Towards a theory unifying implicative interestingness measures and critical values consideration in M_{GK}.

Vers une théorie unificatrice des mesures implicatives d'intérêt et considération des valeurs critiques sur M_{GK} .

ANDRE TOTOHASINA[1]

Abstract

The present paper shows the possibility and the benefit to compute statistical freshold for the so-called Guillaume-Kenchaff interestingness measure M_{GK} of association rule and compares it with other measures as Confidence, Lift and Lovinger's one. Afterwards, it proposes a theory of normalized interestingness measure unifying a set of rule quality measures in a binary context and being surprisingly centered on M_{GK}.

Keywords : *Association rule, Binary context, Statistical implication, Unifying theory, Critical values, MGK.*

Resume

Le présent papier montre la possibilité et l'avantage de calculer les valeurs statistiques critiques de ladite mesure d'intérêt d'une règle d'association M_{GK} de Guillaume-Kenchaff, effectue une étude comparative de cette dernière avec d'autres mesures de la qualité telles Confiance, Lift et celle de Lovinger. Ensuite, il propose une théorie de mesure normalisée qui unifie un ensemble des mesures de qualité des règles dans un contexte binaire et qui a une propriété d'être centrée sur M_{GK}.

Mots-clés : *Règle d'association, Contexte binaire, Implication statistique, Théorie unificatrice, Valeurs critiques, M_{GK}.*

Introduction

Association rules reveal attributes occurring together frequently in a database, their relevance being commonly assessed by means of interestingness measures. In addition of the standard marketing problem, mining association rules has many application areas like environmental science in extracting spatial patterns from image databases or geo-referenced census data, mathematic education, taxonomy problems, fraud detection, sociology, psychology, epidemiology, medical diagnosis (Alonso et al., 2002), etc. Several interestingness measures have been proposed in the literature (Hilderman, 1999), the most popular of them being the well-known *Support* , *Confidence* , *Lift* , *Conviction* , *Lovinger* . A major problem faced in association rule extracting is the huge number of valid rules, *i.e.*, rules meeting specific

[1] ENSET. University of Antsiranana, 201-Madagascar, totohasina@yahoo.fr

constraints relative to given interestingness measures. Such a situation is generally due to the presence of many redundant and / or trivial rules in the set of valid ones, and, maybe, because of arbitrary threshold adoption.

Put in the topic of the knowledge discovery, information retrieval and statistical implication analysis, the present paper shows the possibility and the benefit to compute statistical freshold for the so-called Guillaume-Kenchaff interestingness measure M_{GK} of association rule and talks about its unifying properties.

Moreover, within a comparing analysis with the famous interestingness measure *Condidence* and others, as mathematical and statistical properties, we explain its intelligibility. It allows comparison between M_{GK} and the traditional measure *Confidence* about pertinence of produced rules. We shall talk about an application on a real data.

Motivations and mathematical modelling

Let us recall that this interestingness measure M_{GK} has been independently proposed by (Guillaume, 2000) during the year 2000 inspired by Loevinger's index and by (Wu et al., 2004) in 2004. Through its mathematical properties, this quality measure receives different names as ION by (Totohasina, 2004, 2005) showing its implicative oriented normalized property (Brin & al.,1997), $CPIR$ by (Wu et al, 2004) because of expressing a Conditional Probability Increment Ratio and of its efficiency to extract non redundant association rules, also $Conf_G$ (Guillaume's Confidence) by S. Ferré (cf. p.139-140 in (Ferré, 2002) showing that it is both more precise and more understandable, of course it appears more convenient with contextualized analysis of logical information system than the standard Agrawal et al.'s $Confidence$ (Agrawal et al.,1993) ($Confidence$ can not distinguish attraction and repulsion), CF(Certainty Factor) in (Sanchez et al., 2008).

Moreover, as shown in the litterature, the association rules extraction techniques make sufficient to dealing with somewhat arbitrary and subjective constraint as $minimalsupport$ for potential itemsets, maybe excepting Gras'statistical implication method in integrating objective threshold for the *intensity of implication* ((Gras & al., 1996), p.42-46). In the present paper, we propose an objective possibility and advantage in integrating critical values for statistical index M_{GK}. As it is still rare works on this interesting association rule quality

measure, its advancing analysis is necessary highlithing. We will show through it has unifying property for a lot of association rule interestingness measures and it allows to build infinity of normalized quality measures.

Among many probabilistic mathematical modelling seen in the litterature (see for example (Lerman, 1984)), here, we consider the context of binary data mining $K = (O, A, R)$, where O is a finite set of entities or objects, A is a non empty finite set of attributes and R a binary relation from O to A. A couple $(o, a) \in O \times A$ in the graph of the relation R means that the object o posses the property a, all attribut beeing identifyed as a function from O to $\{0,1\}$, where the value 1 measures the presence of the attribut in an object of O. Let us write n the cardinality of O ($n = |O|$). All subset X of A is called an *itemset* of A, and its logical negation $\overline{X}$ is the negative itemset of A. Any subset X in A is called an *itemset* of A, and its logical negation written as $\overline{X}$ the *negative itemset of the itemset* X, and any element of O an *object* or an *entity* of O. For all itemset X in A, let us remark the eight following points: $\forall a \in A, \overline{a}$, i.e. $(1-a)$ identifies the absence of the attribut a at an entity ; $\forall e \in O, X(e) = 1 \Leftrightarrow \forall a \in X, eRa, i.e. a(e) = 1$; say $X = \wedge_{a \in X} a$ = the conjunction of presences of a finite number of attributs of X ; $X' = \{e \in O \mid \forall x \in X, eRx\}$, *i.e.* the dual or the *extension* of X ; dually, for any subset E of entities in O, the itemset contained in E, say the *intension* of E, symbolized as E', is defined as $E' = \{a \in A / \forall e \in E, a(e) = 1\}$, say the set of common attributs to the objects belonging to E ; $\overline{X}' = O - X' = \overline{X'}$; this coïncidence explains the calling of negative itemset for $\overline{X}$; $\overline{\overline{X}} = X$: so we find again the involutive property of the negation in formal logic, *i.e.* the De Morgan law; $X \subseteq A$, but $\overline{X} \acute{U} A$. It is easy to see that for two itemsets X and Y in the context, one has: $(X \cap Y)' = X' \cap Y'$ et $(X \cup Y)' = X' \cup Y'$. An association rule of $K = (O, A, R)$ is an ordered pair (X, Y) of itemsets wich are both positive or negative, or alternatively negative and positive, denoted $X \to Y$ and read as "If X, then Y ", where $Y \cap X$ is required to be empty: the itemsets X and Y are respectively called the "*Premice*" and the "*Consequent*" of the association rule $X \to Y$. Since a priori one is not right to refuse it, we naturally consider an hypothesis of equiprobability of atomic events of O. Hence we presently

