# 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 *7* 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, 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 par rapport à l'ensemble des propriétés que nous avons dégagées 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 certaines des classes extraites et valide la classification retenue avec celles dégagées par Benoît Vaillant, Marie-Jeanne Lesot et Maria Rifgi, et pour finir Djamel Zighed, Rafik

[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.

Abdesselam et Ahmed Bounekkar (Vaillant 2007, Lesot et Rifqi 2010, Zighed et al. 2011). L'article se termine par une conclusion et des perspectives.

# 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. Le lecteur pourra trouver la formalisation de ces différentes propriétés en *annexe 1* pour une meilleure compréhension de celles-ci. Ces propriétés sont résumées dans le *tableau 1* suivant.

| N° | Propriétés |
|---|---|
| **$P_1$** | La mesure *m* est non symétrique ($P_1(m)=1$) ou symétrique ($P_1(m)=0$). |
| **$P_2$** | *m* n'égalise pas les règles antinomiques ($P_2(m)=1$) ou les égalise ($P_2(m)=0$). |
| **$P_3$** | *m* évalue de la même façon les règles $X \rightarrow Y$ et $\overline{Y} \rightarrow \overline{X}$ dans le cas de l'implication logique ($P_3(m)=1$) ou non ($P_3(m)=0$). |
| **$P_4$** | *m* est croissante en fonction du nombre d'exemples ($P_4(m)=1$) ou non croissante ($P_4(m)=0$). |
| **$P_5$** | *m* est croissante en fonction du nombre d'individus ($P_5(m)=1$) ou non ($P_5(m)=0$). |
| **$P_6$** | *m* décroissante en fonction de la taille de la conclusion ($P_6(m)=1$) ou non ($P_6(m)=0$). |
| **$P_7$** | *m* a une valeur fixe dans le cas de l'indépendance ($P_7(m)=1$) ou non ($P_7(m)=0$). |
| **$P_8$** | *m* a une valeur fixe dans le cas de l'implication logique ($P_8(m)=1$) ou non ($P_8(m)=0$). |
| **$P_9$** | *m* a une valeur fixe dans le cas de l'équilibre ($P_9(m)=1$) ou non ($P_9(m)=0$). |
| **$P_{10}$** | Valeurs identifiables en cas d'attraction entre *X* et *Y* ($P_{10}(m)=1$) ou non ($P_{10}(m)=0$). |
| **$P_{11}$** | Valeurs identifiables en cas de répulsion entre *X* et *Y* ($P_{11}(m)=1$) ou non ($P_{11}(m)=0$). |
| **$P_{12}$** | *m* est tolérante aux premiers contre-exemples ($P_{12}(m)=2$), non tolérante ($P_{12}(m)=0$) ou indifférente ($P_{12}(m)=1$). |
| **$P_{13}$** | *m* invariante en cas de dilatation de certains effectifs ($P_{13}(m)=1$) ou non ($P_{13}(m)=0$). |
| **$P_{14}$** | *m* oppose les règles $X \rightarrow Y$ et $\overline{X} \rightarrow Y$ ($P_{14}(m)=1$) ou non ($P_{14}(m)=0$). |
| **$P_{15}$** | *m* oppose les règles antinomiques ($P_{15}(m)=1$) ou non ($P_{15}(m)=0$). |
| **$P_{16}$** | *m* égalise les règles $X \rightarrow Y$ et $\overline{X} \rightarrow \overline{Y}$ ($P_{16}(m)=1$) ou non ($P_{16}(m)=0$). |
| **$P_{17}$** | *m* est fondée sur un modèle probabiliste ($P_{17}(m)=1$) ou non ($P_{17}(m)=0$). |
| **$P_{18}$** | *m* est statistique ($P_{18}(m)=1$) ou descriptive ($P_{18}(m)=0$). |
| **$P_{19}$** | *m* est discriminante ($P_{19}(m)=1$) ou non ($P_{19}(m)=0$). |

*Tab 1 : Propriétés des mesures m.*

Quelques précisions concernant la terminologie donnée dans le *tableau 1* précédent :
- exemple : individu qui vérifie à la fois la prémisse *X* de la règle et la conclusion *Y*,
- indépendance : cas où la réalisation de *X* n'augmente pas les chances d'apparition de *Y*,
- implication logique : cas où la probabilité conditionnelle P(*Y*/*X*) est égale à *1*,
- équilibre ou indétermination : cas où lorsque *Y* est réalisé, il y a autant de chances que *X* ou *non X* soit réalisé,
- attraction : lorsque la réalisation de *X* augmente les chances d'apparition de *Y*,
- répulsion : lorsque la réalisation de *X* diminue les chances d'apparition de *Y*.

Les *61* mesures étudiées dans (Guillaume et al., 2010) sont résumées dans le *tableau 2* et regroupées en deux catégories : les mesures symétriques et non symétriques. Les expressions de chacun des indices, accompagnées de leur référence, sont disponibles dans (Guillaume et al., 2009) et les définitions sont rappelées en *annexe 2* dans le *tableau 6*. Les *61* mesures du *tableau 6* étant ordonnées par ordre alphabétique, le numéro des mesures donné dans le *tableau 2* permet de faciliter la recherche de sa définition.

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, ce qui correspond aux branches colorées du dendogramme.

| **Mesures symétriques** | | | | | |
|---|---|---|---|---|---|
| 1 | coefficient de corrélation | 2 | Cohen ou Kappa | 11 | cosinus ou Ochiai |
| 13 | Czekanowski | 20 | force collective | 22 | gain informationnel |
| 24 | Goodman | 33 | indice de vraisemblance du lien (*IVL*) | 34 | intérêt |
| 35 | Jaccard | 38 | Kulczynski | 43 | nouveauté |
| 44 | Pearl | 45 | Piatetsky-Shapiro | 46 | précision |
| 48 | Q de Yule | 50 | ratio des chances | 54 | support |
| 56 | support à double sens | 58 | VT100 | 59 | variation du support |
| 60 | Y de Yule | | | | |
| **Mesures non symétriques** | | | | | |
| 3 | confiance | 4 | confiance causale | 5 | Pavillon |
| 6 | Ganascia | 7 | confiance confirmée causale | 8 | confirmation causale |
| 9 | confirmation descriptive | 10 | conviction | 12 | couverture |
| 14 | dépendance | 15 | dépendance causale | 16 | dépendance pondérée |
| 17 | facteur bayésien | 18 | facteur de certitude | 19 | fiabilité négative |
| 21 | Fukuda | 23 | Gini | 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*) | 36 | J-mesure | 37 | Klosgen |
| 39 | Laplace | 40 | leverage | 41 | mesure de Guillaume-Khenchaf $M_{GK}$ |
| 42 | moindre contradiction | 47 | prévalence | 49 | rappel |
| 51 | risque relatif | 52 | Sebag-Schoenauer | 53 | spécificité |
| 55 | support à sens unique | 57 | taux d'exemples | 61 | Zhang |

*Tab 2 : Mesures étudiées.*

Nous aurions pu également choisir la distance de Manhattan et nous aurions obtenu des résultats sensiblement similaires car la matrice est essentiellement binaire : *18* variables binaires sur *19*, et dans ce cas, la distance de Manhattan est la distance euclidienne au carré. Seule une variable présente *3* valeurs : la propriété $P_{12}$.

Cette classification nous révèle *8* groupes de mesures qui sont les suivants :

- $Gc_1$ = {*indice de vraisemblance du lien (IVL), intensité d'implication (II)*}
- $Gc_2$ = {*IIER, IIE, IPD, IP3E, IPEE*}
- $Gc_3$ = {*variation du support à double sens, Pearl*}
- $Gc_4$ = {*indice d'implication, Fukuda, Gini, J-mesure, dépendance, dépendance pondérée, prévalence, couverture*}
- $Gc_5$ = {*VT100, précision, Jaccard, support, cosinus, rappel, dépendance causale, confirmation causale, confiance causale*}
- $Gc_6$ = {*Sebag, moindre contradiction, confirmation descriptive, taux d'exemples, Ganascia, Laplace, confiance*}
- $Gc_7$ = {*Zhang, $M_{GK}$, Y de Yule, Q de Yule, Goodman, Piatetsky-Shapiro, coefficient de corrélation*}
- $Gc_8$ = {*intérêt, gain informationnel, force collective, Cohen, risque relatif, facteur bayésien, conviction, facteur de certitude, 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 *8* classes au vu des résultats de la *CAH* et nous avons obtenu le partitionnement suivant. Tout en présentant ces *8* 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*}

Ce groupe est très proche du groupe $Gc_1$ puisque nous avons $Gp_1 = Gc_1 \cup \{IIER\}$.

- $Gp_2$ = {*IIE, IPD, IP3E, IPEE*}

Ce groupe est très proche du groupe $Gc_2$ puisque nous avons $Gp_2 = Gc_2 - \{IIER\}$. Nous avons l'égalité suivante : $Gp_1 \cup Gp_2 = Gc_1 \cup Gc_2$ , ce qui montre une certaine cohérence dans les résultats obtenus puisque nous sommes en présence de tous les indices de la famille de la vraisemblance du lien.

- $Gp_3$ = {*variation du support double, Pearl, indice d'implication, Gini, J-mesure, dépendance, prévalence, couverture*}

Ce groupe est proche du groupe $Gc_4$ puisque nous avons :

$Gp_3 = Gc_3 \cup Gc_4 \cup \{Fukuda, dépendance\ pondérée\}$. Il est à noter que le groupe $Gc_3$ , composé des mesures *variation du support double* et *Pearl*, est le groupe le plus proche du groupe $Gc_4$ (*voir le dendogramme de la figure 1*).

