# Catégorisation des mesures d'intérêt pour l'extraction des connaissances

Sylvie Guillaume*, ***, Dhouha Grissa**, ***, **** et Engelbert Mephu Nguifo**, ***

Clermont Université, Université d'Auvergne* et Université Blaise Pascal**, LIMOS, BP 10448, F-63000 Clermont-Ferrand
CNRS***, UMR 6158, LIMOS, F-63173 AUBIERE
URPAH****, Département d'Informatique, Faculté des Sciences de Tunis, Campus Universitaire, 1060 Tunis, Tunisie
guillaum@isima.fr, dgrissa@isima.fr, mephu@isima.fr

**Résumé.** La recherche de règles d'association intéressantes est un domaine de recherche important et actif en fouille de données. Les algorithmes de la famille *Apriori* reposent sur deux mesures pour extraire les règles, le support et la confiance. Bien que ces deux mesures possèdent des vertus algorithmiques accélératrices, elles génèrent un nombre prohibitif de règles dont la plupart sont redondantes et sans intérêt. Il est donc nécessaire de disposer d'autres mesures filtrant les règles inintéressantes. Des travaux ont été réalisés pour dégager les "*bonnes*" propriétés des mesures d'extraction des règles et ces propriétés ont été évaluées sur *61* mesures. L'objectif de cet article est de dégager des catégories de mesures afin de répondre à une préoccupation des utilisateurs : le choix d'une ou plusieurs mesures lors d'un processus d'extraction des connaissances dans le but d'éliminer les règles valides non pertinentes extraites par le couple (*support, confiance*). L'évaluation des propriétés sur les *61* mesures a permis de dégager *9* classes de mesures, classes obtenues grâce à deux techniques : une méthode de la classification ascendante hiérarchique et une version de la méthode de classification non-hiérarchique des *k*-moyennes.

## 1 Introduction

Les algorithmes d'extraction de règles d'association (Agrawal et Srikant 1994), fondés sur les mesures *support* et *confiance*, ont tendance à générer un nombre important de règles. Ces deux mesures ne sont pas suffisantes pour extraire uniquement les règles réellement intéressantes et ce constat a été mis en évidence dans de nombreux travaux comme par exemple (Sese et Morishita 2002, Carvalho et al. 2005). Une étape supplémentaire d'analyse des règles extraites est donc indispensable et différentes solutions ont été proposées. Une première solution consiste à restituer facilement et de façon synthétique l'information extraite grâce à des techniques de représentation visuelle (Hofmann et Wilhelm 2001, Blanchard et al. 2003). Une seconde voie consiste à réduire le nombre de règles extraites. Certains auteurs (Zaki 2000, Zaman Ashrafi et al., 2004, Ben Yahia et al. 2009) éliminent les règles

redondantes, d'autres évaluent et ordonnent les règles grâce à d'autres mesures d'intérêt (Lenca et al., 2008). Dans cet article, nous nous intéressons à cette dernière voie : le recours à d'autres mesures pour éliminer les règles inintéressantes. De nombreux travaux de synthèse ont comparé les différentes mesures objectives rencontrées dans la littérature selon plusieurs points de vue : les propriétés sous-jacentes à une "*bonne*" mesure d'intérêt (Tan et al. 2002, Lallich et Teytaud 2004, Geng et Hamilton 2007, Feno 2007 et Vaillant 2007, Guillaume et al., 2010). Ces articles synthétiques ont mis en évidence un grand nombre de mesures présentes dans la littérature (*plus d'une soixantaine*) avec de nombreuses propriétés (*une vingtaine*).

L'objectif de cet article est d'aider l'utilisateur dans le choix d'une ou plusieurs mesures complémentaires afin d'éliminer les règles non pertinentes[1] extraites par le couple (*support, confiance*). Pour cela, nous souhaitons détecter des groupes de mesures ayant des propriétés similaires, ce qui permettra à l'utilisateur, d'une part, de restreindre le nombre de mesures à choisir, et d'autre part, d'orienter son choix en fonction des propriétés qu'il souhaite que celles-ci vérifient.

Ce travail s'appuie sur les travaux synthétiques qui ont été réalisés sur les mesures et leurs propriétés et plus particulièrement sur les travaux de Guillaume et al. (Guillaume et al., 2010) car étant l'article le plus récent dans ce domaine, c'est le plus complet puisque c'est une synthèse des travaux de (Tan et al. 2002, Lallich et Teytaud 2004, Huyng et al. 2005, Geng et Hamilton 2007, Feno 2007 et Vaillant 2007). Ce travail de synthèse de Guillaume et al. (Guillaume et al., 2010) a répertorié une soixantaine de mesures d'intérêt et une vingtaine de propriétés. Ce travail s'est terminé par l'évaluation de *19* propriétés sur *61* mesures.

L'objectif de cet article est de dégager des classes de mesures ayant des comportements similaires mais en aucun cas d'expliquer les propriétés et les mesures répertoriées dans la littérature, explications pouvant être trouvées dans les articles de synthèse (Tan et al. 2002, Lallich et Teytaud 2004, Geng et Hamilton 2007, Feno 2007 et Vaillant 2007). La recherche de ces classes de mesures a été effectuée en utilisant des techniques bien connues comme une des méthodes de la classification ascendante hiérarchique utilisant le critère de Ward et une version de la méthode de classification non-hiérarchique des *k*-moyennes. Un consensus sera dégagé à partir des résultats obtenus avec ces deux techniques. Avant de lancer cette recherche de classes, il nous est apparu essentiel de vérifier que cette matrice de *61* mesures $\times$ *19* propriétés ne pouvait pas être simplifiée en recherchant des groupes de mesures aux comportements totalement similaires par rapport aux *19* propriétés et également, s'il n'y avait pas de propriétés redondantes.