consider the discrete probabilized space $(O, P(O), P)$, P beeing the intuitive uniform probability. Consequently, for all X in $P(A)$, writing $n_X = |X'|$ the cardinality of X', $Supp(X) = \frac{n_X}{n}$ represents an estimation of the probability $P(X')$ of the event X' that X would be contained in n_X entities. Moreover, as justified by the duality between *extension* and *intension* , it appears natural to adopt the following definitions: two itemsets are said to be independent (resp. dependent), if their respective extensions are independent (resp. dependent) in the probabilized space $(O, P(O), P)$.

Between the two measures M_{GK} and *Confidence*.

According to the present probabilistic modelling, the following elementary properties allow us to easily build the so called quality measure M_{GK}.

Remark 1 It is obvious that, for any itemsets X and Y, one has the following double inequalities:

- If X favors Y , (i.e. $P(Y' \mid X') > P(Y')$) , then $0 < P(Y' \mid X') - P(Y') \leq 1 - P(Y')$.
- If X disfavors Y , (i.e. $P(Y' \mid X') < P(Y')$) then $-P(Y') \leq P(Y' \mid X') - P(Y') < 0$.
- " X disfavors Y is equivalent to" X favors $\overline{Y}$ "; thus $1 - P(Y') < 1 - P(Y' \mid X')$

if and only if $P(\overline{Y'}) < P(\overline{Y'} \mid X')$.

Hence, one puts the following definition.

Definition 1 Let X and Y be two itemsets in a data mining context. One defines:

$$M_{GK}(X \to Y) = \begin{cases} \dfrac{P(Y' \mid X') - P(Y')}{1 - P(Y')}, & \text{if } X \text{ favors } Y \\ \dfrac{P(Y' \mid X') - P(Y')}{p(Y')}, & \text{if } X \text{ disfavors } Y. \end{cases} \quad (1)$$

So $M_{GK}(X \to Y) = M_{GK}{}^{f}(X \to Y) \times 1_f(X \to Y) + M_{GK}{}^{d}(X \to Y) \times 1_d(X \to Y)$,

where 1_f represents the indicator of the event "*Premice favors Consequent*", 1_d the indicator of the event "*Premice disfavors Consequent*".

The expression of M_{GK} depending on *Confidence* is given by:

$$\mathrm{M}_{\mathrm{GK}}(X \to Y) = \frac{\mathrm{conf}(X \to Y) - \mathrm{supp}(Y)}{1 - \mathrm{supp}(Y)} \times 1_f(X \to Y) + \frac{\mathrm{conf}(X \to Y) - \mathrm{supp}(Y)}{\mathrm{supp}(Y)} \times 1_d(X \to Y).$$

Thus M_{GK} is composed of the favoring component $\mathrm{M}_{\mathrm{GK}}{}^{f}$ and of the disfavoring one $\mathrm{M}_{\mathrm{GK}}{}^{d}$. But for two non independent itemsets X and Y, we have one of the two alternatives: there is mutual attraction, then we talk about a positive dependence, or about negative dependence in case of repulsion between the two itemsets X and $\overline{Y}$; thus we consider $X \to \overline{Y}$ in the first hand, and between $\overline{X}$ and Y we consider $\overline{X} \to Y$ in the other hand. In the both cases, we always will consider a positive dependence. As it is obvious, the favoring component $\mathrm{M}_{\mathrm{GK}}{}^{f}$ guides the semantic of M_{GK}. Let us remark that $\mathrm{M}_{\mathrm{GK}}{}^{f}$ is the only component of M_{GK} coinciding with the Lovinger's measure, but $\mathrm{M}_{\mathrm{GK}}{}^{d} = Lift - 1$

Proposition 1

– If X favors Y, the we obtain the equivalence relation of two counteropposite rules:

$$\mathrm{M}_{\mathrm{GK}}{}^{f}(\overline{Y} \to \overline{X}) = \mathrm{M}_{\mathrm{GK}}{}^{f}(X \to Y). \quad (2)$$

– If X disfavors Y, then we have the relation:

$$\mathrm{M}_{\mathrm{GK}}{}^{d}(\overline{Y} \to \overline{X}) = a(X \to Y)\mathrm{M}_{\mathrm{GK}}{}^{d}(X \to Y) \quad (3)$$

where $a(X \to Y) = \dfrac{P(X')P(Y')}{(1 - P(X'))(1 - P(Y'))}$.

Thus, according to the above remark, on can consider that M_{GK} is favorly implicative, unlike confidence is not implicative. As illustration, the five following tables (see table 1: (1) & (2), table 2: (3), table 3: (4) & (5)), highlith the evaluating modes of dependency degree between two itemsets, in five references of situation: positive dependence, negative dependence, independence, incompatibility, et the logical implication. Unlikely a χ^2, it appears that the measure M_{GK} computes the strength of the oriented dependence on the bounded intervall $[-1,+1]$. For instance, in the table 1(2), the very significant dependence between the two itemsets as revealed by χ^2 is in fact a negative dependence and since $\mathrm{M}_{\mathrm{GK}}{}^{f}(X \to \overline{Y}) = 0{,}1$

, thus negligible when compared with $M_{GK}{}^{f}(\overline{X} \to Y) = 4/9 = 0{,}444{,}1$: so, only the left-hand negative rule $\overline{X} \to Y$ is significantly valid. Let us notice that here the $\mathrm{conf}(\overline{X} \to Y) = 0{,}75$ is sufficently high too, but its M_{GK}- value is relatively significantless (thus, the saturation ratio $M_{GK} = \frac{1 - M_{GK}}{1 - 0} = \frac{5}{9}$;

$0.556 > (\frac{1 - \mathrm{conf}}{1 - 0} = \frac{1}{4} = 0.25))$. Really, it is immediate that in case of positive dependence partialy implicative, one has $0 < \frac{M_{GK}{}^{f}}{\mathrm{conf}} < 1$. Thus M_{GK} is more discriminant than the standard *confidence*.

