On selecting interestingness measures for association rules: user oriented description and multiple criteria decision aid

Philippe Lenca[a], Patrick Meyer[b], Benoît Vaillant[a], and Stéphane Lallich[c]

[a]GET/ENST Bretagne – CNRS UMR 2872 TAMCIC
Technopôle de Brest Iroise – CS 83818, 29238 Brest Cedex, France
philippe.lenca@enst-bretagne.fr, benoit.vaillant@enst-bretagne.fr

[b]University of Luxembourg - Applied Mathematics Unit
162a, avenue de la Faïencerie, L-1511 Luxembourg
patrick.meyer@uni.lu

[c]University of Lyon 2, ERIC Laboratory
5 avenue Pierre Mendès-France, 69676 Bron Cedex, France
stephane.lallich@univ-lyon2.fr

Data mining algorithms, especially those used for unsupervised learning, generate a large quantity of rules. In particular this applies to the APRIORI family of algorithms for the determination of association rules. It is hence impossible for an expert in the field being mined to sustain these rules. To help carry out the task, many measures which evaluate the interestingness of rules have been developed. They make it possible to filter and sort automatically a set of rules with respect to given goals. Since these measures may produce different results, and as experts have different understandings of what a *good* rule is, we propose in this article a new direction to select the *best* rules: a two-step solution to the problem of the recommendation of one or more user-adapted interestingness measures. First, a description of interestingness measures, based on meaningful classical properties, is given. Second, a multicriteria decision aid process is applied to this analysis and illustrates the benefit that a user, who is not a data mining expert, can achieve with such methods.

Keywords: data mining, interestingness measures, association rules, multiple criteria analysis.

1. Introduction

One of the main objectives of Knowledge Discovery in Databases (KDD) is to produce interesting rules with respect to some user's point of view. This user is not assumed to be a data mining expert, but rather an expert in the field being mined. Moreover, it is well known that the interestingness of a rule is difficult to evaluate with objectivity. Indeed, this estimation greatly depends on the expert user's interests (Klemettinen et al., 1994; Hilderman and Hamilton, 2001). Ideally, a rule should be *valid, new and comprehensive* (Fayyad et al., 1996) but these generic terms cover a large number of various

situations according to the context. It is also well-known that data mining algorithms may produce huge numbers of rules and that the end user is then unable to analyse them manually.

In this context, interestingness measures play an essential role in KDD processes in order to find the *best rules* (in a post-processing step). Depending on the user's objectives, the data mining experts should choose an appropriate interestingness measure in order to filter the huge amount of rules. Nevertheless, as this study shows, this choice is not easy and can be facilitated by the use of a Multiple Criteria Decision Aid (MCDA) approach.

In fact, this choice is hard to make since rule interestingness measures have many different qualities or flaws (Tan et al., 2002; Lenca et al., 2004). What is more, some of these properties are incompatible. Therefore there is no *optimal* measure, and a way of solving this problem is to try to find good compromises (Lenca et al., 2003b; Francisci et al., 2003). A well-known example of such a controversial measure is the support. On the one hand, it is greatly used for filtering purposes in KDD algorithms (Agrawal et al., 1993; Pasquier et al., 1999), since its antimonotonicity property simplifies the large lattice that has to be explored. On the other hand, it has almost all the flaws a user would like to avoid, such as variability of the value under the independence hypothesis or for a logical rule (Piatetsky-Shapiro, 1991). Bayardo and Agrawal (1999), Tan and Kumar (2000), Hilderman and Hamilton (2003), Lallich and Teytaud (2004), McGarry (2005), Blanchard et al. (2005), Lenca et al. (2006), Suzuki (2006) for instance, have formally extracted several specificities of measures/interestingness.

The importance of objective evaluation criteria of interestingness measures has already been studied by Piatetsky-Shapiro (1991) and Freitas (1999) on restricted sets of measures and properties. However, the relevance of these criteria for the selection of the right measure is still difficult to establish. In Tan et al. (2002), the authors provide a comparative study according to certain properties and an original approach to the selection of measures by an expert. However, this approach does not exploit the above-mentioned comparative study: from the set of rules resulting from a data mining algorithm, the authors propose to extract a small subset of rules where the measures give very different results. The authors experimentally establish that the diversity of the results on the subset of rules enables the user to efficiently select an appropriate measure.