L'article s'organise donc de la façon suivante. La *section 2* expose brièvement la matrice des *mesures* $\times$ *propriétés* sur laquelle nous recherchons les classes et étudie si celle-ci ne peut pas être simplifiée. La *section 3* restitue les résultats de la classification obtenue par la première technique : une méthode de la classification ascendante hiérarchique utilisant le critère de Ward. La *section 4* donne les résultats dégagés par la deuxième technique : une version de la méthode de classification non-hiérarchique des *k*-moyennes et discute de la cohérence des résultats obtenus par ces deux techniques. La section se termine par une classification consensuelle. Pour finir, la *section 5* essaye de trouver une sémantique à chacune des classes extraites et valide la classification retenue avec celle dégagée par Benoît Vaillant (Vaillant, 2007). L'article se termine par une conclusion et des perspectives.

[1] La pertinence ou l'intérêt d'une règle se mesure par rapport au problème étudié, et certaines règles pertinentes peuvent ne pas être valides du fait de la mesure utilisée.

## 2 Évaluation des propriétés sur les mesures

Comme nous l'avons mentionné, notre travail s'appuie sur les résultats de recherche de Guillaume et al. (Guillaume et al., 2010), à savoir une matrice évaluant *19* propriétés sur *61* mesures *m*. Nous nous contentons uniquement de rappeler ces propriétés qui ont été dégagées et formalisées sans les expliquer. Ces propriétés sont les suivantes :

- **P1** : la mesure est non symétrique ($P_1(m)=1$) ou symétrique ($P_1(m)=0$)
- **P2** : la mesure est non symétrique au sens de la négation de la conclusion ($P_2(m)=1$) ou symétrique ($P_2(m)=0$)
- **P3** : la mesure évalue de la même façon la règle $X \rightarrow Y$ et $\overline{Y} \rightarrow \overline{X}$ dans le cas de l'implication logique ($P_3(m)=1$) ou pas de la même façon ($P_3(m)=0$)
- **P4** : la mesure est croissante en fonction du nombre d'exemples[2] ($P_4(m)=1$) ou non croissante ($P_4(m)=0$)
- **P5** : la mesure est croissante en fonction du nombre d'individus ($P_5(m)=1$) ou non ($P_5(m)=0$)
- **P6** : la mesure est décroissante en fonction de la taille du conséquent *Y* ($P_6(m)=1$) ou non ($P_6(m)=0$)
- **P7** : la mesure a une valeur fixe dans le cas de l'indépendance[3] ($P_7(m)=1$) ou non ($P_7(m)=0$)
- **P8** : la mesure a une valeur fixe dans le cas de l'implication logique[4] ($P_8(m)=1$) ou non ($P_8(m)=0$)
- **P9** : la mesure a une valeur fixe dans le cas de l'équilibre[5] ($P_9(m)=1$) ou non ($P_9(m)=0$)
- **P10** : la mesure a des valeurs identifiables en cas d'attraction[6] entre *X* et *Y* ($P_{10}(m)=1$) ou non ($P_{10}(m)=0$)
- **P11** : la mesure a des valeurs identifiables en cas de répulsion[7] entre *X* et *Y* ($P_{11}(m)=1$) ou non ($P_{11}(m)=0$)
- **P12** : la mesure est tolérante aux premiers contre-exemples ($P_{12}(m)=2$), non tolérante ($P_{12}(m)=0$) ou indifférente ($P_{12}(m)=1$)
- **P13** : la mesure est invariante en cas de dilatation de certains effectifs ($P_{13}(m)=1$) ou non ($P_{13}(m)=0$)
- **P14** : la mesure a la relation $m(\overline{X} \rightarrow Y) = -m(X \rightarrow Y)$ entre $X \rightarrow Y$ et $\overline{X} \rightarrow Y$ ($P_{14}(m)=1$) ou non ($P_{14}(m)=0$)
- **P15** : la mesure a la relation $m(X \rightarrow \overline{Y}) = -m(X \rightarrow Y)$ entre $X \rightarrow Y$ et $X \rightarrow \overline{Y}$ ($P_{15}(m)=1$) ou non ($P_{15}(m)=0$)
- **P16** : la mesure a la relation $m(\overline{X} \rightarrow \overline{Y}) = m(X \rightarrow Y)$ entre $X \rightarrow Y$ et $\overline{X} \rightarrow \overline{Y}$ ($P_{16}(m)=1$) ou non ($P_{16}(m)=0$)
- **P17** : la mesure est fondée sur un modèle probabiliste ($P_{17}(m)=1$) ou non ($P_{17}(m)=0$)
- **P18** : la mesure est statistique ($P_{18}(m)=1$) ou descriptive ($P_{18}(m)=0$)
- **P19** : la mesure est discriminante ($P_{19}(m)=1$) ou non ($P_{19}(m)=0$)

---

[2] individu qui vérifie à la fois la prémisse *X* de la règle et la conclusion *Y*.
[3] cas où la réalisation de *X* n'augmente pas les chances d'apparition de *Y*.
[4] cas où la probabilité conditionnelle P(*Y*/*X*) est égale à *1*.
[5] cas où lorsque *Y* est réalisé, il y a autant de chances que *X* ou *non X* soit réalisé.
[6] lorsque la réalisation de *X* augmente les chances d'apparition de *Y*.
[7] lorsque la réalisation de *X* diminue les chances d'apparition de *Y*.

