{\frac{(v_X + m_X^2)(v_Y + (1-m_Y)^2)}{n}}}.$$

Contrairement aux deux mesures précédentes, l'intensité de propension est une mesure implicative et sélective pour les données volumineuses. Cependant, celle-ci n'est valable que pour les variables quantitatives à valeurs dans [*0..1*]. Nous pouvons l'utiliser en adaptant les variables quantitatives par la transformation $(X-x_{min})/(x_{max}-x_{min})$ avec x_{min} et x_{max} les valeurs minimale et maximale de la variable X mais nous préférons évaluer les implications sur l'ensemble des variables initiales car une telle transformation est pénalisante pour des données volumineuses et surtout, cette mesure privilégie certains individus [7] ce qui a pour conséquence d'éliminer des règles pertinentes. C'est pourquoi nous avons adapté et modifié l'intensité de propension aux variables quantitatives à valeurs dans tout intervalle $[x_{min}..x_{max}]$ de l'ensemble des réels.

3. INTENSITÉ D'IMPLICATION ORDINALE

Dans cette section, nous présentons l'intensité d'implication ordinale qui généralise l'intensité de propension et l'intensité d'implication aux variables quantitatives à valeurs dans tout intervalle de l'ensemble des réels et aux variables qualitatives nominales après un codage approprié des valeurs de celles-ci dans l'ensemble des réels.

Soient X et Y deux variables quantitatives à valeurs x_i et y_i ($i \in \{1,..,N\}$) dans respectivement les intervalles $[x_{min}..x_{max}]$ et $[y_{min}..y_{max}]$.

L'intensité d'implication ordinale mesure si le nombre des individus ne vérifiant pas fortement la règle $X \rightarrow Y$, c'est-à-dire le nombre des individus ayant une valeur élevée pour X et faible pour Y, est significativement faible comparativement à ce que l'on obtiendrait si par hypothèse ces deux variables étaient indépendantes. Nous appelons ce nombre d'individus, le nombre des contre-exemples ou encore la mesure brute ordinale.

En s'inspirant des travaux de [6] et [8], nous proposons la mesure ordinale brute suivante :

$$q_o = \sum_{i=1}^{N} (x_i - x_{\min})(y_{\max} - y_i)$$

Dans [8], J.B. Lagrange retient un indice moyen $q_0' = \frac{\sum_{i=1}^{N} x_i(1-y_i)}{N}$ pour sa mesure de propension et, dans [6] R. Gras propose l'indice brut $q_0'' = \frac{\sum_{i=1}^{N} x_i(y_{\max}-y_i)}{x_{\max} y_{\max}}$. Comme la mesure de [8] n'est valable que pour des variables à valeurs dans l'intervalle [*0..1*], nous avons la relation $q_0'' = N q_0'$ entre ces deux derniers indices.

L'indice ordinal brut $\sum_{i=1}^{N} x_i(y_{\max}-y_i)$ aurait pu être retenu mais il est important de faire intervenir la valeur minimale de X. En effet, si nous considérons le terme x_i au lieu du terme (x_i-x_{min}) dans la mesure brute, nous donnons de l'importance aux individus ayant des faibles valeurs pour X et ce, d'autant plus, que la valeur minimale prise par X est élevée. L'indice $\sum_{i=1}^{N} x_i(y_{\max}-y_i)$ a une valeur supérieure à q_0 et peut rejeter des règles valides. Ceci est illustré par la figure 1 donnant les courbes des fonctions $x_i(y_{max}-y_i)$ (*courbe de gauche*) et $(x_i-x_{min})(y_{max}-y_i)$ (*courbe de droite*) pour $x \in [100..150]$ et $y \in [50..90]$.

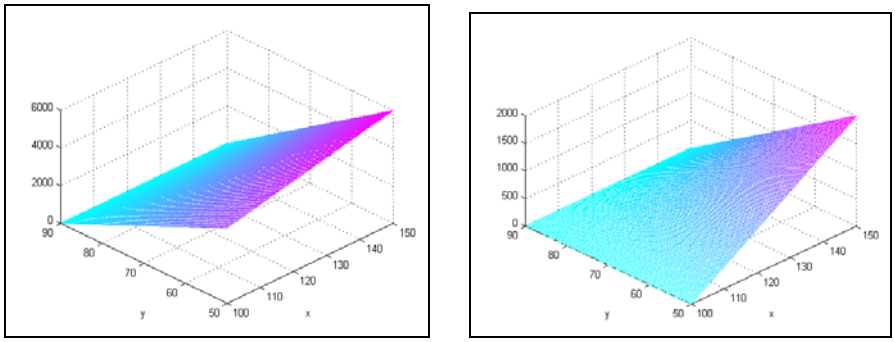

Figure 1 : Courbes des mesures brutes avec (*courbe droite*) et sans (*courbe gauche*) la prise en compte de la valeur minimale de X.