Table 1. Case of Positive dependence and wake M_{GK}-value against Negative dependence and negative heavy M_{GK}-value.

	Y	$\overline{Y}$	(1)
X	3000	2000	5000
$\overline{X}$	2500	2500	5000
(1)	5500	4500	10000

	Y	$\overline{Y}$	(2)
X	1000	3000	4000
$\overline{X}$	4500	1500	6000
(2)	5500	4500	10000

(1) Positive dependence with $\chi^2 = 101$ & $M_{GK} = +0.11$.

(2) (Negative dependence with $\chi^2 = 2424$ & $M_{GK} = -0.54$

Table 2. Case of Independence.

	Y	$\overline{Y}$	(3)
X	2200	1800	4000
$\overline{X}$	3300	2700	6000
(3)	5500	4500	10000

(3) Independence with $\chi^2 = 0$ & $M_{GK} = 0$

Table 3. Case of Incompatibility and M_{GK}-value=-1 against Logical implication and M_{GK}-value = 1.

	Y	$\overline{Y}$	(4)
X	0	2000	2000
$\overline{X}$	6000	2000	8000
(4)	6000	4000	10000

	Y	$\overline{Y}$	(5)
X	3000	0	3000
$\overline{X}$	3000	4000	7000
(5)	6000	4000	10000

(4) Incompatibility with $\chi^2 = 3750$ & $M_{GK} = -1$

(5) logical Implication with $\chi^2 = 2857$ & $M_{GK} = +1$

The proposition 2 below proves that the interestingness measure M_{GK} is favorly non symetric.

Proposition 2

– If X favors Y, then one has the relation:

$$\mathrm{M}_{\mathrm{GK}}{}^{f}(Y \to X) = \frac{1-P(Y')}{1-P(X')}\frac{P(X')}{P(Y')}\mathrm{M}_{\mathrm{GK}}{}^{f}(X \to Y). \quad (4)$$

– If X disfavors Y, then one has the relation:

$$\mathrm{M}_{\mathrm{GK}}{}^{d}(Y \to X) = \mathrm{M}_{\mathrm{GK}}{}^{d}(X \to Y) \quad (5)$$

Concerning the right-hand side negative rules, we have:

Proposition 3 *For any two positive items X and Y, one has the equality and the equivalence below:*

$$\begin{aligned} &\mathrm{M}_{\mathrm{GK}}{}^{f}(X \to \overline{Y}) = -\mathrm{M}_{\mathrm{GK}}{}^{d}(X \to Y). Et\ \forall \alpha \in]0,1[, \\ &on\ a : (-1 < \mathrm{M}_{\mathrm{GK}}{}^{d}(X \to Y) < -\alpha \Leftrightarrow \alpha < \mathrm{M}_{\mathrm{GK}}{}^{f}(X \to \overline{Y}) < 1) \end{aligned} \quad (6)$$

Thus, the more the degree of quasi-incompatibility between the two itemsets is high, the more the quality of the correspondent negative rule is favorly the best; this equivalence allows to prun directly all right-hand side negative rule candidate whose the MGK-value is negative and located in the interval[-1, 0] for a fixed freshold in [0, 1]. About the left-handside negative rule, one has the proposition below.

Proposition 4 : *For any two itemsets X and Y, one has the following inequalities:*

[(1)] *If* X *disfavors* Y, *then* : $\mathrm{M}_{\mathrm{GK}}{}^{f}(\overline{X} \to Y) = \lambda_1(X \to Y)\mathrm{M}_{\mathrm{GK}}{}^{f}(X \to \overline{Y})$ S

[(2)] If X favors Y (*i.e.*, X disfavors $\overline{Y}$ and also $\overline{X}$ disfavors Y), then:

$$\mathrm{M}_{\mathrm{GK}}{}^{d}(\overline{X} \to Y) = \lambda_2(X \to Y)\mathrm{M}_{\mathrm{GK}}{}^{d}(X \to \overline{Y}) \quad (7)$$

[(3)] For all itemsets $X_1, X_2, ..., X_i, X_{i+1}, ..., X_p$ such that $X_1 \subseteq X_2 \subseteq ... \subseteq X_i \subseteq X_{i+1} \subseteq ... \subseteq X_p$.

– If $X_1 \to X_p$ est $(\mathrm{M}_{\mathrm{GK}}, \alpha)$-valid then $\forall i, j \in \{1,...,p\}$ with $i < j$, $X_i \to X_j$ est $(\mathrm{M}_{\mathrm{GK}}, \alpha)$-valid.

– If it exists $i, j \in \{1,...,p\}$ such that $X_i \to X_j$ is non $(\mathrm{M}_{\mathrm{GK}}, \alpha)$-valid then $\forall l, k \in \{1,...,p\}$ such that $l \le i$ et $j \le k$, $X_l \to X_k$ is also non $(\mathrm{M}_{\mathrm{GK}}, \alpha)$-valid.

where: $\lambda_1(X \to Y) = \dfrac{P(X')}{1-P(X')} \dfrac{P(Y')}{1-P(Y')}$, and

$\lambda_2(X \to Y) = \dfrac{P(X')}{(1-P(X'))} \dfrac{1-P(Y')}{P(Y')}$.

From the two precedent propositions 3 and 4, one deduces the relation between a left-hand side negative rule and the right-hand side positive corresponding one.