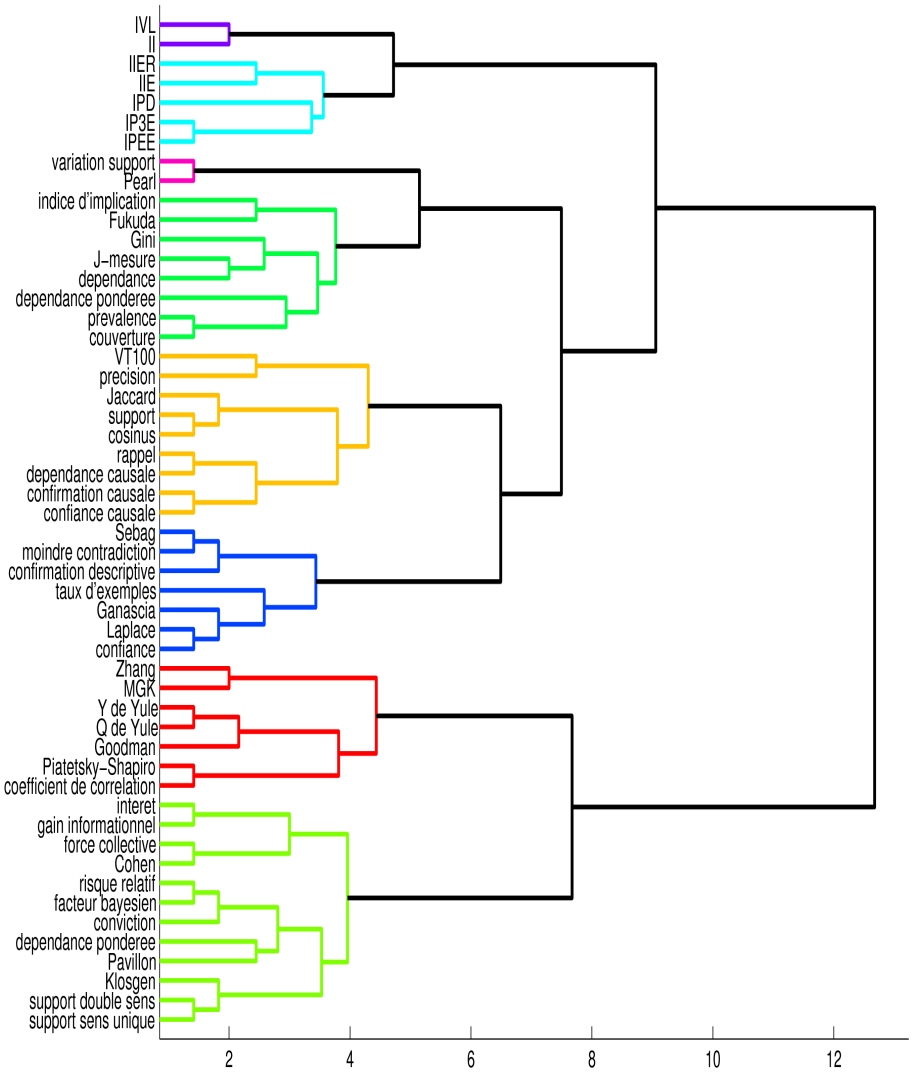


*Fig. 1 : Classification ascendante hiérarchique utilisant le critère de Ward.*

- $Gp_4$ = {*précision, Jaccard, support, cosinus, rappel, dépendance causale, confirmation causale, confiance causale*}

Ce groupe est similaire au groupe $Gc_5$ puisque nous avons :

$Gp_4 = Gc_5 \cup$ {*Fukuda*, *dépendance pondérée*} - {*VT100*}.

- $Gp_5$ = {*Sebag, moindre contradiction, confirmation descriptive, Fukuda*}

Ce groupe est identique au groupe $Gc_6$.

- $Gp_6$ = {*Zhang,* $M_{GK}$, *Y de Yule, Q de Yule,*}

Ce groupe est similaire au groupe $Gc_7$ puisque nous avons :

$Gc_7 = Gp_6 \cup$ {*Piatetsky-Shapiro*, *coefficient de corrélation*}

- $Gp_7$ = {*intérêt, gain informationnel, risque relatif, facteur bayésien, conviction, facteur de certitude, Pavillon, Klosgen, support à double sens, support à sens unique*}

Le groupe $Gp_7$ est très proche du groupe $Gc_8$ puisque nous avons *10* mesures en commun sur *12*. Nous avons l'égalité suivante : $Gc_8 = Gp_7 \cup$ {*force collective, Cohen*}.

- $Gp_8$ = {*VT100, Piatetsky-Shapiro, coefficient de corrélation, force collective, Cohen*}

Contrairement aux autres groupes $Gp_i$ ($i$ = {*1*, .., *7*}), ce groupe n'est similaire à aucun des groupes $Gc_j$ ($j$ = {*1*, .., *8*}) puisque ces *5* mesures sont issues des groupes $Gc_5$, $Gc_7$ et $Gc_8$.

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 $C_7$ de mesures extraites communes aux deux techniques. Nous mentionnons également les mesures pour lesquelles un consensus n'a pas été trouvé et donnons, lorsque c'est possible, les deux groupes (*ou classes*) d'appartenance de ces mesures. Nous avons étiqueté les flèches par "*c*" et "*p*" pour indiquer quelle technique les a rassemblé dans le groupe pointé (*c = classification hiérarchique ou p = partitionnement ou classification non hiérarchique*). 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 à certaines classes extraites et validons cette classification avec celles dégagées par Benoît Vaillant, Marie-Jeanne Lesot et Maria Rifgi, et pour finir Djamel Zighed, Rafik Abdesselam et Ahmed Bounekkar (Vaillant 2007, Lesot et Rifqi 2010, Zighed et al. 2011).

# 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. Deux classes restent cependant facilement interprétables, ce sont les classes $C_1$ et $C_2$ où nous retrouvons tous les indices de la famille de l'indice de vraisemblance du lien (Lerman, 1970), indice fondateur. La classe $C_1$ possède les indices d'origine : l'indice de vraisemblance du lien (*IVL*) et l'intensité d'implication (*II*) (Gras, 1979). 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 classe $C_2$, nous retrouvons les mesures d'intensité d'implication entropiques (*IIE* (*Gras et al., 2001*), *IP3E* (*Blanchard et al., 2005b*)) avec l'indice probabiliste d'écart à l'équilibre (*IPEE* (*Blanchard et al., 2005a*))

ainsi que l'indice probabiliste discriminant *IPD* (Lerman et Azé, 2007). 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* (*Lallich et al., 2005*)*, IIE, IP3E*) avec un indice entropique afin que la mesure soit discriminante dans le cas de données volumineuses. Quant à l'*IPD*, cet indice normalise l'intensité d'implication afin que celle-ci soit discriminante dans le cas de données volumineuses en évaluant une règle par rapport à l'ensemble des règles valides.

Afin de tenter d'expliquer chacune de ces classes $C_i$ ($i=1,..,7$), nous résumons dans le *tableau 3* toutes les propriétés vérifiées par chacune des *7* 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 $P_j$.

En résumant l'ensemble des propriétés vérifiées par chacune des *7* 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.

| **Classes** | **$P_1$** | **$P_2$** | **$P_3$** | **$P_4$** | **$P_5$** | **$P_6$** | **$P_7$** | **$P_8$** | **$P_9$** | **$P_{10}$** | **$P_{11}$** | **$P_{12}$** | **$P_{13}$** | **$P_{14}$** | **$P_{15}$** | **$P_{16}$** | **$P_{17}$** | **$P_{18}$** | **$P_{19}$** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $C_1$ | **?** | **1** | **1** | **1** | **1** | **1** | **1** | **0** | **0** | **1** | **1** | **2** | **?** | **0** | **0** | **0** | **1** | **1** | **0** |
| $C_2$ | **1** | **1** | **1** | **1** | **0?** | **1?** | **0** | **0** | **?** | **0** | **0** | **2** | **0?** | **0** | **0** | **0** | **1** | **1** | **1?** |
| $C_3$ | **1** | **?** | **0?** | **0** | **0** | **0** | **?** | **0** | **0** | **0?** | **0** | **?** | **0** | **0** | **0** | **?** | **0** | **0?** | **?** |
| $C_4$ | **?** | **1** | **?** | **1?** | **?** | **1?** | **0** | **?** | **0** | **0** | **0** | **?** | **0** | **0** | **0** | **0?** | **0** | **0** | **1** |
| $C_5$ | **1** | **1** | **?** | **1** | **0** | **0** | **0** | **?** | **1** | **0** | **0** | **?** | **0** | **0** | **?** | **0** | **0** | **0** | **1?** |
| $C_6$ | **?** | **1** | **1** | **1** | **1** | **0** | **1** | **1** | **0** | **1** | **1** | **?** | **?** | **?** | **1** | **?** | **0** | **0** | **1** |
| $C_7$ | **?** | **1** | **?** | **?** | **1** | **1?** | **1** | **0?** | **0** | **1** | **1** | **?** | **0?** | **0** | **0?** | **0** | **0** | **0** | **1** |

*Tab 3 : Caractéristiques des 7 classes détectées.*

Quant à la recherche d'une sémantique pour chacune de ces classes, ce tableau synthétique peut être le support à une interprétation comme nous allons l'illustrer pour les classes $C_4$ et $C_6$. Nous allons donc nous focaliser sur ces classes et tenter d'en donner une interprétation. Nous commençons par la classe $C_6$.

## 5.1 Étude de la classe $C_6$

La classe $C_6$ est composée des cinq mesures suivantes : *Zhang* (Zhang, 2000), $M_{GK}$ (Guillaume, 2000), *Y* et *Q* de *Yule* (Yule, 1900) et Goodman (Tan et al., 2002). Nous savons d'après le *tableau 3* que celles-ci vérifient les propriétés suivantes :

- non symétrie au sens de la négation de la conclusion ($P_2=1$),
- évaluation identique de $X \rightarrow Y$ et $\overline{Y} \rightarrow \overline{X}$ dans le cas de l'implication logique ($P_3=1$),
- croissance en fonction du nombre d'exemples ($P_4=1$),