Les *61* mesures étudiées dans (Guillaume et al., 2010) sont les suivantes : (1) coefficient de corrélation, (2) Cohen ou Kappa, (3) confiance, (4) confiance causale, (5) confiance centrée ou Pavillon, (6) confiance confirmée descriptive ou Ganascia, (7) confiance confirmée causale, (8) confirmation causale, (9) confirmation descriptive, (10) conviction, (11) cosinus ou Ochiai, (12) couverture, (13) Czekanowski ou F-mesure, (14) dépendance, (15) dépendance causale, (16) dépendance pondérée, (17) facteur bayésien, (18) facteur de certitude ou Loevinger ou satisfaction, (19) fiabilité négative, (20) force collective, (21) Fukuda, (22) gain informationnel, (23) Gini, (24) Goodman, (25) indice d'implication, (26) intensité probabiliste d'écart à l'équilibre (IPEE), (27) intensité probabiliste entropique d'écart à l'équilibre (IP3E), (28) indice probabiliste discriminant (IPD), (29) information mutuelle, (30) intensité d'implication (II), (31) intensité d'implication entropique (IIE), (32) intensité d'implication entropique révisée (IIER), (33) indice de la vraisemblance du lien (IVL), (34) intérêt, (35) Jaccard, (36) J-mesure, (37) Klosgen, (38) Kulczynski ou indice d'accord et de désaccord, (39) Laplace, (40) Leverage, (41) mesure de Guillaume-Khenchaf $M_{GK}$, (42) moindre contradiction ou surprise, (43) nouveauté, (44) Pearl, (45) Piatetsky-Shapiro, (46) précision ou support causal, (47) prévalence, (48) Q de Yule, (49) rappel, (50) ratio des chances, (51) risque relatif, (52) Sebag-Schoenauer, (53) spécificité, (54) support, (55) support à sens unique, (56) support à double sens, (57) taux d'exemples, (58) VT100, (59) variation du support, (60) Y de Yule, (61) Zhang. Les expressions de chacun des indices sont disponibles dans (Guillaume et al., 2009).

Après avoir présenté les données sur lesquelles nous allons réaliser une classification, nous allons maintenant nous assurer que celles-ci ne peuvent pas être restreintes en recherchant des groupes de mesures aux comportements identiques et si des propriétés ne sont pas redondantes.

Dans un premier temps, nous avons donc recherché toutes les mesures dont les valeurs pour chacune des *19* propriétés sont identiques. Nous avons trouvé les *7* groupes suivants : $G_1$ = {*coefficient de corrélation, nouveauté*}, $G_2$ = {*confiance causale, confiance confirmée causale, fiabilité négative*}, $G_3$ = {*cosinus, Czekanowski-Dice*}, $G_4$ = {*dépendance causale, Leverage, spécificité*}, $G_5$ = {*force collective, ratio des chances*}, $G_6$ = {*Gini, information mutuelle*} et $G_7$ = {*Jaccard, Kulczynski*}.

Suite à la détection de ces *7* groupes de mesures, nous sommes donc maintenant en présence d'une matrice de *52* mesures puisque nous gardons une seule mesure par groupe.

Nous recherchons maintenant si des propriétés ne sont pas redondantes. Pour cela, nous avons recherché si une propriété avait des valeurs identiques pour chacune des *52* mesures avec une autre propriété. Nous n'avons trouvé aucune relation de ce type, ce qui nous révèle qu'il n'y a pas de propriétés identiques.

Nous sommes donc à présent avec une matrice de *52* mesures et *19* propriétés, propriétés qui sont des variables qualitatives nominales. Afin de lancer deux versions d'algorithmes de classification, versions nécessitant des variables binaires, nous effectuons un codage disjonctif complet, ce qui nous conduit à l'obtention de *39* variables binaires. Nous sommes donc pour finir en présence d'une matrice de *52* mesures × *39* variables binaires.

Après avoir discuté des données et transformé celles-ci pour pouvoir appliquer les algorithmes choisis, nous étudions la première classification de mesures obtenue avec une méthode de classification ascendante hiérarchique.

# 3 Classification obtenue par une méthode de CAH

Nous avons effectué une classification ascendante hiérarchique (*CAH*) avec le logiciel *Matlab* sur ces *52* mesures en utilisant la distance euclidienne entre paires de mesures puis la distance de Ward pour la phase d'aggrégation. La *figure 1* restitue cette classification pour la distance de Ward. Comme la perte d'inertie interclasse doit être la plus faible possible, nous avons coupé le dendogramme à un niveau où la hauteur des branches est élevée, c'est-à-dire pour la valeur *4,5*, ce qui correspond aux branches colorées du dendogramme.