La courbe de gauche (*figure 1*) montre que les individus ayant une faible valeur pour X et Y ont un poids aussi important que certains individus ayant une forte valeur pour X et une faible valeur pour Y contrairement à la courbe de droite (*figure 1*) où les individus ayant une faible valeur pour X ont un poids négligeable. Ainsi, pour la courbe de gauche, les individus vérifiant simultanément $x=100$ et $y=50$, ont un poids équivalent à ceux vérifiant simultanément $x=150$ et $y=82$ c'est-à-dire un poids de 400 ce qui n'est pas le cas pour la courbe de droite car ces individus ont un poids nul. C'est pourquoi nous avons retenu la mesure ordinale brute q_0.

Comme pour tous les indices statistiques, on se réfère à une échelle de mesure c'est-à-dire à une échelle probabiliste. Il faut donc déterminer la loi de probabilité de la variable aléatoire Q dont cette mesure brute q_o est une réalisation.

Soit une épreuve aléatoire E consistant à prélever un échantillon E de n individus parmi la population Ω. Soient U et W deux variables aléatoires indépendantes prenant respectivement leurs valeurs u_i et w_i dans les intervalles $[u_{min}..u_{max}]$ et $[w_{min}..w_{max}]$ $(i \in \{1,..,n\})$. Afin d'effectuer une comparaison avec les variables X et Y définies ci-dessus, ces deux variables aléatoires U et W doivent avoir la même moyenne et variance que celles respectivement de X et Y. Soient m_X, m_Y, m_U et m_W les moyennes respectivement de X, Y, U et W et soient v_X, v_Y, v_U et v_W les variances respectivement de X, Y, U et W; nous avons $m_X = m_U$, $m_Y = m_W$, $v_X = v_U$, et $v_Y = v_W$.

Considérons la variable aléatoire Q égale à $\sum_{i=1}^{n}(U_i-u_{\min})(w_{\max}-W_i)$ dont $q_o = \sum_{i=1}^{n}(u_i-u_{\min})(w_{\max}-w_i)$ est une réalisation. Cette variable aléatoire Q suit asymptotiquement la loi normale $N(m,v)$ avec $m=n\,(m_X-x_{min})(y_{max}-m_Y)$ et $v^2=n[v_X+(m_X-x_{min})^2][v_Y+(y_{max}-m_Y)^2]$ [7].

Un test statistique unilatéral va déterminer la validité de l'hypothèse d'indépendance des deux variables X et Y.

Soient H_0 l'hypothèse d'indépendance entre X et Y et H_1 l'hypothèse alternative. Soit α le risque de première espèce. La règle de décision est la suivante :

Si $Pr(Q \leq q_o) > \alpha$ ***alors*** *on accepte* H_0

sinon *on rejette H_0 et on accepte H_1*

Si la probabilité $Pr(Q \leq q_o)$ d'avoir un nombre inférieur ou égal à q_o est élevée, nous pouvons en conclure que q_o n'est pas significativement faible car pouvant se produire assez fréquemment et par conséquent l'implication $X \rightarrow Y$ n'est pas pertinente. Par contre, si la probabilité $Pr(Q \leq q_o)$ est faible, nous pouvons en conclure que l'implication $X \rightarrow Y$ a un sens puisqu'il est fort improbable d'obtenir un nombre aussi faible.

Afin de mesurer cette implication de façon croissante, l'indice $\varphi(X \rightarrow Y) = Pr(Q > q_o)$ est retenu. Ainsi, l'implication $X \rightarrow Y$ est admissible au niveau de confiance $(1-\alpha)$ si et seulement si $Pr(Q \leq q_o) \leq \alpha$ ou $Pr(Q > q_o) \geq 1-\alpha$.

L'intensité d'implication ordinale est donc :

$$\varphi(X \rightarrow Y) = \frac{1}{\sqrt{2\pi\sigma}} \int_{q_0}^{+\infty} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

4. REGLES ORDINALES

Dans cette section, nous expliquons la signification des règles extraites par l'intensité d'implication ordinale, les règles ordinales.

La figure 2 (*courbe de gauche*) représente les individus e_i ($i \in \{1, .., n\}$) de E suivant les valeurs de X et Y. On suppose que la variable X prend r valeurs distinctes $x_1 = x_{min} < x_2 < \ldots < x_{r-1} < x_r = x_{max}$ et la variable Y possède s valeurs distinctes $y_1 = y_{min} < y_2 < \ldots < y_{s-1} < y_s = y_{max}$.