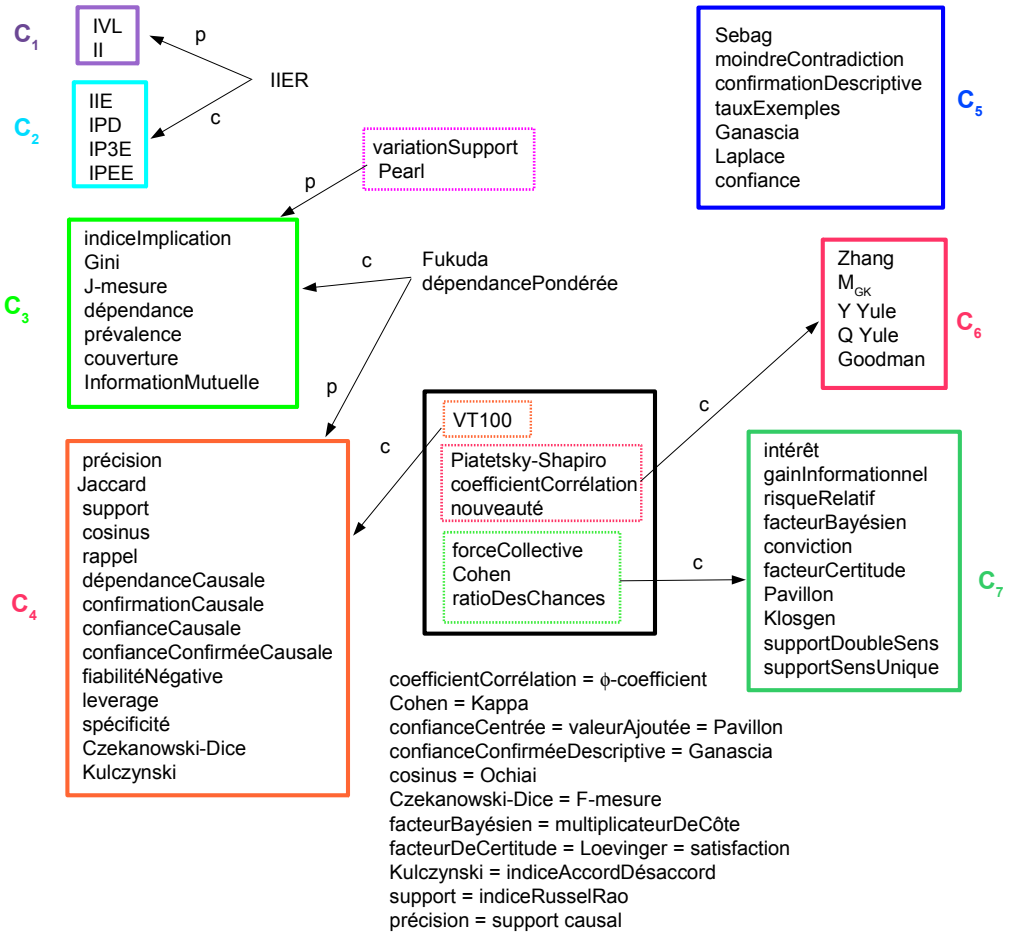


*Fig. 2 : Groupes ou classes de mesures.*

- croissance en fonction de la taille des données ($P_5=1$),
- valeur fixe dans le cas de l'indépendance ($P_7=1$),
- valeur fixe dans le cas de l'implication logique ($P_8=1$),
- valeurs identifiables lorsque la réalisation de la prémisse augmente les chances d'apparition de la conclusion ($P_{10}=1$),
- valeurs identifiables lorsque la réalisation de la prémisse diminue les chances d'apparition de la conclusion ($P_{11}=1$),
- valeurs opposées pour les règles antinomiques $X \rightarrow Y$ et $X \rightarrow \overline{Y}$ ($P_{15}=1$),
- discriminante dans le cas de données volumineuses ($P_{19}=1$).

Grâce à l'ensemble des propriétés vérifiées, nous pouvons donner une première sémantique à cette classe $C_6$. Ces mesures sont des indices normalisés puisqu'ils prennent des valeurs fixes pour les cas d'indépendance ($P_7=1$) et d'implication logique ($P_8=1$) et que les valeurs prises par ces indices permettent de savoir si la règle est dans la zone attractive ($P_{10}=1$) ou dans la zone répulsive ($P_{11}=1$).

La *figure 3* permet de vérifier cette première sémantique donnée à ces indices. Nous avons tracé l'évolution des cinq mesures lorsque le nombre d'exemples augmente passant ainsi de l'état d'incompatibilité (*aucun individu vérifie à la fois la prémisse et la conclusion ou encore $n_{XY} = 0$, avec $n_{XY}$ le nombre d'individus vérifiant à la fois la prémisse X et la conclusion Y*) jusqu'à l'implication logique (*l'ensemble des individus vérifiant la prémisse est inclus dans l'ensemble des individus vérifiant la conclusion ou encore $n_{XY} = n_X$ avec $n_X$ le nombre d'individus vérifiant la prémisse X*). Nous avons également indiqué sur la *figure 3* les trois états caractéristiques d'une règle : l'incompatibilité, l'indépendance et l'implication logique ainsi que les zones d'attraction et de répulsion. La taille de l'ensemble prémisse retenue pour réaliser ces courbes est de *174*, la taille de l'ensemble conclusion est de *400* et pour finir la taille de l'ensemble des données est de *600* ($n_X = 174$, $n_Y = 400$ *et* $n = 600$). Nous aurions pu choisir des tailles différentes pour ces différents ensembles et aurions obtenu des courbes similaires avec la contrainte suivante respectée : $n_X \leq n_Y \leq n$.

Cette *figure 3* ainsi qu'une étude complémentaire nous permettent d'affiner la sémantique apportée à cette classe $C_6$. Ce sont des 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, l'indépendance et l'implication logique. De plus, elles ont non seulement des valeurs identifiables dans la zone d'attraction et de répulsion mais ces valeurs sont comprises entre *0* et *1* dans la zone d'attraction et ces valeurs sont comprises entre *-1* et *0* dans la zone de répulsion. Pour finir, le signe de la mesure nous renseigne sur la zone d'appartenance de la règle. Nous pouvons donc en déduire que ces mesures évaluent une certaine distance par rapport à l'indépendance : distance entre l'indépendance et l'implication logique dans le cas de valeurs positives et une distance entre l'indépendance et l'incompatibilité dans le cas de valeurs négatives.

| **Mesure** | **$P_1$** | **$P_2$** | **$P_3$** | **$P_4$** | **$P_5$** | **$P_6$** | **$P_7$** | **$P_8$** | **$P_9$** | **$P_{10}$** | **$P_{11}$** | **$P_{12}$** | **$P_{13}$** | **$P_{14}$** | **$P_{15}$** | **$P_{16}$** | **$P_{17}$** | **$P_{18}$** | **$P_{19}$** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Zhang | **1** | **1** | **1** | **1** | **1** | **0** | **1** | **1** | **0** | **1** | **1** | **2** | **1** | **0** | **1** | **0** | **0** | **0** | **1** |
| $M_{GK}$ | **1** | **1** | **1** | **1** | **1** | **0** | **1** | **1** | **0** | **1** | **1** | **1** | **0** | **0** | **1** | **0** | **0** | **0** | **1** |
| Y Yule | **0** | **1** | **1** | **1** | **1** | **0** | **1** | **1** | **0** | **1** | **1** | **0** | **1** | **1** | **1** | **1** | **0** | **0** | **1** |
| Q Yule | **0** | **1** | **1** | **1** | **1** | **0** | **1** | **1** | **0** | **1** | **1** | **2** | **1** | **1** | **1** | **1** | **0** | **0** | **1** |
| Goodman | **0** | **1** | **1** | **1** | **1** | **0** | **1** | **1** | **0** | **1** | **1** | **1** | **0** | **1** | **1** | **1** | **0** | **0** | **1** |
| Classe 6 | **?** | **1** | **1** | **1** | **1** | **0** | **1** | **1** | **0** | **1** | **1** | **?** | **?** | **?** | **1** | **?** | **0** | **0** | **1** |

*Tab 4 : Évaluation des propriétés sur les mesures de la classe 6.*

Lorsque nous observons la *figure 1* révélant la classification ascendante hiérarchique en utilisant le critère de *Ward*, nous avons une proximité plus forte entre les indices *Y*, *Q* et *Goodman*, et une proximité également plus forte entre *Zhang* et $M_{GK}$. Les discordances mises en évidence dans le *tableau 3*, c'est-à-dire les cas où nous trouvons le symbole "*?*" pour les propriétés étudiées, peuvent nous renseigner sur ces deux proximités plus prononcées entre les mesures. Le *tableau 4* détaille les différentes propriétés vérifiées par les cinq mesures de ce groupe et rappelle les caractéristiques générales de cette classe. La première propriété où ce symbole apparaît et qui permet d'expliquer ces deux proximités est la symétrie des mesures ($P_1$). *Y*, *Q* et *Goodman* sont des mesures symétriques (*évaluation identique des règles symétriques $X \rightarrow Y$ et $Y \rightarrow X$ : $P_1 = 0$*) alors que les mesures *Zhang* et $M_{GK}$ sont des

mesures non symétriques (*évaluation différente des règles symétriques* $X \rightarrow Y$ *et* $Y \rightarrow X$ *:* $P_1 = 1$).