Except the above mentioned five references situations, (Blanchard & al., 2005) consider an other reference situation, that is the balancing position or maximum uncertainty position (*i.e.*, $|X' \cap Y'| = |X' \cap \overline{Y'}|$) : A quality measure is said "measuring *equilibrium deviation*" if it takes a constante value in case of equality between the number of examples and the number of counter-examples of the rule (Blanchard et al.,2005). Since at the equilibrium position, one has asymptoticaly: $\mathrm{M}_{\mathrm{GK}}{}^{f}(X \to Y) \approx \dfrac{1}{2}$ (Diatta & al., 2007). Let $\mathrm{M}_{\mathrm{GK}}^{\mathrm{cr}}$ be the freshold of $\mathrm{M}_{\mathrm{GK}}{}^{f}$: in a favoring case, a rule $X \to Y$ is valid for the fixed freshold α, i.e. $(\mathrm{M}_{\mathrm{GK}}, \alpha)$-valid, if $\mathrm{M}_{\mathrm{GK}}{}^{f}(X \to Y) > \mathrm{M}_{\mathrm{GK}}^{\mathrm{cr}}$; these critical values are computed from χ^2's freshold read at the same fixed freshold α .

Proposition 4 *The significance of* M_{GK} *depends on three integer parameters, say the size* n *of the sample, the occurences* n_X *and* n_Y *respectively of the itemsets* X *and* Y *.*

If $0 < n_X \le n_Y$ and X favors Y, then

$$\chi^2 > \chi^2_{cr} \Leftrightarrow \mathrm{M}_{\mathrm{GK}}{}^{f}(X \to Y) > \sqrt{\frac{n_{\overline{X}}.n_Y}{n_X.n_Y}\chi^2_{cr}} \qquad (8)$$

where χ^2_{cr} is critical value obtained at a fixed freshold χ^2 of independence of one degree of freedom.

From the usual relation $\chi^2 = n.\rho^2(X,Y)$, one deduces: $\mathrm{M}_{\mathrm{GK}}{}^{f}(X \to Y) = \sqrt{\dfrac{n_{\overline{X}}.n_Y}{n.n_X.n_{\overline{Y}}}\chi^2}$.

This last equality has the advantage to give us the critical values $\mathrm{M}_{\mathrm{GK}}^{\mathrm{cr}}$ of M_{GK}, via the critical

values χ^2_{cr} of the statistic Khi-square of 1 degree of freedom, without normality condition: the critical values of M_{GK} are obtained by replacing $\mathrm{M}_{\mathrm{GK}}{}^{f}(X \to Y)$ by

$$\mathrm{M}^{\mathrm{cr}}_{\mathrm{GK}}(X \to Y) = \sqrt{\frac{n_{\overline{X}}.n_Y}{n.n_X.n_{\overline{Y}}}\chi^2_{cr}} \tag{9}$$

For any itemsets that are fiting together, let us remark that M_{GK} as statistic is writable under the form:

$$\mathrm{M}_{\mathrm{GK}}{}^{f}(X \to Y) = \frac{n\dfrac{n_{XY}}{n_X} - n_Y}{n - n_Y}, \tag{10}$$

The algorithm below gives the computation of the critical values of M_{GK}.

Algorithm 1 (Gen-Rules)

Entrance : l_k, H_m

Exit: R set of association rules [1] $k > m+1$

$H_m \leftarrow Apriori - Gen(H_m)$

$h_{m+1} \in H_{m+1}$

$$\mathrm{M}^{\mathrm{cr}}_{\mathrm{GK}} \leftarrow \sqrt{\frac{(n - \mathrm{supp}(l_k - h_{m+1}))\mathrm{supp}(h_{m+1})}{n * \mathrm{supp}(l_k - h_{m+1})(n - \mathrm{supp}(h_{m+1}))}\chi^2_{cr}}$$

$$\mathrm{M}_{\mathrm{GK}} \leftarrow \frac{n * \mathrm{supp}(l_k) - \mathrm{supp}(l_k - h_{m+1})\mathrm{supp}(h_{m+1})}{\mathrm{supp}(l_k - h_{m+1})(n - \mathrm{supp}(h_{m+1}))}$$

$\mathrm{M}_{\mathrm{GK}} \geq \mathrm{M}^{\mathrm{cr}}_{\mathrm{GK}}$

$\mathsf{R} \leftarrow \mathsf{R} \cup \{r : l_k - h_{m+1} \to h_{m+1}\}$

$H_{m+1} \leftarrow H_{m+1} - \{h_{m+1}\}$

$Gen - rules(l_k, H_{m+1})$

Return R

Proposition 5 *Let* $X_1, X_2, ..., X_i, X_{i+1}, ..., X_p$ *be itemsets such that* $X_1 \subseteq X_2 \subseteq ... \subseteq X_i \subseteq X_{i+1} \subseteq ... \subseteq X_p$.

– if $X_1 \to X_p$ est $(\mathrm{M}_{\mathrm{GK}}, \alpha)$-valid then $\forall i, j \in \{1, ..., p\}$ avec $i < j$,

$X_i \rightarrow X_j$ is $(\mathrm{M_{GK}}, \alpha)$-valid.

– If it exists $i, j \in \{1,...,p\}$ such that $X_i \rightarrow X_j$ is not $(\mathrm{M_{GK}}, \alpha)$-valid then $\forall l, k \in \{1,...,p\}$ such that $l \leq i$ et $j \leq k$, $X_l \rightarrow X_k$ is also not $(\mathrm{M_{GK}}, \alpha)$-valid.

As an example, for a binary context of 10 objects and 9 items, below we can see the critical values of $\mathrm{M_{GK}}$ at 2,5 % threshold (see Table 4 and Table 5).

Table 4. Critical values of $\mathrm{M_{GK}}$ at 0.025 threshold for a context of 10 objects and 9 items

n_X \ n_Y	1	2	3	4	5
	0.70879				
	0.47252	0.70879			
	0.36090	0.54135	0.70879		
	0.28936	0.43404	0.56829	0.70879	
	0.23626	0.35439	0.46401	0.57872	0.70879
	0.19290	0.28936	0.37886	0.47252	0.57872
	0.15467	0.23200	0.30376	0.37886	0.46401
	0.11813	0.17719	0.23200	0.28936	0.35439
	0.07875	0.11813	0.15467	0.19290	0.23626

Table 5. Critical values of $\mathrm{M_{GK}}$ at 0.025 threshold for a context of 10 objects and 9 items

n_X \ n_Y	6	7	8	9
	0.70879			
	0.56829	0.70879		
	0.43404	0.54135	0.70879	
	0.28936	0.36090	0.47252	0.70879