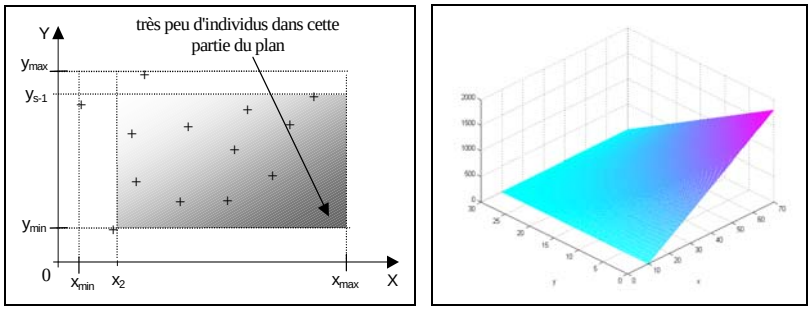

Figure 2. Répartition des individus de E suivant les valeurs de X et Y *(courbe de gauche)* et poids attribué à ces individus *(courbe de droite)*

L'intensité d'implication ordinale évalue, pour la règle $X \rightarrow Y$, si le nombre d'individus dans la zone grisée de la figure 2 (*courbe gauche*) est statistiquement faible. Cependant tous les individus de cette zone n'ont pas la même importance et les individus du coin inférieur droit ont un poids plus important, comme le montre la courbe de droite de la figure 2. Cette dernière, où on a supposé que la variable X prend ses valeurs dans $[10..70]$ et Y dans $[0..30]$, montre la courbe de la fonction $g(x,y) = (x-x_{min})(y_{max}-y)$ où le nombre $(x_i-x_{min})(y_{max}-y)$ pour l'individu e_i de E correspond à $g(x_i,y_i)$.

5. ÉVALUATION SUR DES DONNÉES BANCAIRES

Dans cette section, nous présentons les règles ordinales découvertes sur une base de données bancaires. Tout d'abord, nous décrivons la base de données et ensuite donnons les règles ordinales.

5.1. DONNEES BANCAIRES

La base de données bancaires se composent de *47 112* individus décrits par *52* variables dont *96%* d'entre elles sont des variables quantitatives.

Les variables peuvent être répertoriées en trois catégories :

- Informations concernant le client (*âge, ancienneté,* …),
- Informations sur les différents comptes du client (*actions, obligations, PEL,* …),
- Statistiques sur les différents comptes (montant des ressources, montant des encours prêt, …).

Les informations concernant les comptes du client peuvent être divisées en deux nouvelles catégories :

- Variables mémorisant les encours de chaque type *Z* de comptes ouverts par le client, c'est-à-dire les sommes d'argent déposées sur ces comptes (*encoursZ*),
- Variables comptabilisant le nombre de comptes ouverts pour un type de produit donné (*nombreZ*) dont le montant total est enregistré dans les variables précédentes.

5.2. RESULTATS

208 règles ordinales au seuil minimal de *0,95* ont été découvertes par l'intensité d'implication ordinale.

Cette extraction a permis de découvrir les services pouvant intéresser un client détenteur d'un type de compte donné. Ainsi, nous avons découvert que les clients possédant, par exemple, un livret d'épargne sont potentiellement intéressés par un plan épargne logement (*avec une intensité d'implication égale à 1*), un compte épargne logement (*avec une intensité égale à 1*), un plan d'épargne populaire (*avec une intensité égale à 1*), une carte bleue (*avec une intensité égale à 0,96*) et un disponible permanent (*avec une intensité égale à 0,95*). Ce type d'implication a été détecté pour tous les produits financiers offerts par les banques, ce qui représente *76* règles.

D'autres exemples de règles découvertes : les personnes qui possèdent le plus grand nombre de comptes actions sont les personnes les plus anciennes dans la banque (*la valeur de l'intensité d'implication ordinale est égale à 1*) et les clients qui possèdent le plus grand nombre de plan épargne populaire sont les personnes les plus âgées (*intensité d'implication ordinale égale à 1*).

Les fortes relations découvertes entre les variables sont les suivantes.

- Une équivalence entre l'âge du client et son ancienneté dans la banque ce qui indique que plus l'âge du client est avancé, plus son ancienneté est grande et vice-versa.

- Des équivalences entre le montant des encours pour un type de compte donné *Z* (*encoursZ*) et le nombre de comptes ouverts pour ce même type de compte (*nombreZ*). Ces équivalences sont vérifiées pour *44%* des produits financiers offerts par la banque, comme par exemple le prêt épargne logement et le plan d'épargne populaire (*PEP*). Ce n'est pas vérifié, par exemple, pour les obligations, les actions et le livret d'épargne.