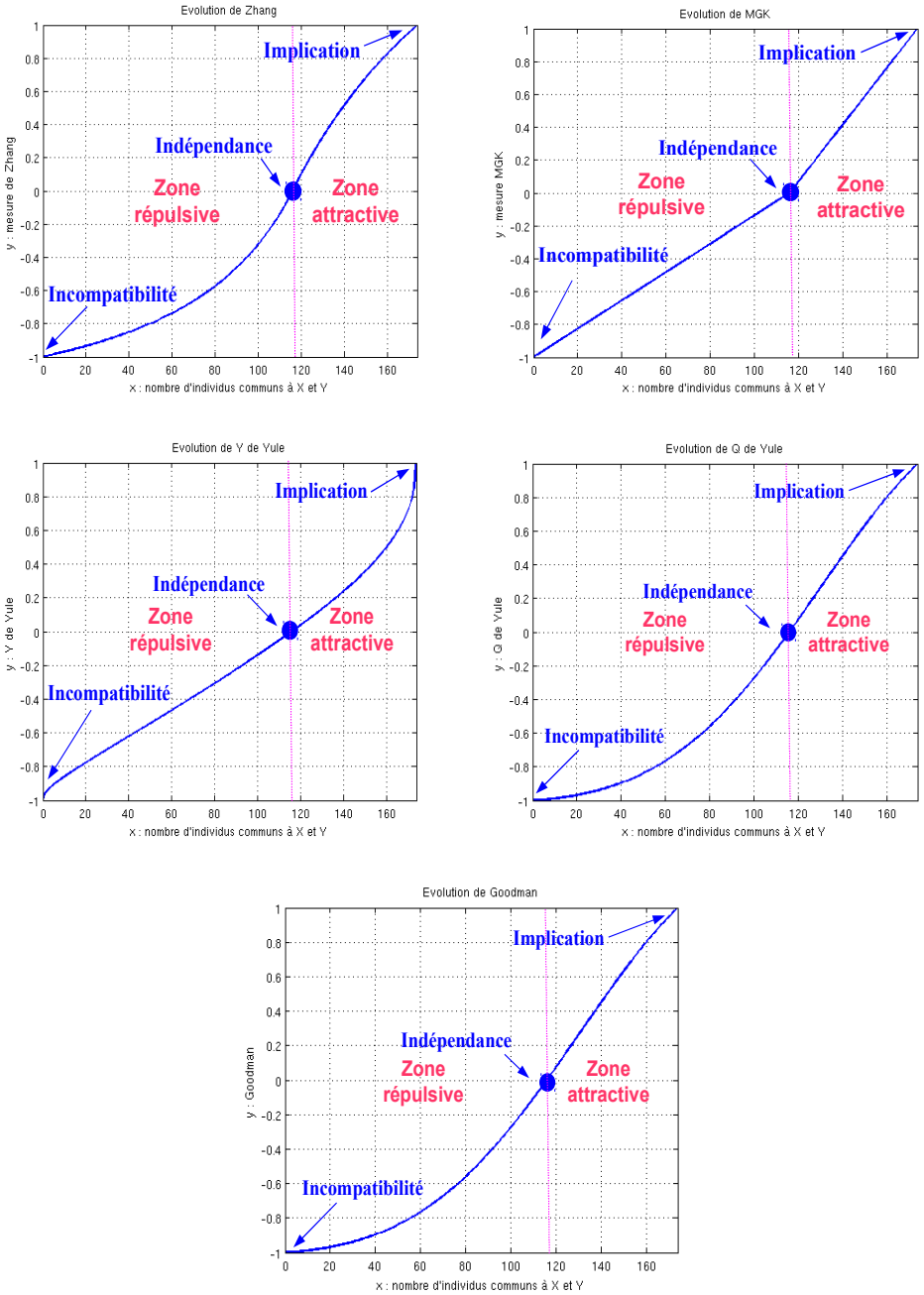


*Fig. 3 : Évolution des cinq mesures de la classe* $C_6$ *en fonction du nombre d'exemples.*

Les propriétés $P_{14}$ (*valeurs opposées ou non pour les règles* $X \to Y$ *et* $\overline{X} \to Y$) et $P_{16}$ (*valeurs identiques ou non pour les règles* $X \to Y$ *et* $\overline{X} \to \overline{Y}$) permettent également d'expliquer ces deux proximités. Les indices *Y, Q* et *Goodman* ont des valeurs opposées pour les règles $X \to Y$ et $\overline{X} \to Y$ et des valeurs identiques pour les règles $X \to Y$ et $\overline{X} \to \overline{Y}$. Les mesures *Zhang* et $M_{GK}$ vérifient la négation de ces deux dernières propriétés.

Nous allons maintenant faire une étude de la classe $C_4$.

## 5.2 Étude de la classe $C_4$

La classe $C_4$ contient les indices suivants : *précision* (Tan et al., 2002), *Jaccard* (Jaccard, 1908), *support* (Russel et Rao, 1940), *cosinus* (Ochiai, 1957), *rappel* (Lavrac et al., 1999), *dépendance causale* (Tan et al., 2002), *confiance causale* (Kodratoff, 2001), *confiance confirmée causale* (Kodratoff, 2001), *fiabilité négative* (Lavrac et al., 1999), *leverage* (Piatetsky-Shapiro, 1991 ), *spécificité* (Tan et al., 2002), *Czekanowski-Dice* (Czekanowski, 1913) et *Kulczynski* (Kulczynski, 1928).

D'après le *tableau 3*, ces *14* mesures vérifient les *12* propriétés suivantes :
- non symétrie au sens de la négation de la conclusion ($P_2=1$),
- discriminante dans le cas de données volumineuses ($P_{19}=1$),
- valeurs non fixes dans le cas de l'indépendance ($P_7=0$) et de l'équilibre ($P_9=0$),
- valeurs non identifiables en cas d'attraction ($P_{10}=0$) et de répulsion ($P_{11}=0$),
- non invariante en cas de dilatation de certains effectifs ($P_{13}=0$),
- deux relations entre les différentes règles négatives n'est pas présente ($P_{14}=0$) ($P_{15}=0$),
- non fondée sur un modèle probabiliste ($P_{17}=0$),
- mesures descriptives ($P_{18}=0$).

Étudions maintenant les propriétés vérifiées par presque toutes les mesures à l'exception d'une seule :
- croissance en fonction du nombre d'exemples ($P_4=1$) à l'exception du *support*,
- croissance en fonction de la taille de la conclusion ($P_6=1$) à l'exception du *support*,
- mesures n'égalisant pas les règles $X \to Y$ et $\overline{X} \to \overline{Y}$ ($P_{16}=0$) à l'exception de la *précision*.

Vu le nombre relativement important de mesures présentes dans cette classe (*c'est la classe dont la cardinalité est la plus élevée*), il est difficile de trouver une sémantique aussi précise que pour la classe précédente $C_6$. Nous pouvons cependant en donner une pour un ensemble plus restreint de mesures : *Jaccard*, *support*, *cosinus*, *Czekanowski-Dice*, *Kulczynski* et *rappel*. Ces mesures sont fonction de *P*(*XY*) et symétriques (*à l'exception du rappel*). Nous rappelons les expressions de ces *6* mesures :

- *Jaccard* : $\frac{P(XY)}{P(X)+P(Y)-P(XY)} = \frac{P(XY)}{P(X\overline{Y})+P(Y)}$

- *support* : $P(XY)$

- *cosinus* : $\frac{P(XY)}{\sqrt{P(X)P(Y)}}$

- *Czekanowski-Dice* : $\frac{2P(XY)}{P(X)+P(Y)}$

- *Kulczynski* : $\frac{P(XY)}{P(X\overline{Y})+P(\overline{X}Y)}$

- *rappel* : $\frac{P(XY)}{P(Y)}$ .

Nous pouvons donc en déduire que ces mesures auront une valeur fixe égale à *0* dans le cas de l'incompatibilité ($P(XY) = 0$). Nous comprenons également la non croissance trouvée en fonction de la taille de l'ensemble des données ($P_5=0$) à la vue de ces différentes formules comme le montre le *tableau 5* qui restitue les propriétés vérifiées par ces *6* mesures. Nous avons une invariance de ces mesures (*sauf pour le support*) en fonction de la taille *n* de l'ensemble des données puisque cela revient à augmenter la probabilité $P(\overline{XY})$ . Quant au *support*, il est décroissant en fonction de la taille *n* de l'ensemble des données.

| **Mesure** | **$P_1$** | **$P_2$** | **$P_3$** | **$P_4$** | **$P_5$** | **$P_6$** | **$P_7$** | **$P_8$** | **$P_9$** | **$P_{10}$** | **$P_{11}$** | **$P_{12}$** | **$P_{13}$** | **$P_{14}$** | **$P_{15}$** | **$P_{16}$** | **$P_{17}$** | **$P_{18}$** | **$P_{19}$** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Jaccard | **0** | **1** | **0** | **1** | **0** | **1** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **1** |
| support | **0** | **1** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **1** | **0** | **0** | **0** | **0** | **0** | **0** | **1** |
| cosinus | **0** | **1** | **0** | **1** | **0** | **1** | **0** | **0** | **0** | **0** | **0** | **1** | **0** | **0** | **0** | **0** | **0** | **0** | **1** |
| rappel | **1** | **1** | **0** | **1** | **0** | **1** | **0** | **0** | **0** | **0** | **0** | **1** | **0** | **0** | **0** | **0** | **0** | **0** | **1** |
| Czekanowski | **0** | **1** | **0** | **1** | **0** | **1** | **0** | **0** | **0** | **0** | **0** | **1** | **0** | **0** | **0** | **0** | **0** | **0** | **1** |
| Kulczynski | **0** | **1** | **0** | **1** | **0** | **1** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **1** |
| sous-ensemble | **0?** | **1** | **0** | **1?** | **0** | **1?** | **0** | **0** | **0** | **0** | **0** | **?** | **0** | **0** | **0** | **0?** | **0** | **0** | **1** |

*Tab 5 : Évaluation des propriétés d'un sous-ensemble de mesures de la classe 4.*

Comme pour la classe précédente $C_6$, nous allons étudier l'évolution de ces différentes mesures en fonction du nombre d'exemples. La *figure 4* restitue cette évolution. Nous avons retenu les mêmes cardinalités que précédemment pour les ensembles prémisse, conclusion et l'ensemble des données ($n_X = 174$, $n_Y = 400$ et $n = 600$).

Nous vérifions la valeur nulle prise par ces mesures dans le cas de l'incompatibilité. Nous obtenons deux types de courbes :
- une droite pour les mesures *support*, *cosinus*, *Czekanowski-Dice* et *rappel*,
- une demi-parabole pour les mesures *Jaccard* et *Kulczynski*.

Après avoir étudié plus précisément certaines classes et tenté de donner une interprétation à celles-ci, nous validons maintenant notre travail par une comparaison avec des classifications existantes (Vaillant 2007, Lesot et Rifqi 2010, Zighed et al. 2011).