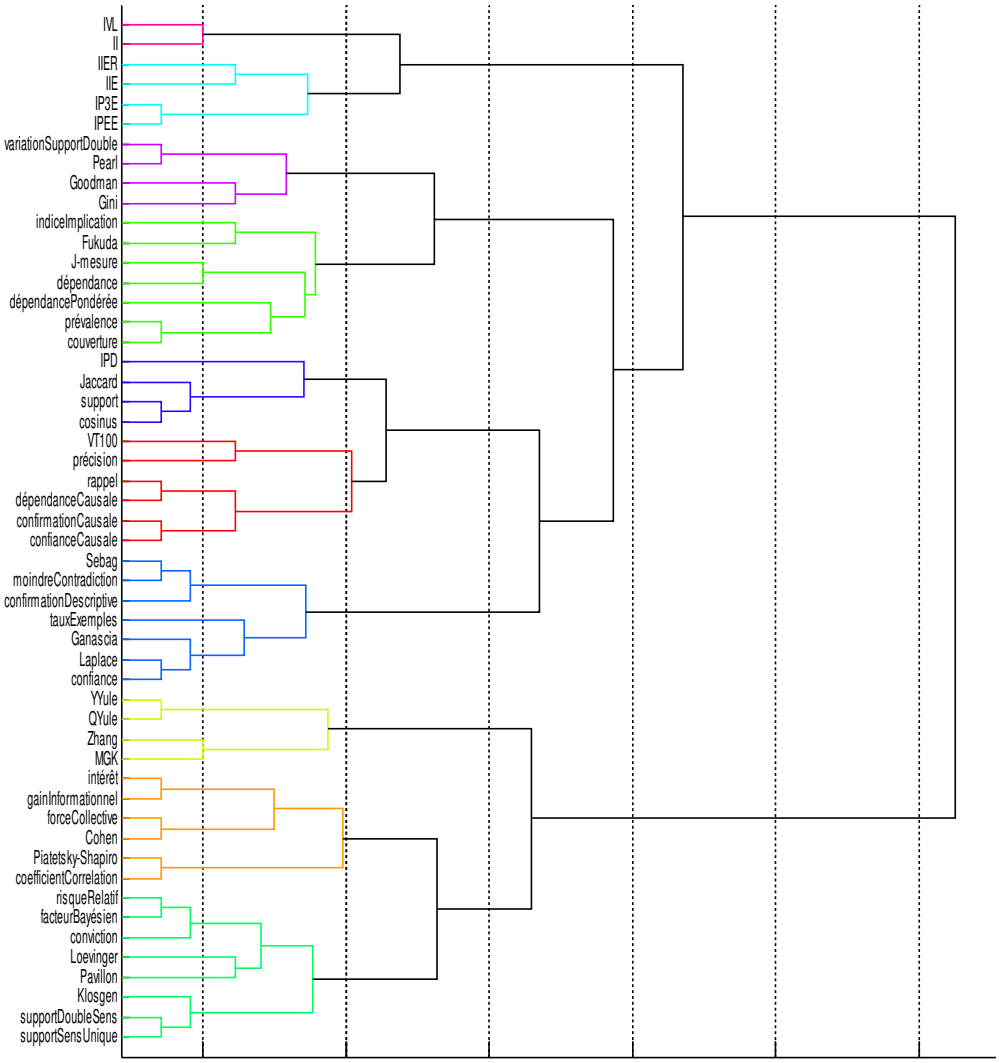


*Fig. 1 : Classification ascendante hiérarchique utilisant le critère de Ward.*

Cette classification nous révèle *10* groupes de mesures qui sont les suivants :

- $Gc_1$ = {*indice de vraisemblance du lien (IVL), intensité d'implication (II)*}
- $Gc_2$ = {*IIER, IIE, IP3E, IPEE*}
- $Gc_3$ = {*variation du support à double sens, Pearl, Goodman, Gini*}
- $Gc_4$ = {*indice d'implication, Fukuda, J-mesure, dépendance, dépendance pondérée, prévalence, couverture*}
- $Gc_5$ = {*indice probabiliste discriminant, Jaccard, support, cosinus*}
- $Gc_6$ = {*VT100, précision, rappel, dépendance causale, confirmation causale, confiance causale*}
- $Gc_7$ = {*Sebag, moindre contradiction, confirmation descriptive, taux d'exemples, Ganascia, Laplace, confiance*}
- $Gc_8$ = {*Y de Yule, Q de Yule, Zhang, $M_{GK}$*}
- $Gc_9$ = {*intérêt, gain informationnel, force collective, Cohen, Piatetsky-Shapiro, coefficient de corrélation*}
- $Gc_{10}$ = {*risque relatif, facteur bayésien, conviction, Loevinger, Pavillon, Klosgen, support à double sens, support à sens unique*}

Après avoir effectué cette première classification des mesures, nous allons confronter ces résultats avec une deuxième technique : une version de la méthode des *k*-moyennes et nous discuterons des différents résultats obtenus afin de dégager un consensus.

## 4 Classification obtenue par une version des *k*-moyennes et classification définitive

Nous avons effectué un partitionnement avec la méthode des *k*-moyennes grâce au logiciel *Matlab* en retenant également la distance euclidienne. Nous avons choisi *10* classes au vu des résultats de la *CAH* et nous avons obtenu le partitionnement suivant. Tout en présentant ces *10* nouvelles classes obtenues, nous discutons de la cohérence des résultats obtenus avec la première technique.

- $Gp_1$ = {*indice de vraisemblance du lien, intensité d'implication, IIER, IIE, IP3E, IPEE*}

Ce groupe rassemble les groupes $Gc_1$ et $Gc_2$ puisque nous avons $Gp_1 = Gc_1 \cup Gc_2$. Nous sommes en présence des indices de la famille de la vraisemblance du lien (*l'indice fondateur*).

- $Gp_2$ = {*indice d'implication, J-mesure, dépendance, dépendance pondérée, prévalence, couverture, Gini*}

Ce groupe est très proche du groupe $Gc_4$ puisque les *6* mesures suivantes sont présentes dans les deux groupes : *indice d'implication, J-mesure, dépendance, dépendance pondérée, prévalence, couverture*. Dans le groupe $Gc_4$, nous avons en plus la mesure *Fukuda* et dans ce groupe $Gp_2$, la mesure *Gini*. Nous avons donc $Gc_4 - \{Fukuda\} = Gp_2 - \{Gini\}$.

- $Gp_3$ = {*variation du support double, Pearl, Goodman*}

Ce groupe est proche du groupe $Gc_3$ puisque nous avons $Gc_3 = Gp_3 \cup \{Gini\}$.

- $Gp_4$ = {*indice probabiliste discriminant (IPD), Jaccard, support, cosinus, rappel*}

Ce groupe est similaire au groupe $Gc_5$ puisque nous avons $Gp_4 = Gc_5 \cup \{rappel\}$.

- $Gp_5$ = {*VT100, précision, dépendance causale, confirmation causale, confiance causale*}

Ce groupe est très proche du groupe $Gc_6$ puisque nous avons $Gc_6 = Gp_5 \cup \{rappel\}$.