This article can be seen as an alternative contribution to Tan et al. (2002). We propose a two-step process. First, we provide a comparative description of a set of measures through the expression of a list of properties. These properties partly differ from those evaluated in Tan et al. (2002), since some of the latter ones do not apply efficiently, in our opinion, to the interestingness of association rules, and others do not make any distinction between the different interestingness measures which are studied. In addition, we introduce and study new properties, such as for example the easiness to fix a threshold, or intelligibility. Second, we propose to use a MCDA method on some classical measures and the previously identified properties to help select a measure which is concordant with the user's objectives. MCDA methods have already proved their utility in different fields (Roy, 1996; Roy and Bouyssou, 1993). We argue in this paper that an MCDA method could be profitable for the specific problem of the selection of an appropriate interestingness measure.

This paper is organised as follows. In Section 2 we briefly recall the context of associ-

ation rule discovery. We introduce in Section 3 a representative list of existing measures, frequently used in the scientific context of association rules. In Section 4, we report some experimental results that underline the diversity of ranks obtained by the different measures. In Section 5 we define the problem within an MCDA context. We propose in Section 6 a list of 8 meaningful properties (from the user's point of view) and evaluate the previous list of measures according to them. Section 7 is dedicated to the use of the MCDA method PROMETHEE, using different users' preferences scenarios. Finally, we conclude in Section 8.

2. On association rule mining

As defined in Agrawal et al. (1993), given a typical market-basket (transactional) database $\mathtt{E}$, an association rule $\mathtt{A} \rightarrow \mathtt{B}$ means *if someone buys the set of items* $\mathtt{A}$*, then he/she probably also buys item* $\mathtt{B}$. Such sets of items are usually called itemsets.

The problem of mining for association rules involves discovering all the rules that correlate the presence of one itemset with another under minimum support and minimum confidence conditions:

- an association rule is an assertion $\mathtt{A} \rightarrow \mathtt{B}$ where $\mathtt{A}$ and $\mathtt{B}$ are two itemsets and $\mathtt{A} \cap \mathtt{B} = \emptyset$,
- the support of $\mathtt{A} \rightarrow \mathtt{B}$ is the percentage of transactions that contain $\mathtt{A}$ and $\mathtt{B}$,
- the confidence of $\mathtt{A} \rightarrow \mathtt{B}$ is the ratio of the number of transactions that contain $\mathtt{A}$ and $\mathtt{B}$ against the number of transactions that contain $\mathtt{A}$.

The well-known APRIORI algorithm (Agrawal et al., 1993) proceeds in two steps within the support-confidence framework (minimum support and confidence thresholds have to be fixed by the user) in order to extract association rules:

- find frequent itemsets (the sets of items which occur more frequently than the minimum support threshold) with the frequent itemset property (any subset of a frequent itemset is frequent; if an itemset is not frequent, none of its supersets can be frequent) for efficiency reasons. Thus starting from $k = 1$, APRIORI generates itemsets of size $k + 1$ from frequent itemsets of size k,
- generate rules from frequent itemsets and filter them with the minimum confidence threshold.

Unfortunately APRIORI tends to generate a large number of rules. It is hence impossible for an expert of the field being mined to sustain these rules. The validation of the knowledge extracted within a KDD process by a field expert requires a filtering step. One of the classical methods relies on the use of subjective and objective interestingness measures. Subjective measures are *user-driven* in the sense that they take into account the user's *a priori* knowledge while objective measures are said to be *data-driven* and only take into account the data cardinalities. We focus in this study on objective measures. For a discussion about subjective aspects of rule interestingness measures, the reader can refer to Silberschatz and Tuzhilin (1995), Liu et al. (1997) and Liu et al. (2000).

Strong rules (interesting rules within the support and confidence framework) satisfy the minimum support and minimum confidence thresholds. Nevertheless, they are not necessarily interesting either from an expert's point of view or from a statistical one. For example, high confidence should not be confused with high correlation, nor with causality between the antecedent and the consequent of a rule (Brijs et al., 2003).