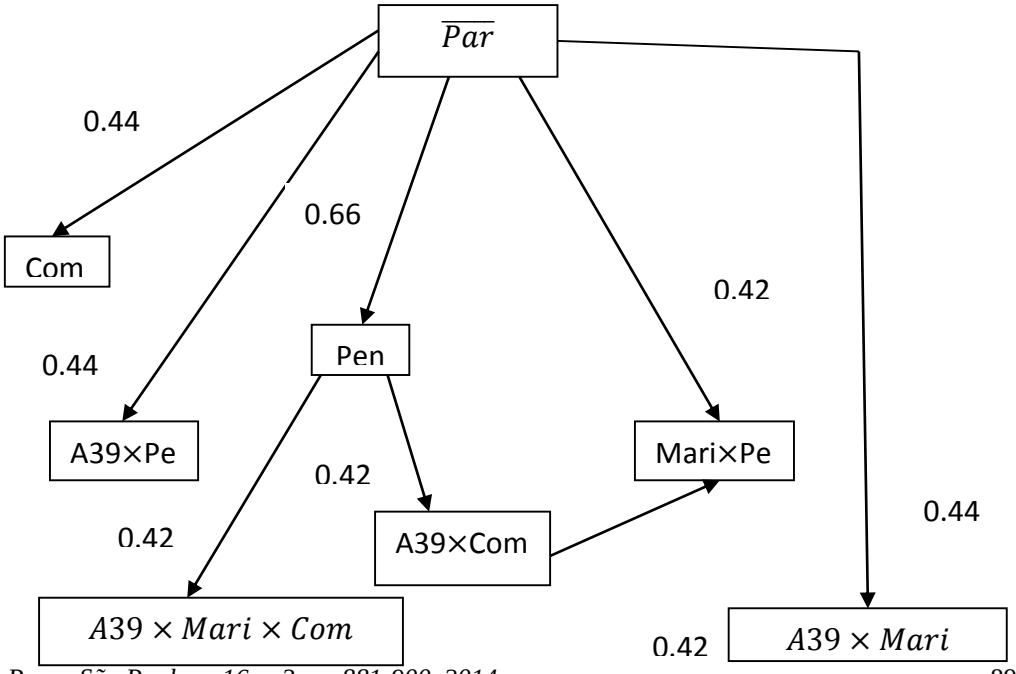
As an illustration, let us consider the Table 6 below taken from (Guillaume, 2000) presenting a database of bank about its customers'behavior : it is about 10 clients observed on four variables extended in 9 binary modalities, say : âge in the two classes]20 ; 29],]29 ; 39], and modalities of matrimonial situation, say : Married, Occupation and Category.

Table 6. Bank data

Variables	Age	Maried	Occupation	Category
Entités				
e_1	24	yes	artist	bad
e_2	23	no	guide	medium
e_3	32	yes	teaching	medium
e_4	35	yes	artist	good
e_5	39	yes	teaching	bon
e_6	31	yes	artist	good
e_7	29	yes	teaching	good
e_8	30	yes	teaching	medium
e_9	38	yes	teaching	good
e_{10}	36	yes	artist	bad

The corresponding result is expressed under the form of valued implicative graph (or valued directed graph) (see Figure 1): it interprets a part of the knowledge contained in the data, where modalities are represented like following : A29: âge $\in$]20;29], A39: âge $\in$]29;39], Mari: married, Par: profession artist, Pgu: profession guide, Pte: profession teaching, Cba: category bad, Cme: category medium, Cgo: category good.

Figure 1. Illustration on implicative graph of 2,5% threshold.



Interpretation of the implicative graph: The majority of the regular customers of the bank that are not artist are significantly teachers of the medium category, married and 39 years old.

Extension and unifying view

Normalized rule interestingness measure

Definition 2 *An interestingness measure* μ *is said to be normalized if it satisfies the five following conditions, say for any association rule* $X \to Y$, *one has: (i)* $\mu(X \to Y) = -1$, *if* $P(Y'/X') = 0$ *; (ii)* $-1 < \mu(X \to Y) < 0$, *if* $0 \neq P(Y'/X') < P(Y')$ *(i.e.* X *and* Y *are negatively dependent (or partially repulsive) ; (iii)* $\mu(X \to Y) = 0$, *if* $P(Y'/X') = P(Y')$ *(i.e.* X *and* Y *are independent) ; (iv)* $0 < \mu(X \to Y) < 1$, *if* $1 \neq P(Y'/X') > P(Y')$, *i.e.if* X *favors* Y, *or* X *and* Y *attract one another partially*; *(v)* $\mu(X \to Y) = 1$, *if* $P(Y'/X') = 1$ *(either if* X *totally implies* Y.

Thus, a normalized interestingness measure has the semantic of an oriented link, which can be interpreted as taxonomy, that is implicite in a sylogysm as "if X, then Y ". It is a quasi-implication index. Let us notice C(N) the set of such continued normalized probabilistic quality measures of association rule, that is continued function of the number of counter-examples (or examples) of the rule.

Remark 2 We have added two other conditions to the three Piatetsky-Shapiro's conditions (See (Piatetsky-Shapiro, 1991), (Hilderman & al., 1999), (Freitaas & al., 1999)), say the value of -1 in case of incompatibility which is considered as the limit of negative dependence and the value of $+1$ in case of logical implication which is considered as the limit of positive dependence. These two extreme values of all normalized probabilistic interestingness measure of association rule allows comparison of strentgh of rules. For instance, an association rule whose normalized quality measure is near $+1$ (resp. -1) indicates that there is strong attraction (resp. repulsion) between premise and consequent.

Between the standard measure *Confidence* and M_{GK}, one has the proposition below.

Proposition 6

• M_{GK} is normalized, but *confidence* is not ; at fixed margins, of course *Confidence* and $\mathrm{M}_{\mathrm{GK}}{}^{f}$ are both increasing functions of the number of examples of the association rule, however $\mathrm{M}_{\mathrm{GK}}{}^{f}$ is more slowly increasing than *Confidence*.

• $\forall(X \to Y)$ such that X favors Y, one has:

- If their extensions are such that $X' \subseteq Y'$, then

$\mathrm{conf}(X \to Y) = \mathrm{M}_{\mathrm{GK}}{}^{f}(X \to Y) = 1 : X \to Y$ is said to be an exact rule .