- $Gp_6 = \{Sebag, moindre\ contradiction, confirmation\ descriptive, Fukuda\}$
- $Gp_7 = \{taux\ d'exemples, Ganascia, Laplace, confiance\}$

Le groupe $Gp_6$ est très proche du groupe $Gc_7$ puisque nous retrouvons les trois premières mesures. Le reste des mesures du groupe $Gc_7$ est présent dans le groupe $Gp_7$. Nous avons $Gc_7 = Gp_6 \cup Gp_7 - \{Fukuda\}$.

- $Gp_8 = \{Y\ de\ Yule, Q\ de\ Yule, Zhang, M_{GK}\}$

Nous retrouvons le groupe $Gc_8$ de la *CAH*. Nous avons donc $Gp_8 = Gc_8$.

- $Gp_9 = \{force\ collective, Cohen, Piatetsky\text{-}Shapiro, coefficient\ de\ corrélation\}$

Ce groupe est très proche du groupe $Gc_9$ puisque nous avons $Gp_9 \subset Gc_9$. Les deux autres mesures de $Gr_9$ (*intérêt et gain informationnel*) sont dans le groupe $Gp_{10}$ suivant. Nous constatons sur la *figure 1* que les deux groupes $Gc_9$ et $Gc_{10}$ sont relativement proches et que si nous avions augmenté le seuil de la distance de Ward, ces deux groupes de mesures auraient été rassemblés dans le même groupe.

- $Gp_{10} = \{risque\ relatif, facteur\ bayésien, conviction, Loevinger, Pavillon, Klosgen, support\ à\ double\ sens, support\ à\ sens\ unique, gain\ informationnel, intérêt\}$

Ce groupe est très proche du $Gc_{10}$ puisque les *8* mesures suivantes sont présentes dans les deux groupes : *risque relatif, facteur bayésien, conviction, Loevinger, Pavillon, Klosgen, support à double sens* et *support à sens unique*. Nous avons en plus dans le groupe $Gp_{10}$ les *2* mesures suivantes : *gain informationnel, intérêt.* Nous avons donc $Gp_{10} = Gc_{10} \cup \{gain\ informationnel, intérêt\}$.

Après cette discussion sur la cohérence des résultats obtenus par les deux techniques, nous dégageons un consensus sur la classification. La *figure 2* révèle ce consensus et nous restitue les classes $C_1$ à $C_9$ de mesures extraites communes aux deux techniques. Nous mentionnons également les mesures pour lesquelles un consensus n'a pas été trouvé et donnons les deux groupes (*ou classes*) d'appartenance de ces mesures. Nous avons étiqueté les flèches par "*c*" et "*p*" pour indiquer quelle technique (*c = classification hiérarchique ou p = partitionnement ou classification non hiérarchique*) les a rassemblé dans le groupe pointé. Pour finir, dans le cadran inférieur droit, nous rappelons les mesures identiques mais portant des noms différents.

Après avoir synthétisé les résultats obtenus (*figure 2*), nous essayons dans la section suivante de donner une sémantique aux différentes classes extraites et validons cette classification avec celle dégagée par Benoît Vaillant (Vaillant, 2007).

# 5 Étude des classes et validation

Il n'est pas aisé de donner une sémantique à chacune des classes extraites en regardant uniquement les définitions de ces mesures. Une classe reste cependant facilement interprétable, c'est la classe $C_6$ où nous retrouvons tous les indices de la famille de l'indice de vraisemblance du lien, indice fondateur. Cette classe a été subdivisée en deux autres classes $C_{61}$ (*ou* $Gc_1$) et $C_{62}$ (*ou* $Gc_2$) par la *CAH*. La première sous-classe, la classe $C_{61}$, possède les indices d'origine : l'indice de vraisemblance du lien (*IVL*) et l'intensité d'implication (*II*). Nous savons que ces deux mesures sont très proches puisque l'indice de vraisemblance du lien recherche si le nombre d'exemples (*ceux qui vérifient à la fois la prémisse et la conclusion*) est significativement élevé alors que l'intensité d'implication évalue si le nombre

de contre-exemples (*ceux qui vérifient la prémisse mais qui ne vérifient pas la conclusion*) est significativement faible. Pour la deuxième sous-classe, la classe $C_{62}$, nous retrouvons toutes les mesures d'intensité d'implication entropiques (*IIER, IIE, IP3E*) avec l'indice probabiliste d'écart à l'équilibre (*IPEE*). Ces mesures sont issues d'une idée commune : évaluer la significativité d'un nombre (*nombre d'exemples ou de contre-exemples*), en le combinant pour certaines mesures (*IIER, IIE, IP3E*) avec un indice entropique afin que la mesure soit discriminante dans le cas de données volumineuses.