As an illustrative example, consider a classical dataset of 10.000 transactions in a shop. 6.000 transactions include computer games, 7.500 include movies and 4.000 include both items. Let the minimum support be 30% and the minimum confidence be 60%. Thus, the strong rule *buy computer games* $\rightarrow$ *buy movies* is indeed retained with a support of 40% and a confidence of 66%. However, this strong rule is misleading since the probability of purchasing movies is 75%.

The data mining experts should select and apply an interestingness measure which is compatible with the user's objectives. Nevertheless, the measures have many different and conflicting qualities and flaws. Moreover, on a given set of rules, they may generate different rankings and hence highlight different pieces of information.

In the next section we present 20 association rule interestingness measures that will be used in our future discourse.

3. Selected measures

The 20 measures we list here evaluate the interestingness of association rules.

It is very important to differentiate between the association rule $\mathtt{A} \rightarrow \mathtt{B}$, which focuses on cooccurrence and gives asymmetric meaning to $\mathtt{A}$ and $\mathtt{B}$, and the logical implication $\mathtt{A} \Rightarrow \mathtt{B}$ or the equivalence $\mathtt{A} \Leftrightarrow \mathtt{B}$ (see Lallich and Teytaud (2004)).

$\mathtt{A}\backslash\mathtt{B}$	0	1	total
0	$p_{\bar{a}\bar{b}}$	$p_{\bar{a}b}$	$p_{\bar{a}}$
1	$p_{a\bar{b}}$	p_{ab}	p_a
total	$p_{\bar{b}}$	p_b	1

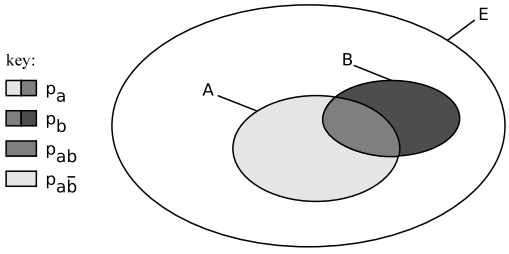


Figure 1. Notations

Interesting measures for association rules are usually defined using frequency counts or relative frequencies as presented in Figure 1. This kind of analysis is a particular case of the analysis of a contingency table, introduced by Hajek et al. (1966) within the GUHA method and developed later on by Rauch and Simunek (2001) in the 4FT-MINER tool. Given a rule $\mathtt{A} \rightarrow \mathtt{B}$, we note:

- $n = |\mathtt{E}|$ the total number of records
- $n_a = |\mathtt{A}|$ the number of records satisfying $\mathtt{A}$

- $n_b = |\mathtt{B}|$ the number of records satisfying $\mathtt{B}$
- $n_{ab} = |\mathtt{A} \cap \mathtt{B}|$ the number of records satisfying both $\mathtt{A}$ and $\mathtt{B}$ (the examples of the rule)
- $n_{a\bar{b}} = |\mathtt{A} \cap \bar{\mathtt{B}}|$ the number of records satisfying $\mathtt{A}$ but not $\mathtt{B}$ (the counter-examples of the rule)

For $\mathtt{X} \subset \mathtt{E}$, we note p_x instead of n_x/n when we consider relative frequencies rather than absolute frequencies.

It is clear that, given n, n_a and n_b (or p_a and p_b), knowing one cell of the table of Figure 1 is enough to deduce the other ones. For example, if one knows p_{ab}, then $p_{a\bar{b}} = p_a - p_{ab}$, $p_{\bar{a}\bar{b}} = 1 - p_a - p_b + p_{ab}$ and $p_{\bar{a}b} = p_b - p_{ab}$.

We have restricted the list of measures evaluated in this paper to decreasing ones with respect to $n_{a\bar{b}}$, all marginal frequencies being fixed. This choice reflects the common assertion that the fewer counter-examples ($\mathtt{A}$ true and $\mathtt{B}$ false) to the rule there are, the higher the interestingness of the rule is. Therefore, some measures like χ^2, Pearson's r^2, Goodman and Smyth's J-measure or Pearl's measure have been excluded from this list.

For a given decreasing monotonic measure μ it is then possible to select interesting rules by fixing a threshold α and keeping only the rules satisfying $\mu(A \rightarrow B) \geq \alpha$. Note that the value of this threshold α has to be fixed by the expert. The same threshold is considered for all the rules extracted during the data mining process. It is hence an important issue. A well known situation of such a critical point is the determination of a minimal support and confidence threshold in the Apriori algorithm (Agrawal et al., 1993).