- If $X' \cup Y'$, then $0 < \mathrm{M}_{\mathrm{GK}}{}^{f}(X \to Y) < \mathrm{conf}(X \to Y) < 1$, or

$\frac{1 - \mathrm{M}_{\mathrm{GK}}{}^{f}}{1 - \mathrm{conf}}(X \to Y) > 1 : X \to Y$ is said to be an approximate rule.

Let us write $EI(X \to Y) = (P(Y'/X') - P(Y'))$ the deviation from independence of Y to X.

Proposition 7 *For all normalized interestingness measure* μ *, one has:*

$\mu(X \to Y) = f(n, P(X'), P(Y'), P(X' \cap Y'))EI(X \to Y)$, if X favors Y ;

$g(n, P(X'), P(Y'), P(X' \cap Y'))EI(X \to Y)$, if X disfavors Y, where f et g are two real functions strictly positive and less or equall than $+1$.

Corollary 1 *All continued normalized interestingness measure produces most pertinent association rules than the standard measure Confidence*.

Finally, one obtains the canonical decomposition of any continued normalized interestingness measure.

Proposition 8 *All continued normalized interestingness measure* μ *is canonically decomposed depending of* M_{GK} *as :*

$\mu = \lambda \times \mathrm{M}_{\mathrm{GK}}{}^{f} \times 1_{f} + \beta \times \mathrm{M}_{\mathrm{GK}}{}^{d} \times 1_{d}$,

where 1_{f} is the indicator function of the event "*Premise favors Consequent*", 1_{d} the indicator function of the event "*Premise disfavors Consequent*", λ *and* β *being two real function belonging to*]0,1].

It is now obvious that we can define many algebra operations in the set $C(N)$.

Definition 3

- **Addition :** $\forall \mu, \nu \in C(N)$,

$$\mu \oplus \nu = \frac{(\mu^f + \nu^f)}{2} + \frac{(\mu^d + \nu^d)}{2}$$

- **Barycentric addition:** $\forall \mu, \nu \in C(N)$,

$$\forall a, b \in R_+^*,\ a\mu \oplus_B b\nu = \frac{(a\mu^f + b\nu^f)}{a+b} + \frac{(a\mu^d + b\nu^d)}{a+b}.$$

- **Product:** $\mu\mu' = 1_f\, \mu^f\, \mu'^f - 1_d\, \mu^d\, \mu'^d$

- **Power:** for $\alpha, \beta > 1$,

$$\mu^\alpha = 1_f (\mu^f)^\alpha + 1_d (-1)^{\alpha-1} (\mu^d)^\alpha$$

$\mu^{(\alpha,\beta)} = 1_f (\mu^f)^\alpha + 1_d (-1)^\gamma (\mu^d)^\beta$, with $\gamma = 1$, if β is even and 0 if not.

- **Supremum et Infimum:**

$$\mu \vee \mu' = 1_f\, \mu^f \vee \mu'^f + 1_d\, \mu^d \vee \mu'^d, \quad \mu \wedge \mu' = 1_f\, \mu^f \wedge \mu'^f + 1_d\, \mu^d \wedge \mu'^d,$$

$$\mu \hat{a} \mu' = 1_f\, \mu^f \vee \mu'^f + 1_d\, \mu^d \wedge \mu'^d.$$

Let us remark that the addition $\oplus$ is a particular case of the linear convex combination $\oplus_B$.

Proposition 9

- $C(N$ is closed in all these algebra operations defined above. Moreover, one has:

$|\mu^\alpha| < |\mu^{\alpha-1}|$ and $|\mu\mu'| < |\mu \wedge \mu'|$

- $C(N$ is closed in both supremum envelopping and infimum envelopping:

$\forall \mu, \nu \in C(N,\ \ max(\mu,\nu) = max(\mu^f, \nu^f)1_f + max(\mu^d, \nu^d)1_d \in C(N$ and

$min(\mu,\nu) = min(\mu^f, \nu^f)1_f + min(\mu^d, \nu^d)1_d \in C(N$

-

$$\forall \mu \in C(N), \forall m \in N, \forall n \in N^*, (\mu^f)^n 1_f + (\mu^d)^{2m+1} 1_d \in C(N), \mu^n \in C(N), \mu^{(n,m)} \in C(N).$$

- $C(N)$ is closed in $\oplus_B$, product, $\vee$ and $\wedge$.

By the above proposition 9, we deduce the infinity of the set $\mathsf{C}(\mathsf{N})$ and how to construct normalized quality measure. And M_{GK} appears playing an important basis role in $\mathsf{C}(\mathsf{N})$: M_{GK} is likely the simplest continued normalized interestingness measure. Such measures have the advantage to be convenient for mining both positive and negative association rules (Antonie & al., 2004)) In addition, as $\forall n \in \mathsf{N}, 0 < \frac{(\mu^f)^{n+1}}{(\mu^f)^n} = \mu^f < 1$, it is possible to construct a more selective continued normalized quality measure than $\mu^f, \forall \mu \in \mathsf{C}(\mathsf{N})$. However, the optimization problem in choosing the power n and the freshold must be solved.

Corollary 2 *For two continued normalized measures μ and ν, the canonical components of their "sum" are such that: $\forall a,b \in \mathsf{R}_+^*$, one has:*

$$(a\mu \oplus_B b\nu)^f = \frac{(a\mu^f + b\nu^f)}{a+b} = (\frac{(a\lambda_\mu + b\lambda_\nu)}{a+b})\mathrm{M}_{\mathrm{GK}}{}^f$$

and $$(a\mu \oplus_B b\nu)^d = \frac{(a\mu^d + b\nu^d)}{a+b} = (\frac{(a\beta_\mu + b\beta_\nu)}{a+b})\mathrm{M}_{\mathrm{GK}}{}^d.$$