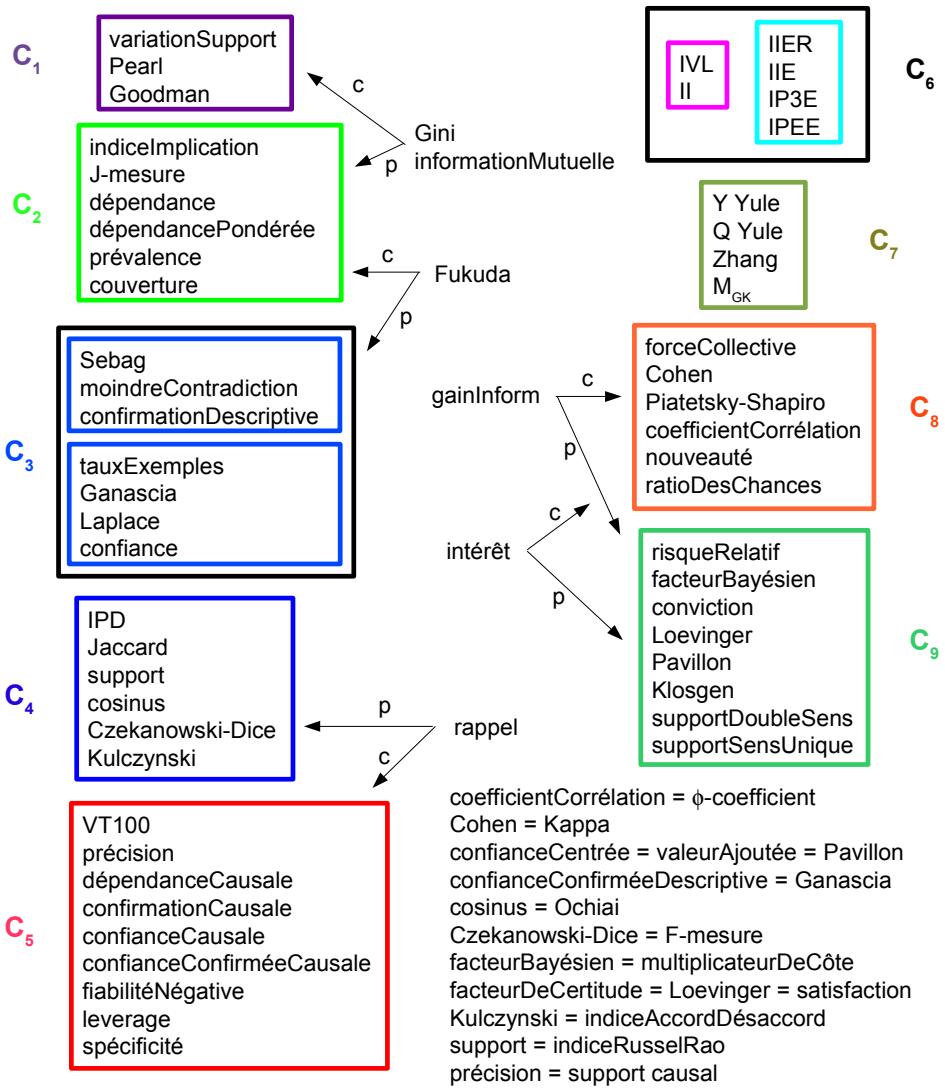


*Fig. 2 : Groupes ou classes de mesures.*

Afin de tenter d'expliquer chacune de ces classes $C_i$ ($i=1,..,9$), nous résumons dans le *tableau 1* toutes les propriétés vérifiées par chacune des *9* classes. Nous rajoutons un symbole par rapport à la matrice d'origine, le caractère "*?*", qui a la signification "*indéterminé*" c'est-à-dire que les mesures de la classe $C_i$ prennent différentes valeurs pour la propriété $P_j$ ($j=1,..,19$) concernée. Dans le cas où la propriété est contredite une seule fois, nous indiquons la valeur de la propriété majoritaire. Ainsi "*0?*" signifie que toutes les mesures de la classe $C_i$ sauf une seule mesure, prennent la valeur "*0*" pour la propriété $P_j$. Dans le *tableau 1*, nous avons fait apparaître les deux sous-classes, $C_{31}$ et $C_{32}$, de $C_3$ et également les deux sous-classes, $C_{61}$ et $C_{62}$, de $C_6$.

| Classes | $P_1$ | $P_2$ | $P_3$ | $P_4$ | $P_5$ | $P_6$ | $P_7$ | $P_8$ | $P_9$ | $P_{10}$ | $P_{11}$ | $P_{12}$ | $P_{13}$ | $P_{14}$ | $P_{15}$ | $P_{16}$ | $P_{17}$ | $P_{18}$ | $P_{19}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $C_1$ | 0 | 0 | 1 | 0 | 0 | 0 | 1? | 0 | 0 | 1? | 0 | 1? | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| $C_2$ | 1 | ? | 0? | 0 | 0? | 0 | ? | 0 | 0 | 0? | 0 | ? | 0 | 0 | 0 | 0 | 0 | 0? | ? |
| $C_{31}$ | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | ? | 0 | 0 | ? | 0 | 0 | 0 | 1 |
| $C_{32}$ | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1? | 1 | 0 | 0 | ? | 0 | 0 | ? | 0 | 0 | 0 | ? |
| $C_3$ | 1 | 1 | ? | 1 | 0 | 0 | 0 | ? | 1 | 0 | 0 | ? | 0 | 0 | ? | 0 | 0 | 0 | ? |
| $C_4$ | 0 | 1 | 0 | 1 | 0 | 1? | 0 | 0 | 0 | 0 | 0 | ? | 0? | 0 | 0 | 0 | 0? | 0? | 1 |
| $C_5$ | ? | 1 | ? | 1 | 1 | 1 | 0 | ? | 0 | 0 | 0 | 1 | 0 | 0? | 0? | ? | 0? | 0 | 1 |
| $C_{61}$ | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | ? | 0 | 0 | 0 | 1 | 1 | 0 |
| $C_{62}$ | 1 | 1 | 1 | 1 | ? | 0 | 0? | 0 | ? | 0? | 0? | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1? |
| $C_6$ | ? | 1 | 1 | 1 | ? | ? | ? | 0 | ? | ? | ? | 1 | ? | 0 | 0 | 0 | 1 | 1 | ? |
| $C_7$ | ? | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | ? | 1? | ? | 1 | ? | 0 | 0 | 1 |
| $C_8$ | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | ? | 0 | ? | ? | 1 | 0 | 0? | 1 |
| $C_9$ | 1? | 1 | ? | ? | 1 | 1? | 1 | 0? | 0 | 1 | 1 | ? | 0? | 0 | 0? | 0 | 0 | 0 | 1 |

*Tab 1 : Caractéristiques des 9 classes détectées.*

En résumant l'ensemble des propriétés vérifiées par chacune des *9* classes dans ce tableau, nous aidons l'utilisateur dans le choix de ses mesures puisqu'il n'a plus qu'à consulter une matrice beaucoup moins importante que celle d'origine. De plus, s'il souhaite des mesures très différentes, son choix est également facilité avec la consultation de ce tableau, aide complétée par le dendogramme de la *figure 1* où apparaît une notion de proximité entre les mesures. Pour finir, cette classification peut également empêcher de choisir des mesures trop similaires en évitant de prendre des indices appartenant à la même classe.