The interestingness measures are given with their bibliographical references in Table 1. Their formulae in absolute and relative frequencies are given in Table 2. Their values for various reference situations are given in Table 3.

We kept the well-known support and confidence: these are the two most frequently used measures in association rule extraction algorithms based on the selection of frequent itemsets (Agrawal et al. (1993), Pasquier et al. (1999)).

Many other measures are linear transformations of the confidence, enhancing it, since they enable comparisons with p_b. This transformation is generally achieved by centring the confidence on p_b, using different scale coefficients (centred confidence, Piatetsky-Shapiro's measure, Loevinger's measure, Zhang's measure, correlation, implication index and least contradiction). In case the of the lift, the confidence is divided by p_b.

Other measures, like Sebag-Schoenauer's or the examples and counter-examples rate, are monotonically increasing transformations of the confidence, while the information gain is a monotonically increasing transformation of the lift. Thus, measures that are monotonically increasing transformations of the confidence inherit the confidence's properties, and so on.

Therefore, such measures will rank the rules according to the same order (Lallich (2002)). However, they are different according to the user's points of view studied later in this article. A user will therefore be able to choose one of these "equal" measures on the basis of his/her preferences. For example, the conviction and Loevinger rank similarly, nevertheless, they differ on the linearity criterion which is introduced in Section 6.

Table 1
List of selected measures

	Name	References
BF	Bayes factor	Jeffreys (1935)
CenConf	centred confidence	
Conf	confidence	Agrawal et al. (1993)
Conv	conviction	Brin et al. (1997b)
ECR	examples and counter-examples rate	
IG	information gain	Church and Hanks (1990)
- ImpInd	implication index	Lerman et al. (1981)
IntImp	intensity of implication	Gras et al. (1996)
Kappa	Kappa coefficient	Cohen (1960)
Lap	Laplace	Good (1965)
LC	least contradiction	Azé and Kodratoff (2002)
Lift	Lift	Brin et al. (1997a)
Loe	Loevinger	Loevinger (1947)
PDI	probabilistic discriminant index	Lerman and Azé (2003)
PS	Piatetsky-Shapiro	Piatetsky-Shapiro (1991)
r	Pearson's correlation coefficient	Pearson (1896)
Seb	Sebag and Schoenauer	Sebag and Schoenauer (1988)
Sup	support	Agrawal et al. (1993)
TEII	truncated entropic intensity of implication	Lallich et al. (2005)
Zhang	Zhang	Zhang (2000)