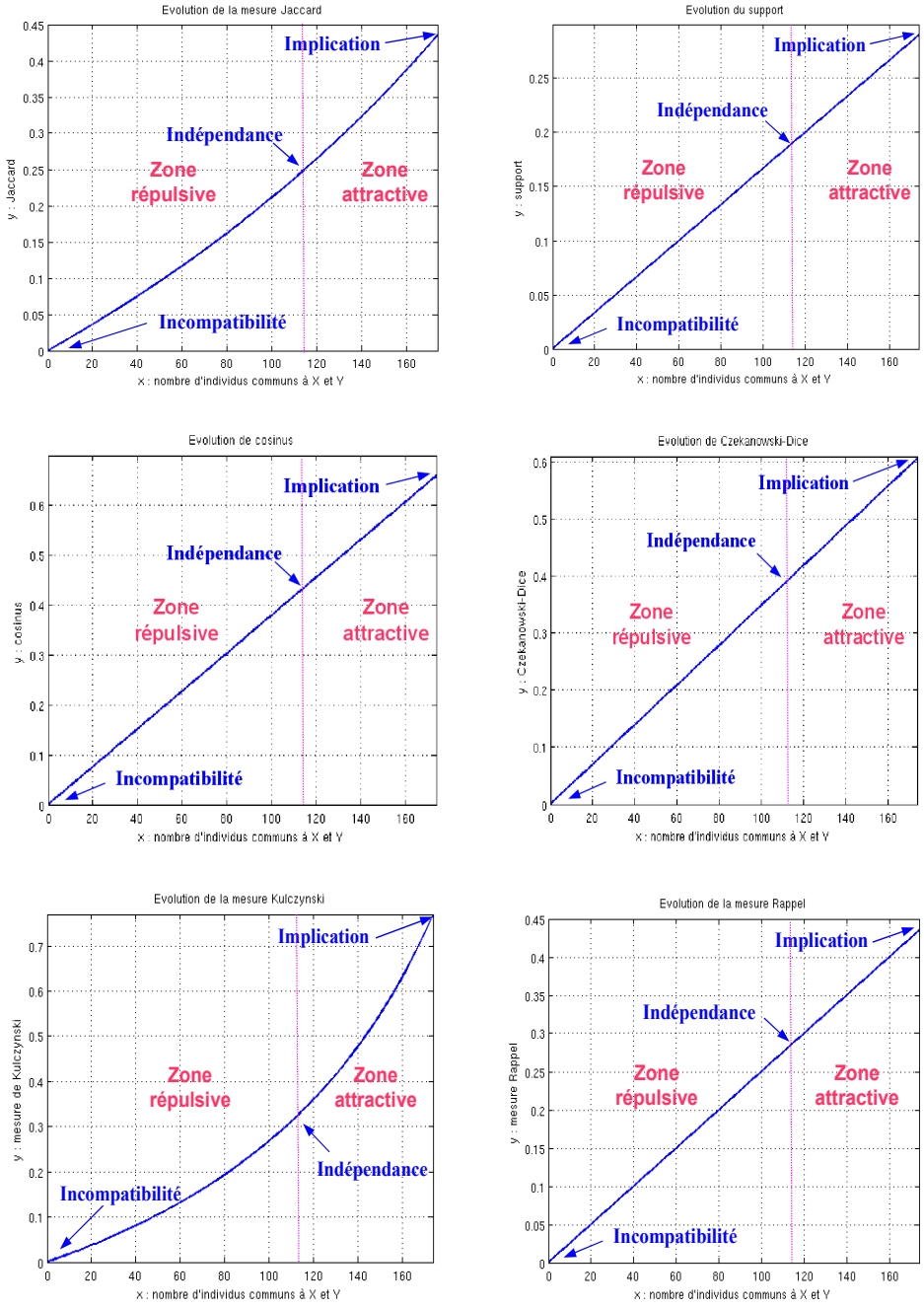


*Fig. 4 : Évolution des six mesures de la classe $C_4$ en fonction du nombre d'exemples.*

### 5.3 Validation

Nous comparons tout d'abord 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), celles-ci étant 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 :

$ClBV_1$={*support, moindre contradiction, Laplace*},
$ClBV_2$={*confiance, Sebag, taux exemples*},
$ClBV_3$={*coefficient corrélation, Piatetsky-Shapiro, Pavillon, intérêt, indice d'implication, Cohen, gain informationnel*},
$ClBV_4$={*Loevinger, facteur bayésien, conviction*} et
$ClBV_5$={*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 :
$ClBV_2 \subset C_5$ , $ClBV_4 \subset C_7$ ,
et nous avons les relations suivantes entre les groupes :
$ClBV_1 - \{support\} \subset C_5$, $ClBV_3 - \{indice\ implication\} \subset Gp_8 \cup C_7$ et
$ClBV_5 - \{Zhang\} \subset C_1 \cup C_2$

Le groupement où le désaccord est le plus important est le groupe $ClBV_3$ puisque nous avons du faire apparaître le groupe $Gp_8$ présent uniquement avec l'une des techniques : une version des k-moyennes. Quant au groupe $ClBV_5$, il regroupe, à l'exception de la mesure *Zhang*, toutes les mesures de la famille de l'intensité d'implication.
Nous avons étudié *12* propriétés supplémentaires, ce qui explique que nous ne retrouvons pas tous les résultats de Benoît Vaillant.

Une autre classification a été effectuée sur des mesures de distance et de similarité par Marie-Jeanne Lesot et Maria Rifgi (Lesot et Rifqi, 2010). Les auteurs ont étudié l'ordre induit par les mesures et non pas les valeurs numériques obtenues puisque leur contexte d'étude est la recherche d'information. Cette étude a porté sur des mesures dédiées aux données binaires et aux données numériques en réalisant des expérimentations sur à la fois des données réelles et sur des données artificielles. Les auteurs ont obtenu une liste de mesures équivalentes (*mesures qui induisent toujours le même ordre*) et pour les mesures non équivalentes, les auteurs ont quantifié ce désaccord par un degré d'équivalence basé sur le coefficient de Kendall généralisé. Sur les *10* mesures étudiées et destinées aux données binaires, *5* sont communes à nos deux études. Ces mesures sont les suivantes : *Czekanowski-Dice*, *Jaccard*, *Ochiai* et *Q* et *Y* de *Yule*. Les auteurs ont trouvé que les mesures *Q* et *Y* de *Yule* sont des mesures équivalentes. Ce résultat est également confirmé par notre étude puisque ces deux mesures sont dans la classe $C_6$ et comme nous l'avons déjà mentionné, très

proche sur le dendogramme de la *figure 1*. Les auteurs ont également trouvé que les mesures *Czekanowski-Dice* et *Jaccard* sont des mesures équivalentes. Ces deux mesures ont été également affectées à la même classe : la classe $C_4$, et nous les retrouvons avec une proximité relativement importante dans le dendogramme de la *figure 1* (*nous avons choisi la mesure cosinus comme mesure représentative sur le dendogramme puisque comme nous l'avons vu dans la section 2, les mesures cosinus et Czekanowski-Dice ont des valeurs identiques pour les 19 propriétés ce qui a conduit à la constitution du groupe* $G_3$). Pour finir, nous avons regroupé également la mesure *Ochiai* (*ou cosinus*) avec les mesures *Czekanowski-Dice* et *Jaccard*, dans la classe $C_4$. Les auteurs (Lesot et Rifqi, 2010) ont trouvé un degré d'équivalence entre la mesure *Ochiai* et la classe d'équivalence {*Czekanowski-Dice*, *Jaccard*} de *0.99*, ce qui conforte nos résultats.

Une dernière classification a été proposée par Djamel Zighed, Rafik Abdesselam et Ahmed Bounekkar (Zighed et al., 2011) sur *13* mesures de proximité dont seulement deux mesures sont communes à nos deux études : le *cosinus* et le *coefficient de corrélation*. Cette classification proposée par les auteurs est basée sur l'équivalence topologique et fait appel à la structure de voisinage local. Les deux mesures sont apparues très proches dans cette classification, contrairement à nos travaux puisque nous les retrouvons dans les classes $C_4$ et $Gp_8$. L'ensemble des mesures étudiées étant tellement différent, les classes trouvées par chacune des techniques ne peuvent être que difficilement comparables. De plus, comme les auteurs l'ont souligné lors de la présentation de leur travaux, la classification qu'ils ont obtenue est faiblement représentative puisque réalisée sur un seul jeu de données : les *Iris* de *Fisher*.

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, Lesot et Rifqi 2010, Zighed et al. 2011).

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 : *7* classes qui ont été en partie validées par des classifications existantes.

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.

**Remerciements**. Nous remercions Israël-César Lerman pour ses remarques constructives concernant cet article. De plus, ce travail est partiellement soutenu par le projet franco-tunisien PHC Utique 11G1417 : EXQUI (*EXtraction, QUalité et Ingénierie des connaissances dans les environnements hétérogènes*).

# Annexe 1 : Propriétés des mesures

**<u>Propriété 1</u>** : **Mesure non symétrique**

| | |
|---|---|
| $P_1(m) = 0$ si $m$ est symétrique | i.e. si $\forall\ X \to Y\ m(X \to Y) = m(Y \to X)$ |
| $P_1(m) = 1$ si $m$ est non symétrique | i.e. si $\exists\ X \to Y\ /\ m(X \to Y) \neq m(Y \to X)$ |

**<u>Propriété 2</u>** : **Mesure non symétrique au sens de la négation de la conclusion ou Mesure n'égalisant pas les règles antinomiques**

| | |
|---|---|
| $P_2(m) = 0$ si $m$ est *nc*-symétrique | i.e. si $\forall\ X \to Y\ m(X \to Y) = m(X \to \overline{Y}\ )$ |
| $P_2(m) = 1$ si $m$ est non *nc*-symétrique | i.e. si $\exists\ X \to Y\ /\ m(X \to Y) \neq m(X \to \overline{Y}\ )$ |

**<u>Propriété 3</u>** : **Mesure évaluant de la même façon $X \rightarrow Y$ et $\overline{Y} \rightarrow \overline{X}$ dans le cas de l'implication logique.**

$P_3(m) = 0$ si $\exists\, X \rightarrow Y \,/\, P(Y/X) = 1$ et $m(X \rightarrow Y) \neq m(\overline{Y} \rightarrow \overline{X})$
$P_3(m) = 1$ si $\forall\, X \rightarrow Y\; P(Y/X) = 1 \Rightarrow m(X \rightarrow Y) = m(\overline{Y} \rightarrow \overline{X})$

**<u>Propriété 4</u>** : **Mesure croissante en fonction du nombre d'exemples ou décroissante en fonction du nombre de contre-exemples.**

$P_4(m) = 0$ si $m$ n'est pas croissante en fonction de $n_{XY}$ i.e. si $\exists\, X_1 \rightarrow Y_1$ , $\exists\, X_2 \rightarrow Y_2$ /
$n_{X_1} = n_{X_2}$ et $n_{Y_1} = n_{Y_2}$ et ( $n_{X_1} n_{Y_1} < n_{X_2} n_{Y_2}$ ou $n_{X_1} n_{\overline{Y_1}} > n_{X_2} n_{\overline{Y_2}}$ ) et $m(X_1 \rightarrow Y_1) > m(X_2 \rightarrow Y_2)$
$P_4(m) = 1$ si $m$ est croissante en fonction de $n_{XY}$ i.e. si $\forall\, X_1 \rightarrow Y_1$ , $\forall\, X_2 \rightarrow Y_2$
[ $n_{X_1} = n_{X_2}$ et $n_{Y_1} = n_{Y_2}$ et $(n_{X_1 Y_1} < n_{X_2 Y_2}$ ou $n_{X_1 \overline{Y_1}} > n_{X_2 \overline{Y_2}})$ ] $\Rightarrow m(X_1 \rightarrow Y_1) \leq m(X_2 \rightarrow Y_2)$ et
$\exists\, X_1 \rightarrow Y_1$ , $\exists\, X_2 \rightarrow Y_2$ / $n_{X_1} = n_{X_2}$ et $n_{Y_1} = n_{Y_2}$ et $(n_{X_1 Y_1} < n_{X_2 Y_2}$ ou $n_{X_1 \overline{Y_1}} > n_{X_2 \overline{Y_2}})$
et $m(X_1 \rightarrow Y_1) < m(X_2 \rightarrow Y_2)$