Quant à la recherche d'une sémantique pour chacune de ces classes, ce tableau synthétique n'est pas d'une grande aide. Nous pouvons expliquer l'association de certaines mesures au sein d'une classe mais mettre une sémantique pour l'ensemble des mesures n'est pas aisé. Ainsi par exemple, nous pouvons expliquer l'association des mesures $M_{GK}$ et *Zhang* au sein de la classe $C_7$ (*mesures normalisées dont les valeurs sont comprises entre -1 et 1 avec des valeurs fixes égales à -1, 0 et 1 pour respectivement l'incompatibilité*[8]*, l'indépendance et l'implication logique*) ainsi que l'association de *Y* et *Q* de *Yule* (*mesures avec des formules très similaires où apparaissent les mêmes termes, la différence venant du*

[8] cas où lorsque *X* est réalisé, *Y* ne l'est pas.

*fait que Y prend les racines carrés de chacun des termes*) dans également la classe $C_7$ mais le regroupement de ces *4* mesures au sein de la même classe n'est pas facilement interprétable. Le dendogramme de la *figure 1* met bien en évidence la proximité de ces *4* mesures.

Pour finir, nous comparons la classification obtenue avec celle de Benoît Vaillant (Vaillant, 2007) qui a fait son étude sur *20* mesures selon *9* propriétés formelles. Sur ces *9* propriétés, nous en avons *7* en commun car les propriétés "*compréhensibilité de la mesure*" et "*facilité à fixer un seuil d'acceptation*" ont été écartées par (Guillaume et al., 2010) car jugées trop subjectives. Pour effectuer sa classification, Benoît Vaillant a également utilisé le critère de Ward mais il a retenu la distance de Manhattan. L'auteur fait remarquer qu'en utilisant d'autres critères, il a obtenu des résultats semblables. Il a dégagé les *5* classes suivantes :

$Cl_1$={*support, moindre contradiction, Laplace*},
$Cl_2$={*confiance, Sebag, taux exemples*},
$Cl_3$={*coefficient corrélation, Piatetsky-Shapiro, Pavillon, intérêt, indice d'implication, Cohen, gain informationnel*},
$Cl_4$={*Loevinger, facteur bayésien, conviction*} et
$Cl_5$={*Zhang, IIET, intensité d'implication, indice probabiliste discriminant*}.

Nous pouvons assimiler la mesure *IIET* avec la mesure *IIER* car le but de ces deux mesures est le même.

Nous sommes en accord sur les regroupements suivants :

{*moindre contradiction, Laplace*}$\subset Cl_1 \subset C_3$, $Cl_2 \subset C_3$, $Cl_4 \subset C_9$,

{*coefficient corrélation, Piatetsky-Shapiro*} $\subset Cl_3 \subset C_8$ et {*IIER, II*}$\subset Cl_5 \subset C_6$.

Nous sommes également en accord avec le regroupement suivant mais avec une seule des techniques étudiées (*k-moyennes*) :

{*Pavillon, intérêt*} $\subset Cl_3 \subset Cp_{10}$.

Nous avons étudié *12* propriétés supplémentaires, ce qui explique que nous ne retrouvons pas tous les résultats de Benoît Vaillant.

Nous sommes bien conscients que la catégorisation des mesures peut aussi dépendre de plusieurs facteurs parmi lesquels : les données, l'expert-utilisateur, la nature des règles extraites et la procédure de recherche des classes, comme le souligne Suzuki (2008).

Afin d'éviter le biais des données, de l'expert et de la nature des règles extraites, nous avons ici fait le choix d'une étude théorique basée sur des propriétés de mesures (Guillaume et al. 2010), plutôt que sur des données expérimentales (Huynh et al. 2005). Les deux aspects sont bien évidemment complémentaires.

Pour éviter le biais de la procédure de construction de classes, nous avons utilisé deux techniques de classification qui de manière générale ont exhibé de fortes ressemblances entre de nombreuses mesures, et fait ressortir des similitudes et des différences avec des travaux précédents (Vaillant 2007).

Cette étude vient compléter des travaux précédents sur la description d'une vision unificatrice des mesures d'intérêt (Hébert et Cremilleux, 2007), et apporte une contribution supplémentaire à l'analyse de ces mesures.

# 6 Conclusion et perspectives

Cet article a pris comme point de départ un travail de synthèse sur les mesures d'intérêt présentes dans la littérature pour extraire des connaissances et les propriétés jugées pertinentes pour celles-ci. Ce travail de synthèse a conduit à l'évaluation de *19* propriétés jugées intéressantes sur *61* mesures. L'objectif de cet article est la classification de ces mesures afin d'aider l'utilisateur dans le choix de ses mesures complémentaires au couple (*support, confiance*) afin d'éliminer les règles inintéressantes. Dans un premier temps, nous avons analysé ces données (*matrice de 61 mesures × 19 propriétés*) afin de déterminer si une simplification n'était pas envisageable en recherchant tout d'abord tous les groupes de mesures au comportement totalement identique et en détectant ensuite si des propriétés n'étaient pas redondantes. Nous avons détecté *7* groupes de mesures au comportement totalement identique ce qui a permis de réduire nos données de départ pour la recherche de la classification grâce à deux techniques : une méthode de la classification ascendante hiérarchique et une version de la méthode des *k*-moyennes. Les classifications obtenues grâce aux deux techniques ont permis de trouver un consensus : *9* classes qui ont été en partie validées par une classification existante.