Table 2
Association rule interestingness measures

	Absolute definitions	Relative definitions
BF	$\frac{n_{ab}n_{\bar{b}}}{n_b n_{a\bar{b}}}$	$\frac{p_{b/a}/p_{\bar{b}/a}}{p_b/p_{\bar{b}}} = \frac{p_{a/b}}{p_{a/\bar{b}}}$
CenConf	$\frac{n_{ab}}{n_a} - \frac{n_b}{n}$	$p_{b/a} - p_b$
Conf	$\frac{n_{ab}}{n_a}$	$p_{b/a}$
Conv	$\frac{n_a n_{\bar{b}}}{n n_{a\bar{b}}}$	$\frac{p_a p_{\bar{b}}}{p_{a\bar{b}}}$
ECR	$\frac{n_{ab}-n_{a\bar{b}}}{n_{ab}}$	$1 - \frac{p_{a\bar{b}}}{p_{ab}}$
IG	$\log(\frac{n n_{ab}}{n_a n_b})$	$\log \frac{p_{ab}}{p_a p_b}$
-ImpInd	$-\frac{n_a n_b - n n_{ab}}{\sqrt{n n_a n_{\bar{b}}}}$	$-\sqrt{n}\frac{p_{a\bar{b}}-p_a p_{\bar{b}}}{\sqrt{p_a p_{\bar{b}}}}$
IntImp	$P\left[\mathcal{N}(0,1) \geq \text{ImpInd}\right]$	
Kappa	$2\frac{n n_{ab}-n_a n_b}{n n_a + n n_b - 2 n_a n_b}$	$2\frac{p_{ab}-p_a p_b}{p_a+p_b-2p_a p_b}$
Lap	$\frac{n_{ab}+1}{n_a+2}$	$\frac{p_{b/a}+\frac{1}{n p_a}}{1+\frac{2}{n p_a}}$
LC	$\frac{n_{ab}-n_{a\bar{b}}}{n_b}$	$\frac{p_{ab}-p_{a\bar{b}}}{p_b}$
Lift	$\frac{n n_{ab}}{n_a n_b}$	$\frac{p_{b/a}}{p_b}$
Loe	$\frac{n n_{ab}-n_a n_b}{n_a n_{\bar{b}}}$	$\frac{p_{b/a}-p_b}{1-p_b}$
PDI	$P\left[\mathcal{N}(0,1) > \text{ImpInd}^{CR/\mathcal{B}}\right]$	
PS	$n_{ab} - \frac{n_a n_b}{n}$	$n(p_{ab} - p_a p_b)$
R	$\frac{n n_{ab}-n_a n_b}{\sqrt{n_a n_b n_{\bar{a}} . n_{\bar{b}}}}$	$\frac{p_{ab}-p_a p_b}{\sqrt{p_a p_{\bar{a}} p_b p_{\bar{b}}}}$
Seb	$\frac{n_{ab}}{n_{a\bar{b}}}$	$\frac{p_{ab}}{p_{a\bar{b}}}$
Sup	$\frac{n_{ab}}{n}$	p_{ab}
TEII	$[i_t(\mathtt{A} \to \mathtt{B}) \times \text{IntImp}(\mathtt{A} \to \mathtt{B})]^{1/2}$	
Zhang	$\frac{n n_{ab}-n_a n_b}{\max\{n_{ab}n_{\bar{b}}, n_b n_{a\bar{b}}\}}$	$\frac{p_{ab}-p_a p_b}{\max\{p_{ab}p_{\bar{b}}; p_b p_{a\bar{b}}\}}$

$i_t(\mathtt{A} \to \mathtt{B})$ is the (truncated) inclusion index of $\mathtt{A} \to \mathtt{B}$, defined as:

$$i_t(A \subset B) = \left[(1 - H^*(B/A)^\alpha)\left(1 - H^*(\overline{A}/\overline{B})^\alpha\right)\right]^{\frac{1}{2\alpha}}$$

where $H^*(X/Y) = 1$, if $p_{x/y} > \max\{0.5;\ p_x\}$

and $H^*(X/Y) = -p_{x/y}\log_2 p_{x/y} - (1-p_{x/y})\log_2(1-p_{x/y})$ otherwise

$\mathcal{N}(0,1)$ stands for the centred and reduced normal distribution function

$\text{ImpInd}^{CR/\mathcal{B}}$ corresponds to ImpInd, centred reduced (CR) for a rule set $\mathcal{B}$