**<u>Propriété 5</u>** : **Mesure croissante en fonction de la taille de l'ensemble des données**

$P_5(m) = 0$ ($m$ *pas croissante en fonction de* $n$) si $\exists\, (\Omega_1, \Omega_2)\; \exists\, X_1 \rightarrow Y_1\, (\Omega_1)$, $\exists\, X_2 \rightarrow Y_2\, (\Omega_2)$ /
$n_{X_1} = n_{X_2}$ et $n_{Y_1} = n_{Y_2}$ et $n_{X_1 Y_1} = n_{X_2 Y_2}$ et $n_1 < n_2$ et $m(X_1 \rightarrow Y_1) > m(X_2 \rightarrow Y_2)$
$P_5(m) = 1$ ($m$ *croissante en fonction de* $n$) si $\forall\, \Omega_1$ , $\forall\, \Omega_2$, $\forall\, X_1 \rightarrow Y_1\, (\Omega_1)$, $\forall\, X_2 \rightarrow Y_2\, (\Omega_2)$,
( $n_{X_1} = n_{X_2}$ et $n_{Y_1} = n_{Y_2}$ et $n_{X_1 Y_1} = n_{X_2 Y_2}$ et $n_1 < n_2$ ) $\Rightarrow m(X_1 \rightarrow Y_1) \leq m(X_2 \rightarrow Y_2)$ et
$\exists\, \Omega_1, \exists\, \Omega_2, \exists\, X_1 \rightarrow Y_1\, (\Omega_1)$, $\exists\, X_2 \rightarrow Y_2\, (\Omega_2)$ /
$n_{X_1} = n_{X_2}$ et $n_{Y_1} = n_{Y_2}$ et $n_{X_1 Y_1} = n_{X_2 Y_2}$ et $n_1 < n_2$ et $m(X_1 \rightarrow Y_1) < m(X_2 \rightarrow Y_2)$

**<u>Propriété 6</u>** : **Mesure décroissante en fonction de la taille de la conclusion ou de la taille de la prémisse.**

$P_6(m) = 0$ si $m$ n'est pas décroissante en fonction de $n_Y$ i.e. si $\exists\, X_1 \rightarrow Y_1$ , $\exists\, X_2 \rightarrow Y_2$ /
$n_{X_1} = n_{X_2}$ et $n_{X_1 Y_1} = n_{X_2 Y_2}$ et $n_{Y_1} < n_{Y_2}$ et $m(X_1 \rightarrow Y_1) < m(X_2 \rightarrow Y_2)$
$P_6(m) = 1$ si $m$ est décroissante en fonction de $n_Y$ i.e. si $\forall\, X_1 \rightarrow Y_1$, $\forall\, X_2 \rightarrow Y_2$
( $n_{X_1} = n_{X_2}$ et $n_{X_1 Y_1} = n_{X_2 Y_2}$ et $n_{Y_1} < n_{Y_2}$ ) $\Rightarrow m(X_1 \rightarrow Y_1) \geq m(X_2 \rightarrow Y_2)$ et
$\exists\, X_1 \rightarrow Y_1$ , $\exists\, X_2 \rightarrow Y_2$ / $n_{X_1} = n_{X_2}$ et $n_{X_1 Y_1} = n_{X_2 Y_2}$ et $n_{Y_1} < n_{Y_2}$ et $m(X_1 \rightarrow Y_1) > m(X_2 \rightarrow Y_2)$

**<u>Propriété 7</u>** : **Valeur fixe $a$ dans le cas de l'indépendance**

$P_7(m) = 0$ si $\forall\, a \in \Re\;\; \exists\, X \rightarrow Y \,/\, P(Y/X) = P(Y)$ et $m(X \rightarrow Y) \neq a$
$P_7(m) = 1$ (*valeur fixe*) si $\exists\, a \in \Re \,/\, \forall\, X \rightarrow Y\;\; P(Y/X) = P(Y) \Rightarrow m(X \rightarrow Y) = a$

**Propriété 8** : **Valeur fixe *b* dans le cas de l'implication logique**

$P_8(m) = 0$ si $\forall\, b \in \Re\ \exists\, X \to Y / P(Y/X) = 1$ et $m(X \to Y) \neq b$
$P_8(m) = 1$ (*valeur fixe*) si $\exists\, b \in \Re / \forall\, X \to Y\ P(Y/X) = 1 \Rightarrow m(X \to Y) = b$

**Propriété 9** : **Valeur fixe *c* dans le cas de l'équilibre (*ou indétermination*)**

$P_9(m) = 0$ si $\forall\, c \in \Re\ \exists\, X \to Y / P(Y/X) = P(X)/2$ et $m(X \to Y) \neq c$
$P_9(m) = 1$ (*valeur fixe*) si $\exists\, c \in \Re / \forall\, X \to Y\ P(Y/X) = P(X)/2 \Rightarrow m(X \to Y) = c$

**Propriété 10** : **Valeurs identifiables en cas d'attraction entre *X* et *Y***

$P_{10}(m) = 0$ si $\forall\, a \in \Re\ \exists\, X \to Y / P(Y/X) > P(Y)$ et $m(X \to Y) \leq a$
$P_{10}(m) = 1$ (*valeurs identifiables*) si $\exists\, a \in \Re / \forall\, X \to Y\ P(Y/X) > P(Y) \Rightarrow m(X \to Y) > a$

**Propriété 11** : **Valeurs identifiables en cas de répulsion entre *X* et *Y***

$P_{11}(m) = 0$ si $\forall\, a \in \Re\ \exists\, X \to Y / P(Y/X) < P(Y)$ et $m(X \to Y) \geq a$
$P_{11}(m) = 1$ (*valeurs identifiables*) si $\exists\, a \in \Re /\ \forall\, X \to Y\ P(Y/X) < P(Y) \Rightarrow m(X \to Y) < a$

**Propriété 12** : **Tolérance aux premiers contre-exemples**

$P_{12}(m) = 0$ si rejet donc convexe, $\exists\, min_{conf} \in [0, 1] / \forall\, X_1 \to Y_1\ \forall\, X_2 \to Y_2\ \forall\, \lambda \in [0, 1]$

$$n_{X_1Y_1} \geq min_{conf}\, n_{X_1} \text{ et } n_{X_2Y_2} \geq min_{conf}\, n_{X_2}$$
$$\Rightarrow f_{m,n_{XY}}(\lambda n_{X_1Y_1} + (1-\lambda) n_{X_2Y_2}) \leq \lambda f_{m,n_{XY}}(n_{X_1Y_1}) + (1-\lambda) f_{m,n_{XY}}(n_{X_2Y_2})$$

$P_{12}(m) = 1$ si indifférence donc notamment linéaire i.e. $P_{14}(m) \neq 0$ et $P_{14}(m) \neq 2$
$P_{12}(m) = 2$ si tolérance donc concave $\exists\, min_{conf} \in [0, 1] / \forall\, X_1 \to Y_1\ \forall\, X_2 \to Y_2\ \forall\, \lambda \in [0, 1]$

$$n_{X_1Y_1} \geq min_{conf}\, n_{X_1} \text{ et } n_{X_2Y_2} \geq min_{conf}\, n_{X_2}$$
$$\Rightarrow f_{m,n_{XY}}(\lambda n_{X_1Y_1} + (1-\lambda) n_{X_2Y_2}) \geq \lambda f_{m,n_{XY}}(n_{X_1Y_1}) + (1-\lambda) f_{m,n_{XY}}(n_{X_2Y_2})$$

**Propriété 13** : **Invariance en cas de dilatation de certains effectifs**

$P_{13}(m) = 0$ (*variance*) si $\exists\, (k_1, k_2) \in \mathbb{N}^{*2}, \exists\, X_1 \to Y_1\ , \exists\, X_2 \to Y_2\ /$
$[\ n_{X_1Y_1} = k_1 n_{X_2Y_2}$ et $n_{X_1\overline{Y_1}} = k_1 n_{X_2\overline{Y_2}}$ et $n_{\overline{X_1}Y_1} = k_2 n_{\overline{X_2}Y_2}$ et $n_{\overline{X_1}\overline{Y_1}} = k_2 n_{\overline{X_2}\overline{Y_2}}$ et $m(X_1 \to Y_1) \neq m(X_2 \to Y_2)\ ]$ ou $[\ n_{X_1\overline{Y_1}} = k_1 n_{X_2\overline{Y_2}}$ et $n_{\overline{X_1}\overline{Y_1}} = k_1 n_{\overline{X_2}\overline{Y_2}}$ et $n_{X_1Y_1} = k_2 n_{X_2Y_2}$ et $n_{\overline{X_1}Y_1} = k_2 n_{\overline{X_2}Y_2}$ et $m(X_1 \to Y_1) \neq m(X_2 \to Y_2)\ ]$
$P_{13}(m) = 1$ (*invariance*) si $\forall\, (k_1, k_2) \in \mathbb{N}^{*2}\ , \forall\, X_1 \to Y_1\ , \forall\, X_2 \to Y_2$
$[\ (n_{X_1Y_1} = k_1 n_{X_2Y_2}$ et $n_{X_1\overline{Y_1}} = k_1 n_{X_2\overline{Y_2}}$ et $n_{\overline{X_1}Y_1} = k_2 n_{\overline{X_2}Y_2}$ et $n_{\overline{X_1}\overline{Y_1}} = k_2 n_{\overline{X_2}\overline{Y_2}})$
$\Rightarrow m(X_1 \to Y_1) = m(X_2 \to Y_2)\ ]$ et $[\ (n_{X_1\overline{Y_1}} = k_1 n_{X_2\overline{Y_2}}$ et $n_{\overline{X_1}\overline{Y_1}} = k_1 n_{\overline{X_2}\overline{Y_2}}$ et $n_{X_1Y_1} = k_2 n_{X_2Y_2}$ et $n_{\overline{X_1}Y_1} = k_2 n_{\overline{X_2}Y_2})$
$\Rightarrow m(X_1 \to Y_1) = m(X_2 \to Y_2)\ ]$

**<u>Propriété 14</u>** : **Opposition des règles $X \to Y$ et $\overline{X} \to Y$**

$P_{14}(m) = 0$ si $\exists\, X \to Y$ / $m(\overline{X} \to Y) \neq -m(X \to Y)$
$P_{14}(m) = 1$ si $\forall\, X \to Y$ $m(\overline{X} \to Y) = -m(X \to Y)$