Dans le futur, nous souhaiterions conforter les classes de mesures obtenues en étudiant les *N* meilleures règles extraites dans des bases de données différentes et par chacune des mesures afin de vérifier que cet ensemble des *N* meilleures règles est sensiblement le même dans chacune des classes. Pour finir, il serait intéressant de prendre en compte de plus petites classes (*en nous aidant du dendogramme extrait*) afin d'attribuer une sémantique à chacune d'elles, ce qui serait une aide précieuse pour l'utilisateur (*plutôt qu'un ensemble de propriétés vérifiées*), car nous avons pu constater notre incapacité à définir en quelques mots ou phrases chacune de ces classes extraites. Des propriétés complémentaires seraient peut-être à envisager. La notion de robustesse des règles d'association (Le Bras et al. 2010) pourrait aussi être envisagée dans le processus de catégorisation de mesures d'intérêt.

# Références

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th VLDB Conference*, Santiago, Chile, pp. 487-499.

Ben Yahia, S., G. Gasmi, et E. Mephu Nguifo: A new generic basis of "factual" and "implicative" association rules. *Intelligent Data Analysis journal.* 13(4): 633-656 (2009)

Blanchard, J., F. Guillet, et H. Briand (2003). A user-driven and quality-oriented visualization for mining association rules. In 3rd *ICDM*, pp. 493–496. IEEE Computer Society Press, Los Alamitos.

Carvalho, D.R., AA. Freitas et N.F.F. Ebecken (2005). Evaluating the Correlation Between Objective Rule Interestingness Measures and Real Human Interest. In *PKDD*. LNCS 3721, pp. 453–461. Springer, Heidelberg .

Feno, D.J. (2007). *Mesures de qualité des règles d'association : normalisation et caractérisation des bases.* PhD thesis, Université de La Réunion.

Geng, L. et H.J. Hamilton (2007). Choosing the Right Lens: Finding What is Interesting in Data Mining. In *Quality Measures in Data Mining,* pp. 3–24, ISBN 978-3-540-44911-9.

Guillaume, S., D. Grissa et E. Mephu Nguifo (2009). Propriétés des mesures d'intérêt pour l'extraction des règles. Rapport de recherche LIMOS, RR-09-10, 22 pages, 31 décembre 2009.

Guillaume, S., D. Grissa et E. Mephu Nguifo (2010). Propriétés des mesures d'intérêt pour l'extraction des règles. In *Actes de l'atelier QDC de la conférence* EGC, pp. 15-28, Hammamet, Tunisie.

Hébert, C., et B. Crémilleux (2007). A Unified View of Objective Interestingness Measures. *MLDM* 2007, pp. 533-547.

Hofmann, H. et A. Wilhelm (2001). Visual comparison of association rules. *Computational Statistics*, 16(3) pp. 399–415.

Huynh, X.-H, F. Guillet et H. Briand (2005). Clustering Interestingness Measures with Positive Correlation. In *Proceedings of 7th ICEIS*, pp. 248–253.

Lallich, S. et O. Teytaud (2004). Evaluation et validation de mesures d'intérêt des règles d'association. In *Mesures de Qualité pour la Fouille de Données 2004*, Volume RNTI-E-1, pp. 193-217. Cépaduès.

Le Bras, Y., P. Meyer, P. Lenca et S. Lallich (2010). A robustness measure of association rules. In *ECML/PKDD*, 2, pp. 227-242, Springer.

Lenca, P., P. Meyer, B. Vaillant, et S. Lallich (2008). On Selecting Interestingness Measures for Association Rules: User Oriented Description and Multiple Criteria Decision Aid. *European Journal of Operational Research*. 184(2), pp. 610-626.

Sese, J. et S. Morishita (2002). Answering the most correlated n association rules efficiently. In *Proceedings of the 6th PKDD*, pp. 410–422. Springer-Verlag.

Suzuki, E. (2008). Pitfalls for Categorizations of Objective Interestingness Measures for Rule Discovery. In *Statistical Implicative Analysis: Theory and Applications*, 127, pp. 383–395. Springer.

Tan, P.N., V. Kumar, et J. Srivastava (2002). Selecting the right interestingness measure for association patterns. In *8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pp. 32-41.

Vaillant, B. (2007). *Mesurer la qualité des règles d'association : études formelles et expérimentales*. PhD thesis, ENST Bretagne.

Zaki, M. (2000). Generating non-redundant association rules. In *Knowledge Discovery and Data Mining*, pp. 34–43.

Zaman Ashrafi, M., D. Taniar, et K. Smith (2004). A new approach of eliminating redundant association rules. In *DEXA*, LNCS 3180, pp. 465–474, Zaragoza, Spain. Springer.

# Summary

Finding interesting association rules is an important and active research field in data mining. The algorithms of the Apriori family are based on two measures to extract the rules, support and confidence. Although these two measures have accelerators algorithmic virtues, they generate a prohibitive number of rules most of which are redundant and irrelevant It is therefore a need for further measures filtering uninteresting rules. Different reported works were realized to identify "*good*" measures properties for extraction rules and these properties were assessed on *61* measures. The aim of this paper is to identify categories of measures able to reply to users concern: the choice of one measure or more during the knowledge extraction process in order to eliminate valid and irrelevant rules extracted by the pair (support, confidence). The properties evaluation on the *61* measures identifies *9* classes of measures, classes obtained through two techniques : AHC according to the Ward criterion and the clustering k-means method.