Normalization process and characterization

Normalization and normalizability

Since any bounded interval of the type $[a,b]$ is homeomorphic to the interval $[-1,1]$, an affine function being the simplest bijection, it appears natural to search an affine function or partially affine one with dynamical coefficients eventually transforming an arbitrary non normalized interestingness measure. It would be possible to have a unifying view on the set of quality measures used in the litterature. We search a necessary and sufficent condition of such normalizability of a fixed measure μ. Let us write its associate normalized in $\mathsf{C}(\mathsf{N})$ as μ_n. Let us consider an association rule $X \to Y$ from a context. Let x_f and y_f (resp. x_d & y_d) be respectively the multiplying coefficient and centering coefficient of μ, in case of X favoring Y (resp. X disfavoring Y). Thus we have :

$$\mu_n(X \to Y) = \begin{cases} x_f . \mu(X \to Y) + y_f, & if \quad X \quad fav. \quad Y \\ x_d . \mu(X \to Y) + y_d, & if \quad X \quad disfav \quad Y \end{cases}$$

These four coefficients are determined by passing to unilateral limits in the referencing situations as incompatibility, independence (on the left and on the right) and logical implication. That is : Let $\mu_{imp}(X \to Y)$ the value of $\mu(X \to Y)$ at implication, $\mu_{ind}(X \to Y)$ the value of $\mu(X \to Y)$ at independence, and $\mu_{inc}(X \to Y)$ the value of $\mu(X \to Y)$ in case of incompatibility. In case of X favoring Y, one has:

$$\begin{cases} x_f \mu_{imp}(X \to Y) + y_f = 1 & logical\ implication \\ x_f \mu_{ind}(X \to Y) + y_f = 0 & independence\ from\ right \end{cases}$$

In case of X disfavoring Y, one obtains:

$$\begin{cases} x_d \mu_{ind}(X \to Y) + y_d = 0 & independence\ from\ left \\ x_d \mu_{inc}(X \to Y) + y_d = -1 & incompatibility \end{cases}$$

The corresponding equations system is linear and writed as below;

$$\begin{cases} x_f . \mu_{imp}(X \to Y) + y_f = 1 \\ x_f . \mu_{ind}(X \to Y) + y_f = 0 \\ x_d . \mu_{ind}(X \to Y) + y_d = 0 \\ x_d . \mu_{inc}(X \to Y) + y_d = -1 \end{cases} \tag{11}$$

Let M be the corresponding matrix. One has:

$$M = \begin{pmatrix} \mu_{imp}(X \to Y) & 1 & 0 & 0 \\ \mu_{ind}(X \to Y) & 1 & 0 & 0 \\ 0 & 0 & \mu_{ind}(X \to Y) & 1 \\ 0 & 0 & \mu_{inc}(X \to Y) & 1 \end{pmatrix}$$

Since, its determinant is $\det(M) = (\mu_{imp}(X \to Y) - \mu_{ind}(X \to Y))\,(\mu_{ind}(X \to Y) - \mu_{inc}(X \to Y))$, one has the strategic theorem below.

Theorem 1 A quality measure μ normalizable if and only if, for all association rule X→Y, the fowing conditions are satisfyed:

the quantities μ_{imp}(X→Y), μ_{ind}(X→Y) and μ_{inc}(X→Y) are finite;

the following inequalities are satisfied

$\mu_{imp}(X \to Y) \neq \mu_{ind}\ (X \to Y)$;

$\mu_{ind}(X \to Y) \neq \mu_{inc}\ (X \to Y)$.

Corollary 3 For any normalizable measure μ and any association rule X→Y the key coefficients are given by expressions below:

$$x_f = \frac{1}{\mu_{imp}(X \to Y) - \mu_{ind}(X \to Y)}, y_f = -\frac{\mu_{ind}(X \to Y)}{\mu_{imp}(X \to Y) - \mu_{ind}(X \to Y)};$$
$$x_d = \frac{1}{\mu_{ind}(X \to Y) - \mu_{inc}(X \to Y)}, y_d = -\frac{\mu_{ind}(X \to Y)}{\mu_{ind}(X \to Y) - \mu_{inc}(X \to Y)}.$$

;

Remark 3

• The coefficients x_f, x_d, y_f *and* y_d depend only on the margin probabilities $P(X')$ et $P(Y')$, and the quantities $\mu_{imp}(X \to Y), \mu_{ind}(X \to Y)$ and $\mu_{inc}(X \to Y)$ so.

• It is easy to obtain that $\mathrm{M}_{\mathrm{GK}n} = \mathrm{M}_{\mathrm{GK}}$. More generally, for all normalized measure $\mu \in \mathrm{C(N)}$, one has : $\mu_n = \mu$.

• Let us observe that for all μ whose associated normalized measure is $\mu_n = \mathrm{M}_{\mathrm{GK}}$, one has the inverse relation:

$$\mu(X \to Y) = \begin{cases} \dfrac{\mathrm{M}_{\mathrm{GK}}(X \to Y) - y_f}{x_f}, & if \quad X \quad fav. \quad Y \\ \dfrac{\mathrm{M}_{\mathrm{GK}}(X \to Y) - y_d}{x_d}, & if \quad X \quad disfav \quad Y \end{cases}$$

This reciprocal relation will allow comparing, via M_{GK}, two normalizable measures μ and μ' on a fixed association rule R_1 : for instance, if R_1 is valid according to M_{GK}, but non valid according to μ and μ', then these two measures under-evaluate approximate rules ; otherwise, if R_1 is valid according to μ and μ', but not to M_{GK}, then they uper-evaluate rules. Thus, M_{GK} plays important unifying role in the subset of normalizable measures associated to M_{GK}.

Example of normalization

As illustration of the normalization process, below is presented the example of *Confidence*. Let $X \to Y$ be a rule from a context. **Confidence**: $\text{conf}(X \to Y) = P(Y' \mid X')$, $\text{conf}_{inc}(X \to Y) = 0 \neq -1$, $\text{conf}_{ind}(X \to Y) = P(Y') \neq 0$ and $\text{conf}_{imp}(X \to Y) = 1$, so $det(M) = 1 - P(Y') \neq 0$. According to the above theorem?, *Confidence* is normalizable:

$$x_f = \frac{1}{1-P(Y')}, y_f = -\frac{P(Y')}{1-P(Y')} x_d = \frac{1}{p(Y')}, y_d = -1$$

Say

$$\text{conf}_n(X \to Y) = \begin{cases} \dfrac{P(X' \cap Y') - P(X')P(Y')}{P(X')(1-P(Y'))}, if\ X\ fav.Y \\ \dfrac{P(X' \cap Y') - P(X')P(Y')}{P(X')P(Y')}, if\ X\ disfav Y \end{cases}$$

Thus, $\text{conf}_n(X \to Y) = \text{M}_{\text{GK}}(X \to Y)$. Moreover, it is easy to show that $\text{M}_{\text{GK}} \leq Confidence$ and for some itemsets independent X and Y, one can obtain $\text{conf}(X \to Y) > 0.90$ against the natural $\text{M}_{\text{GK}}(X \to Y) = 0$. Thus, *Confidence* uper-evaluates association rules. In extension, it is easy to verify the result below.