**<u>Propriété 15</u>** : **Opposition des règles antinomiques**

$P_{15}(m) = 0$ si $\exists\, X \to Y$ / $m(X \to \overline{Y}) \neq -m(X \to Y)$
$P_{15}(m) = 1$ si $\forall\, X \to Y$ $m(X \to \overline{Y}) = -m(X \to Y)$

**<u>Propriété 16</u>** : **Égalité entre les règles $X \to Y$ et $\overline{X} \to \overline{Y}$**

$P_{16}(m) = 0$ si $\exists\, X \to Y$ / $m(\overline{X} \to \overline{Y}) \neq m(X \to Y)$
$P_{16}(m) = 1$ si $\forall\, X \to Y$ $m(\overline{X} \to \overline{Y}) = m(X \to Y)$

**<u>Propriété 17</u>** : **Mesure fondée sur un modèle probabiliste ou non**

$P_{17}(m) = 0$ (*taille fixe*) si $m$ n'est pas fondée sur un modèle probabiliste
$P_{17}(m) = 1$ (*taille aléatoire*) si $m$ est fondée sur un modèle probabiliste

**<u>Propriété 18</u>** : **Mesure descriptive ou statistique**

$P_{18}(m) = 0$ (*descriptive ou invariante*) si $\forall\, k \in \mathbb{N}^*$, $\forall\, X_1 \to Y_1$, $\forall\, X_2 \to Y_2$, ( $n_{X_1Y_1} = k\, n_{X_2Y_2}$ et $n_{X_1\overline{Y_1}} = k\, n_{X_2\overline{Y_2}}$ et $n_{\overline{X_1}Y_1} = k\, n_{\overline{X_2}Y_2}$ et $n_{\overline{X_1}\overline{Y_1}} = k\, n_{\overline{X_2}\overline{Y_2}}$ ) $\Rightarrow m(X_1 \to Y_1) = m(X_2 \to Y_2)$
$P_{18}(m) = 1$ (*statistique*) si $\exists\, k \in \mathbb{N}^*$, $\exists\, X_1 \to Y_1$, $\exists\, X_2 \to Y_2$ / $n_{X_1Y_1} = k\, n_{X_2Y_2}$ et $n_{X_1\overline{Y_1}} = k\, n_{X_2\overline{Y_2}}$ et $n_{\overline{X_1}Y_1} = k\, n_{\overline{X_2}Y_2}$ et $n_{\overline{X_1}\overline{Y_1}} = k\, n_{\overline{X_2}\overline{Y_2}}$ et $m(X_1 \to Y_1) \neq m(X_2 \to Y_2)$

**<u>Propriété 19</u>** : **Mesure discriminante**

$P_{19}(m) = 0$ (*non discriminante*) si $\exists\, \eta \in \mathbb{N}^*$ / $\forall\, n > \eta$ $\forall\, X_1 \to Y_1$ $\forall\, X_2 \to Y_2$
[ $P(Y_1/X_1) > P(Y_1)$ et $P(Y_2/X_2) > P(Y_2)$ ] $\Rightarrow m(X_1 \to Y_1) \approx m(X_2 \to Y_2)$
$P_{19}(m) = 1$ (*discriminante*) si $\forall\, \eta \in \mathbb{N}^*\, \exists\, n > \eta\, \exists\, X_1 \to Y_1\, \exists\, X_2 \to Y_2$ /
$P(Y_1/X_1) > P(Y_1)$ et $P(Y_2/X_2) > P(Y_2)$ et $m(X_1 \to Y_1) \neq m(X_2 \to Y_2)$

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

## Annexe 2 : Définitions des mesures

| N° | Mesure | Formule |
|---|---|---|
| 1 | coefficient corrélation ou ϕ-coefficient | $\frac{P(XY)-P(X)P(Y)}{\sqrt{P(X)P(Y)P(\overline{X})P(\overline{Y})}}$ |
| 2 | Cohen ou Kappa | $\frac{P(XY)+P(\overline{X}\,\overline{Y})-P(X)P(Y)-P(\overline{X})P(\overline{Y})}{1-P(X)P(Y)-P(\overline{X})P(\overline{Y})}$ $=2\times\frac{P(XY)-P(X)P(Y)}{P(X)+P(Y)-2\mathrm{P}(X)P(Y)}=2\times\frac{P(XY)-P(X)P(Y)}{P(X)P(\overline{Y})+P(\overline{X})P(Y)}$ |
| 3 | confiance ou précision | $P(Y/X)$ |
| 4 | confiance causale | $1-\frac{1}{2}P(\overline{Y}/X)-\frac{1}{2}P(X/\overline{Y})$ |
| 5 | confiance centrée ou valeur ajoutée ou Pavillon | $P(Y/X)-P(Y)=\frac{P(XY)-P(X)P(Y)}{P(X)}=P(\overline{Y})-P(\overline{Y}/X)$ |
| 6 | confiance confirmée descriptive ou Ganascia | $1-2\times P(\overline{Y}/X)=2\,confiance(X\rightarrow Y)-1=2\times\left(\frac{P(XY)}{P(X)}-0{,}5\right)$ |
| 7 | confiance confirmée causale | $1-\frac{3}{2}P(\overline{Y}/X)-\frac{1}{2}P(X/\overline{Y})=confCausale(X\rightarrow Y)-P(\overline{Y}/X)$ |
| 8 | confirmation causale | $P(X)+P(\overline{Y})-4\times P(X\,\overline{Y})$ |
| 9 | confirmation descriptive | $P(X)-2\times P(X\,\overline{Y})$ |
| 10 | conviction | $\frac{P(X)P(\overline{Y})}{P(X\,\overline{Y})}$ |
| 11 | cosinus ou Ochiai | $\frac{P(XY)}{\sqrt{P(X)P(Y)}}$ |
| 12 | couverture | $P(X)$ |
| 13 | Czekanowski-Dice ou F-mesure | $\frac{2\,P(XY)}{P(X)+P(Y)}=\frac{2\,P(XY)}{P(XY)+1-P(\overline{XY})}$ |
| 14 | dépendance | $\lvert P(\overline{Y})-P(\overline{Y}/X)\rvert$ |
| 15 | dépendance causale estimé | $\frac{3}{2}+2\,P(X)-\frac{3}{2}P(Y)-\frac{3}{2}P(\overline{Y}/X)-2\,P(X/\overline{Y})$ |

| | | |
|---|---|---|
| 16 | dépendance pondérée d'intérêt de Gray et Orlowska | $\left(\left(\frac{P(XY)}{P(X)P(Y)}\right)^k - 1\right) \times P(XY)^m$ |
| 17 | facteur bayésien ou multiplicateur de côte | $\frac{P(XY)P(\overline{Y})}{P(X\overline{Y})P(Y)}$ |
| 18 | facteur de certitude ou Loevinger ou satisfaction | $\frac{P(Y/X)-P(Y)}{1-P(Y)} = \frac{P(Y/X)-P(Y)}{P(\overline{Y})} = 1 - \frac{P(X\overline{Y})}{P(X)P(\overline{Y})} = \frac{P(X)P(\overline{Y})-P(X\overline{Y})}{P(X)P(\overline{Y})}$ |
| 19 | fiabilité négative | $P(\overline{X}/\overline{Y}) = confiance(\overline{Y} \rightarrow \overline{X})$ |
| 20 | force collective | $\frac{P(XY)+P(\overline{XY})}{P(X)P(Y)+P(\overline{X})P(\overline{Y})} \times \frac{1-P(X)P(Y)-P(\overline{X})P(\overline{Y})}{1-P(XY)-P(\overline{XY})}$ |
| 21 | Fukuda | $n(P(XY) - minConf \times P(X))$ |
| 22 | gain informationnel ou gain d'information | $\log_2\left(\frac{P(XY)}{P(X)P(Y)}\right)$ |
| 23 | Gini | $P(X)\left[P(Y/X)^2 + P(\overline{Y}/X)^2\right]$ $+P(\overline{X}) \times \left[P(Y/\overline{X})^2 + P(\overline{Y}/\overline{X})^2\right] - P(Y)^2 - P(\overline{Y})^2$ |
| 24 | Goodman-Kruskal | $\frac{\sum_j max_k P(X_j, Y_k) + \sum_k max_j P(X_j, Y_k) - max_j P(X_j) - max_k P(Y_k)}{2 - max_j P(X_j) - max_k P(Y_k)}$ $\frac{n_{XY}n_{\overline{XY}}/n^2 - n_{X\overline{Y}}n_{\overline{X}Y}/n^2}{n_{XY}n_{\overline{XY}}/n^2 + n_{X\overline{Y}}n_{\overline{X}Y}/n^2}$ |
| 25 | indice d'implication | $\sqrt{n}\frac{P(X\overline{Y})-P(X)P(\overline{Y})}{\sqrt{P(X)P(\overline{Y})}}$ |
| 26 | indice probabiliste d'écart à l'équilibre (*IPEE*) | $P\left[N(0,1) \geq \frac{n_{X\overline{Y}} - n_{XY}}{\sqrt{n_X}}\right]$ |
| 27 | indice probabiliste entropique d'écart à l'équilibre (*IP3E*) | $\sqrt{\frac{1}{2}\left[\left((1-h_1(P(X\overline{Y}))^2) \times (1-h_2(P(X\overline{Y}))^2)\right)^{1/4} + 1\right] \times IPEE}$ avec $h_1(t) = -\left(1-\frac{t}{P(X)}\right)\log_2\left(1-\frac{t}{P(X)}\right) - \frac{t}{P(X)}\log_2\left(\frac{t}{P(X)}\right)$ pour $t \in \left[0, \frac{P(X)}{2}\right[$, $h_1(t) = 1$ sinon $h_2(t) = -\left(1-\frac{t}{P(\overline{Y})}\right)\log_2\left(1-\frac{t}{P(\overline{Y})}\right) - \frac{t}{P(\overline{Y})}\log_2\left(\frac{t}{P(\overline{Y})}\right)$ pour $t \in \left[0, \frac{P(\overline{Y})}{2}\right[$, $h_2(t) = 1$ sinon |