- Une très forte relation entre le plan épargne logement (*PEL*) et le compte épargne logement (*CEL*) comme le montre la figure 4. Ces relations indiquent qu'en général un client souscrit à ces deux produits financiers.

6. CONCLUSION ET PERSPECTIVES

Nous avons trouvé une mesure ordinale sélective pour les données volumineuses : *l'intensité d'implication ordinale*. Cette mesure évite l'étape de transformation des variables en variables binaires et permet de découvrir un nouveau type de règles, les règles ordinales, qui nous renseignent sur l'évolution conjointe des variables. Ces règles nous révèlent le comportement général de la population plus facilement qu'avec des règles dont les variables sont partitionnées en intervalles. Cette étude doit se poursuivre avec la recherche de règles d'association ordinales, c'est-à-dire des règles composées de plus d'une variable en prémisse et / ou en conclusion.

BIBLIOGRAPHIE

[1] AGRAWAL R., MANNILA H., SRIKANT R., TOIVONEN H., VERKAMO A.I., Fast discovery of association rules, *Fayyad U.M., Piatetsky-Shapiro G., Smyth P. and Uthurusamy R. eds., Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press*, 1996, p. 307-328.

[2] BRIN S., MOTWANI R., SILVERSTEIN C., Beyond Market Baskets : Generalizing Association Rules to Correlations, *Proceedings of the 1997 ACM SIGMOD Conference*, Tucson, Arizona, mai 1997, p. 265-276.

[3] FLEURY L., BRIAND H., PHILIPPE J., DJERABA C., Rules Evaluations for Knowledge Discovery in Database, *6th International Conference and Workshop on Database and Expert Systems Applications, DEXA*, Londres, Angleterre, 1995.

[4] FREITAS A.A., On Objective Measures of Rules Surprisingness, *Second European Symposium on Principles of Data Mining and Knowledge Discovery PKDD'98*, Nantes, France, 1998, p. 1-9.

[5] GRAS R., Contribution à l'Etude Expérimentale et à l'Analyse de certaines Acquisitions Cognitives et de certains Objectifs Didactiques en Mathématiques, Thèse d'État, Université de Rennes I, Octobre 1979.

[6] GRAS R., ALMOULOUD S.A., BAILLEUL M., LARHER A., POLO M., RATSIMBA-RAJOHN H., TOTOHASINA A., *L'implication Statistique*, La Pensée Sauvage, 1996.

[7] GUILLAUME S., KHENCHAF A., et BRIAND H., *Generalizing Association Rules to ordinal rules*, In proceedings of Information Quality Conference, IQ2000, edited by Barbara D. Klein University of Michigan-Dearborn and Donald F. Rossin University of Michigan-Dearborn, pp. 268-282, Cambridge, Massachusetts, USA, October 2000.

[8] LAGRANGE J.B., Analyse Implicative d'un Ensemble de Variables Numériques; Application au Traitement d'un Questionnaire à Réponses Modales Ordonnées, *Revue de Statistique Appliquée,* I.H.P., Paris, 1997.

[9] LERMAN I.C., *Classification et analyse ordinale des données,* Dunod, 1981.

[10] MAJOR J., MANGANO J., Selecting among Rules Induced from a Hurricane Database, *KDD-93*, p. 28-41.

[11] MANNILA H., TOIVONEN H., VERKAMO A.I., Efficient algorithms for Discovering Association Rules, *Usama M. Fayyad et Ramasamy Uthurusamy, éditeurs, AAAI Workshop on Knowledge Discovery in Databases*, Seattle, Washington, Juillet 1994, p. 181-192.

[12] PIATETSKY-SHAPIRO G., Discovery, Analysis and Presentation of Strong Rules, *In G. Piatetsky-Shapiro & W.J. Frawley, editors, Knowledge Discovery in Databases*, AAAI Press, 1991, p. 229-248.

[13] SAPORTA G., *Théories et méthodes de la statistique*, Editions Technip, 1978.

[14] SILBERSCHATZ A., TUZHILIN A., What makes Patterns interesting in Knowledge Discovery Systems, *IEEE Trans. On Know. And Data Eng.*, vol.8, n°6, p. 970-974.

[15] SRIKANT R., AGRAWAL R., Mining quantitative association rules in large relational tables, *Proceedings 1996 ACM-SIGMOD International Conference Management of Data*, Montréal, Canada, Juin 1996.

[16] SUZUKI E., KODRATOFF Y., Discovery of Surprising Exception Rules Based on Intensity of Implication, *Second European Symposium on Principles of Data Mining and Knowledge Discovery PKDD'98*, Nantes, France, 1998, p. 10-18.