Proposition 10

- • The twenty quality measures M_{GK}, Support, Confidence, Recall, Lift, laverage, Centered-Confidence, Certitude factor, Laplace, ϕ-coefficient, Piatetsky-Shapiro, Cosinus, Accuracy, Little contradiction (Moindre contradiction in french), Lovinger, Kappa, Implication index, Specificity and negative reliability are normalizable and associated to M_{GK};
- • The five measures Jaccard, Zhang, Q-Yule, Y-Yule, J-measure are normalizable but not associated to M_{GK};
- • The seven measures Côte multiplier, Sebag, Conviction, Odd Ratio, Klosgen, Gain informationnel et Ratio of counter-example are not affine homomorphic normalizable.

This last proposition confirms the unifying role of M_{GK}, except likely a little set of few measures (For other results, see (Totohasina, 2008).

Conclusion

In the present work, it is shown that the normalized interestingness measure M_{GK} is more selective and more pertinent, *i.e.*, it does not produce redundant rules and it systematically avoids independence, than the standard *Confidence*, for associations rules of the same kind, M_{GK} measures both the distance from independence and the intensity of statistical (or approximate) implication between two itemsets. Moreover, unlike *Confidence*, since M_{GK} is asymptoticaly satisfying the condition of equilibrium, M_{GK} deals conveniently with large data bases. Regarding its coherence with mutual attraction and mutual repulsion of two itemsets, M_{GK} is less ambiguous and more understandable than the standard independence Khi-square testing and than *Confidence*. Netherveless, regarding the intuitive word *Confidence* in the popular language and because of the concept of conditional probability, we think it is necessary to keep using *Confidence* but only for $M_{GK\text{-valid}}$ association rules. We also think it is profitable to consider critical values of most of interestingness measures depending on contingency table, like the continued normalized interestingness measures, for increasing such valid rules relevance. To end, the present work has shown that M_{GK} plays a central unifying role in the infinite set of such quality measures and in the set of non normalizable probabilistic measures.

References

S. Guillaume(2000), *Traitement des données volumineuses. Mesures et algorithmes d'extraction des règles d'association et règles ordinales*, Universté de Nantes, Phd.

X. Wu,C. Zhang,S. Zhang(2004), Efficient mining of both positive and negative association rules, *ACM Transactions on information Systems*, 381–405.

J. Diatta, H. Ralambondrainy, A. Totohasina(2007), Towards a Unifying Probabilistic ImplicativeNormalized Quality Measure for Association Rules, *Quality Measures in Data Mining book*, 10, 237–250.

S. Ferré(2002), *Systèmes d'information logique : un paradigme logico-contextuel pour interroger, naviguer et apprendre*, Université de Rennes I, Phd.

R. Agrawal T. Imielinski A. Swami(1993), Mining association rules between sets of items in large databases, *Proc. of the ACM SIGMOD International Conference on Management of Data, editor P. Buneman and S. Jajodia, Washington,U.S.A.*.

I.C. Lerman(1984),*Classification et analyse ordinale des donnés*, Dunod.

J. Blanchard, F. Guillet,H. Briand, R. Gras(2005), Assessing rule with a probabilistic measure of deviation from equilbrium, *Proc. of 11th International Symposium on Applied Stochastic Models and Data Analysis ASMDA, ENST, Brest, France*, 74, 191–200.

R. Gras, Almouloud S., M. Bailleul, A. Larher, M. Polo, H. Ratsimba-Rajohn, A. Totohasina(1996), *L'Implication statistique. Nouvelle méthode exploratoire d'analyse des données*, Coll. Recherche en Didactique des Mathématiques, Édit. La Pensée Sauvage, Grenoble.

D.Sánchez, M. Vila, L. Cerdal, J.M. Serrano(2008), Association rules applied to credit card fraud detection, *Expert Systems with Applications(2008)*, doi.10.1016/ j.eswa.2008.02.001

G. Piatetsky-Shapiro(1991), Knowledge discovery in real data bases, *AI Magazine*, 68–70.

R. J. Hilderman,H. J. Hamilton(1999), Knowledge discovery and interestingness measures: A survey, *Technical Report CS 99-04, Department of Computer Science, University of Regina*.

M. L. Antonie, O. R. Zaïane(2004), Mining positive and negative association rules : An approch for confined rules, *Proc. 8th Int. Conf. on Principle and Practice of Knowledge Discovery in Databases (PKDD'04)* 27–38.

S. Brin, R. Motwani, C. Silverstein(1997), Beyond market baskets: Generalizing association rules to correlation, *Proc. of the ACM SIGMOD Conference*, 265–276.

A. Freitas(1999) , On rule interestingness measures, *Journal Knowledge-Based System*, 309–315.

A. Totohasina(2008), *Contributions to studying association rules interestingness measures: normalization under five constraints and case study of* M_{GK} *: properties, rules composites basis and extension for applying objective in Statistic and in Physical Sciences.*, Habilitation to Supervizing Researchs thesis (H.D.R. degree in Malagasy system), University of Antsiranana, Madagasikara (in French)

A. Totohasina, H. Ralambondrainy(2005), ION : a pertinent new measure for mining information from many types of data, *Proceedings of IEEE SITIS'05, Yaoundé, Cameroon,* 202–207.

F. Alonso, J.P. Carença-Valente, A. L. Gonzalez, C. Montes(2002), Combining expert knowledge and data mining in medical diagnosis domain, *Expert Systems with Applications 23(2002), 367-375.*