| | | |
|---|---|---|
| 28 | indice probabiliste discriminant (*IPD*) | $P\left[N(0,1)>II^{CR/B}\right]$ où $II^{CR/B}$ indique que *II* est centré-réduit en fonction des valeurs prises par *II* sur l'ensemble des règles extraites |
| 29 | information mutuelle | $\frac{VS(XY)}{-P(X)\log_2 P(X)-P(\overline{X})\log_2 P(\overline{X})}$ |
| 30 | intensité d'implication (*II*) | $P[Poisson(nP(X)P(\overline{Y}))\geq P(X\overline{Y})]$ |
| 31 | intensité d'implication entropique (*IIE*) | $\sqrt{\left[(1-h_1(P(X\overline{Y}))^2)\times(1-h_2(P(X\overline{Y}))^2)\right]^{1/4}\times II}$ |
| 32 | intensité d'implication entropique révisée (*IIER*) | $\sqrt{\left[(1-h_1(P(X\overline{Y}))^2)\times(1-h_2(P(X\overline{Y}))^2)\right]^{1/4}\times max(2\times II-1\,;0)}$ |
| 33 | indice de vraisemblance du lien (*IVL*) | $P[Poisson(nP(X)P(Y))<P(XY)]$ |
| 34 | intérêt ou lift | $\frac{P(Y/X)}{P(Y)}=\frac{P(XY)}{P(X)P(Y)}$ |
| 35 | Jaccard | $\frac{P(XY)}{P(X)+P(Y)-P(XY)}=\frac{P(XY)}{P(X\overline{Y})+P(Y)}$ |
| 36 | J-mesure | $P(XY)\log_2\frac{P(XY)}{P(X)P(Y)}+P(X\overline{Y})\log_2\frac{P(X\overline{Y})}{P(X)P(\overline{Y})}$ |
| 37 | Klosgen | $\sqrt{P(XY)}\times\lvert P(Y/X)-P(Y)\rvert$ |
| 38 | Kulczynski ou indice d'accord et de désaccord | $\frac{P(XY)}{P(X\overline{Y})+P(\overline{X}Y)}$ |
| 39 | Laplace | $\frac{n_{XY}+1}{n_X+2}=\frac{n\times support(X\rightarrow Y)+1}{\frac{n\times support(X\rightarrow Y)}{confiance(X\rightarrow Y)}+2}$ |
| 40 | leverage | $P(Y/X)-P(X)P(Y)$ |
| 41 | $M_{GK}$ | Si $P(Y/X)\geq P(Y)$ alors $M_{GK}(X\rightarrow Y)=\frac{P(Y/X)-P(Y)}{1-P(Y)}$<br>sinon $M_{GK}(X\rightarrow Y)=\frac{P(Y/X)-P(Y)}{P(Y)}$ |
| 42 | moindre contradiction ou surprise | $\frac{P(XY)-P(X\overline{Y})}{P(Y)}$ |
| 43 | nouveauté | $P(XY)-P(X)P(Y)$ |
| 44 | Pearl | $P(X)\lvert P(Y/X)-P(Y)\rvert$ |

| | | |
|---|---|---|
| 45 | Piatetsky-Shapiro | $n \times P(X)[P(Y/X) - P(Y)] = n\lvert P(XY) - P(X)P(Y)\rvert$ |
| 46 | précision ou support causal | $P(XY) + P(\overline{XY}) = P(X) + P(\overline{Y}) - 2 \times P(X\overline{Y})$ |
| 47 | prévalence | $P(Y)$ |
| 48 | Q de Yule | $\frac{P(XY)P(\overline{XY}) - P(X\overline{Y})P(\overline{X}Y)}{P(XY)P(\overline{XY}) + P(X\overline{Y})P(\overline{X}Y)}$ |
| 49 | rappel | $P(X/Y)$ |
| 50 | ratio des chances | $\frac{P(XY)P(\overline{XY})}{P(X\overline{Y})P(\overline{X}Y)}$ |
| 51 | risque relatif | $\frac{P(Y/X)}{P(Y/\overline{X})}$ |
| 52 | Sebag-Schoenauer | $\frac{P(XY)}{P(X\overline{Y})} = \frac{1}{\frac{1}{confiance(X \to Y)} - 1}$ |
| 53 | spécificité | $P(\overline{Y}/\overline{X})$ |
| 54 | support ou indice de Russel et Rao | $P(XY)$ |
| 55 | support à sens unique de Yao et Liu (*SU*) | $P(Y/X) \times \log_2 \frac{P(XY)}{P(X)P(Y)}$ |
| 56 | support à double sens de Yao et Liu (*SD*) | $P(XY) \times \log_2 \frac{P(XY)}{P(X)P(Y)}$ |
| 57 | taux d'exemples et de contre-exemples | $1 - \frac{P(X\overline{Y})}{P(XY)}$ |
| 58 | valeur test *VT100* | $\phi^{-1}(P[Hypergéométrique(100\,P(X)P(Y)) \leq P(XY)])$ |
| 59 | variation du support à double sens de Yao et Liu (*VS*) | $P(XY) \times \log_2 \frac{P(XY)}{P(X)P(Y)} + P(X\overline{Y}) \times \log_2 \frac{P(X\overline{Y})}{P(X)P(\overline{Y})}$ $+ P(\overline{X}Y) \times \log_2 \frac{P(\overline{X}Y)}{P(\overline{X})P(Y)} + P(\overline{X}\,\overline{Y}) \times \log_2 \frac{P(\overline{X}\,\overline{Y})}{P(\overline{X})P(\overline{Y})}$ |
| 60 | Y de Yule | $\frac{\sqrt{P(XY)P(\overline{X}\,\overline{Y})} - \sqrt{P(X\overline{Y})P(\overline{X}Y)}}{\sqrt{P(XY)P(\overline{X}\,\overline{Y})} + \sqrt{P(X\overline{Y})P(\overline{X}Y)}}$ |
| 61 | Zhang | $\frac{P(XY) - P(X)P(Y)}{max(P(XY)P(\overline{Y}); P(Y)P(X\overline{Y}))}$ |

*Tab 6 : Définitions des 61 mesures.*

# 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 (2009). A new generic basis of "factual" and "implicative" association rules. *Intelligent Data Analysis journal.* 13(4): pp. 633-656.

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.

Blanchard J., Guillet F., Briand H. and R. Gras (2005a). IPEE : Indice Probabiliste d'Ecart à l'Equilibre pour l'évaluation de la qualité des règles. Dans l'Atelier Qualité des Données et des Connaissances, pp. 26–34.

Blanchard J., Guillet F., Briand H. and R. Gras (2005b). Une version discriminante de l'indice probabiliste d'écart à l'équilibre pour mesurer la qualité des règles. In *Troisièmes rencontres internationales de l'Analyse Statistique Implicative* (ASI 05).

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 .

Czekanowski J. (1913). *Zarys metod statystycznych (Die Grundzuge der statischen Metoden)*, Warsaw.

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.

Guillaume, S. (2000), *Traitement des données volumineuses : mesures et algorithmes d'extraction de règles d'association et règles ordinales*, PhD thesis, Université de Nantes.

Gras R. (1979). *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques*, Thèse d'Etat, Université Rennes 1.

Gras R., Kuntz P., Couturier R. and F. Guillet (2001). Une version entropique de l'intensité d'implication pour les corpus volumineux. *Extraction des connaissances et apprentissage (Extraction et Gestion des Connaissances 2001)*, 1(1-2) pp. 69–80.

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.

Kodratoff, Y. (2001). Comparing machine learning and knowledge discovery in databases: An application to knowledge discovery in textx, *Machine Learning and Its Applications*, Advanced Lectures LNCS 2049, pp. 1-21.

Kulczynski, S. (1928). Die Pflanzenassoziationen der Pieninen, *Bull. Int. Acad. Pol. Sci. Lett. Cl. Sci. Math. Nat*., Ser. B, Suppl. II (1927), pp. 57-203.

Jaccard P. (1908), Nouvelles recherches sur la distribution florale. *Bulletin Society Vaud Science National*, (44) pp. 223–270.

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, Cépaduès, pp. 193-217.

Lallich, S., Vaillant, B., and P. Lenca (2005). Parametrised measures for the evaluation of association rule interestingness. In *The XIth International Symposium on Applied Stochastic Models and Data Analysis*, Brest, France, pp. 220–229.

Lavrac N., Flach P., and B. Zupan (1999). Rule evaluation measures : A unifying view. In G. Mineau and B. Ganter, editors, *Ninth International workshop on Inductive Logic Programming*, volume 1634, pp. 174–185.

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.

Lerman, I.-C. (1970). Sur l'analyse des données préalable à une classification automatique (proposition d'une nouvelle mesure de similarité) . In *Mathématiques et sciences humaines*, tome 32, pp. 5-15.

Lerman I.C. (1981). Classification et analyse ordinale des données, Dunod.

Lerman, I.-C. et J. Azé (2007). A new probabilistic measure of interestingness for association rules, based on the likelihood of the link. In Quality measures in data mining 2007, Volume 43 of Studies in Computational Intelligence, pp. 207–236. Springer.

Lesot, M.-J. et M. Rifqi (2010). Order-based equivalence degrees for similarity and distance measures. In Proceedings of *IPMU (International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems)*, pp. 19-28, Springer LNAI 6178.

Ochiai, A. (1957). Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions, *Bull. Jpn. Soc. Sci. Fish*., 22, pp. 526-530.

Piatetsky-Shapiro, G. (1991). Discovery, Analysis and Presentation of Strong Rules, In G. Piatetsky-Shapiro & W.J. Frawley, editors, *Knowledge Discovery in Databases*, AAAI Press, pp. 229-248.

Russell P. F. et T. R. Rao (1940). On habitat and association of species of anopheline larvae in south-eastern Madras, *J. Malar. Inst. India*, 3, pp. 153-178.

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.

Yule, G. U. (1900). On the association of attributes in statistics. In *Philosophical Transactions of the Royal Society of London*, Londra.

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.

Zhang, T. (2000). Association rules. In T. Terano, H. Liu, A.L.P. Chen (Eds), *Actes Conférence PAKDD 2000*, LNAI 1805, pp. 245-256, Springer-Verlag.

Zighed, D., Abdesselam R., et A. Bounekkar (2011). Equivalence topologique entre mesures de proximité. *Actes EGC'2011 Extraction et Gestion des Connaissances*, RNTI E.20, Hermann ISBN 978 27056 8112 8, pp.53-62.

# 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 rule extraction measures, support and confidence. Although these two measures have the virtue of being algorithmically fast, they generate a prohibitive number of rules most of which are redundant and irrelevant. It is therefore necessary to use further measures which filter uninteresting rules. Different reported studies have been carried out to identify "*good* " properties of rule extraction measures and these properties have been assessed on *61* measures. The aim of this paper is to identify categories of measures which can answer a concern raised by users: choosing one or more measure 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 has enabled us to identify *7* classes of measures, classes that we obtained using two techniques: AHC according to the Ward criterion and the clustering k-means method.