Table 3
Interestingness measures reference situations

	value at incompatibility (minimum)	value at independence	value for a logical rule (maximum)
BF	0	1	$+\infty$
CenConf	$-\frac{n_b}{n}$	0	$\frac{n_{\bar{b}}}{n}$
Conf	0	$\frac{n_b}{n}$	1
Conv	$\frac{n_{\bar{b}}}{n}$	1	$+\infty$
ECR	$-\infty$	$\frac{n_b - n_{\bar{b}}}{n_b}$	1
IG	$-\infty$	0	$\log \frac{n}{n_b}$
-ImpInd	$-\frac{\sqrt{n_a} n_b}{\sqrt{n n_{\bar{b}}}}$	0	$\sqrt{\frac{n_a n_{\bar{b}}}{n}}$
IntImp	$P\left[\mathcal{N}(0,1) > \frac{\sqrt{n_a} n_b}{\sqrt{n n_{\bar{b}}}}\right]$	0.5	$P\left[\mathcal{N}(0,1) > -\sqrt{\frac{n_a n_{\bar{b}}}{n}}\right]$
Kappa	$-2\frac{n_a n_b}{n_a n_{\bar{b}} + n_{\bar{a}} n_b}$	0	$2\frac{n_a n_{\bar{b}}}{n_a n_{\bar{b}} + n_{\bar{a}} n_b}$
Lap	$\frac{1}{n_a+2}$	$\frac{n_a n_b + n}{n n_a + 2n}$	$\frac{n_a+1}{n_a+2}$
LC	$-\frac{n_a}{n_b}$	$\frac{n_a(n_b - n_{\bar{b}})}{n n_b}$	$\frac{n_a}{n_b}$
Lift	0	1	$\frac{n}{n_b}$
Loe	$-\frac{n_b}{n_{\bar{b}}}$	0	1
PDI	$P\left[\mathcal{N}(0,1) > \frac{\frac{\sqrt{n_a} n_b}{\sqrt{n n_{\bar{b}}}} - \mu}{\sigma}\right]$	$P\left[\mathcal{N}(0,1) > -\frac{\mu}{\sigma}\right]$	$P\left[\mathcal{N}(0,1) > -\frac{\sqrt{\frac{n_a n_{\bar{b}}}{n}} + \mu}{\sigma}\right]$
PS	$-\frac{n_a n_b}{n}$	0	$\frac{n_a n_{\bar{b}}}{n}$
R	$-\sqrt{\frac{n_a n_b}{n_{\bar{a}} n_{\bar{b}}}}$	0	$\sqrt{\frac{n_a n_{\bar{b}}}{n_{\bar{a}} n_b}}$
Seb	0	$\frac{n_b}{n_{\bar{b}}}$	$+\infty$
Sup	0	$\frac{n_a n_b}{n^2}$	$\frac{n_a}{n}$
TEII	0	0	$\sqrt{\text{IntImp}}$
Zhang	-1	0	1

μ is the mean value of ImpInd on a given rule set
σ is the standard deviation of ImpInd on a given rule set

Similarly, the lift and the information gain are also such measures, but they differ on the linearity and the intelligibility criteria. A last example is the pair composed of Sebag-Schoenauer and the examples and counter-examples rate which differ on linearity and intelligibility.

Some measures focus on counter-examples, like the conviction or the above-cited implication index. In its original definition, ImpInd models the number of counter-examples under null hypothesis. Thus, in order to have a decreasing quality measure with respect to $n_{a\bar{b}}$, we consider -ImpInd. This latter measure is the basis of several different probabilistic measures like the probabilistic discriminant index, the intensity of implication, or its entropic version, which takes into account an entropic coefficient, enhancing the discriminant power of the intensity of implication. For the intensity of implication, the statistical law was approximated using the centred and reduced normal distribution function. In this paper we use the truncated entropic intensity of implication, TEII, presented in Lallich et al. (2005), a more consistent definition of the entropic intensity of implication (Gras et al., 2001). Compared to EII, TEII has a constant value at the independence situation under certain conditions.

Laplace's measure is a variant of the confidence, taking the total number of records n into account.

The Bayes factor, also called sufficiency by Kamber and Shingal (1996), is the ratio of the odd of `B/A` against the prior odd of `B`. It has been thoroughly studied by Kamber and Shingal (1996), Lallich and Teytaud (2004), Greco et al. (2004).

The following section presents a comparison of the preorders generated by the measures on an experimental dataset. This comparison highlights the problem of selecting the "best" rules, and thus the necessity of using a measure adapted to the user's needs.

4. Experimental comparison of total preorders

In order to get an idea of the difficulty of selecting the subset of the N best rules, we study the total preorders induced by the measures' values on rule sets.

This comparison is based on counts over all the possible couples of rules. We simplify the mathematical background introduced in Giakoumakis and Monjardet (1987) to its simplest form, and consider 4 different situations, for any two rules, and two measures:

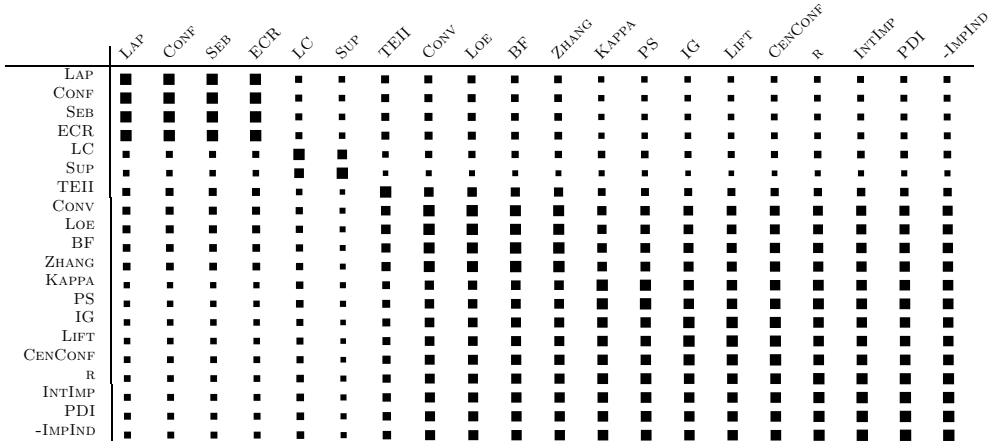
- there is *strict agreement* if both measures assign a strictly higher value to one of the rules over the other,
- there is *semi-agreement* if only one of the measures evaluates the two rules as of equal quality,
- there is *large agreement* if both measures evaluate the two rules as of equal quality,
- and there is *strict disagreement* if one of the measures evaluates the quality of a rule strictly higher than the other one, the second measure making the inverse evaluation.

In Giakoumakis and Monjardet (1987), 16 coefficients for preorder comparison based on such counts over possible situations are studied. Moreover, Lingoes (1979) defines the τ_1 coefficient, derived from Kendall's τ coefficient. τ_1 takes its values in $[-1; 1]$, the

maximum value being obtained when both preorders are equal. In this case, there are only strict agreements, or large agreements. The minimum value is obtained if, for any couple of different rules, there is either strict disagreement or semi-agreement. In the first case, both measures rank the rules in the same way and the subset of the N best rules is the same, for any N (see for example Table 5, measures Conv and Loe). On the contrary, in the second case, the order of the rules is reversed.

Using the Herbs tool developed by Vaillant et al. (2003), we computed the values of τ_1 for the 20 measures on the `cmc` database (*contraceptive method choice*, Lim et al. (2000), a subset of the 1987 National Indonesia Contraceptive Prevalence Survey). The rule set is composed of 2878 rules, generated by the Apriori algorithm implementation of Borgelt and Kruse (2002) with a support threshold of 5% and a confidence threshold of 60%. The results are presented in Table 4. The length of the side of each square is equal to $\frac{\tau_1+1}{2}$ (a linear transformation of τ_1 into $[0, 1]$). The lines and columns have been reorganised in order to highlight groups of similar measures using the Amado method (Chauchat and Risson, 1998), which is based on the works of Bertin (1977). For more experimental results the reader can refer to Vaillant et al. (2004) in which a thorough study carried out on 10 databases is presented. These results lead to a comparison of classifications built on both experimental and formal aspects.

Table 4
Comparison of total preorders between 20 measures



We have only 8 negative values, the lowest being -0.0331 for (Sup, TEII). The average value of τ_1 is 0.54, and the variance is 0.11. Some of the values are equal to 1, and this could have been predicted as in these cases the measures are monotonically increasing transformations of one another, like for (Conv, Loe) or (IG, Lift).

This means that although some measures do generate the same rankings, there are some significant differences. Thus, the subset of the N best rules may differ, depending on the measure which is used. This is illustrated in Table 5, which presents the number of rules in common within the subsets of N best rules for 4 measures. Clearly, these subsets are of size at least equal to N for measures which rank rules in the same order,

Table 5
Number of best rules in common in the subset of N best rules

N	20	50	70	100	150	200	400
CenConf & Conv	0	0	10	25	62	86	179
CenConf & Loe	0	0	10	25	62	86	179
CenConf & BF	0	0	10	25	64	94	210
Conv & Loe	62	62	70	100	150	201	401
Conv & BF	62	62	70	97	135	170	369
Loe & BF	62	62	70	97	135	170	369

such as Conv and Loe. The reason why this value may be above N is that rules equally evaluated by the measures are all included in the subsets. By taking such a closer look at top ranked rules, we see that CenConf disagrees with the three other measures on the 50 most interesting rules. Moreover, as some measures do generate the same rankings, the user may freely pick out from among them the one that best fits his preferences, without any loss of interesting rules.

These two remarks lead us to develop a description of interestingness measures, based on user preferences, in order to assist him in the task of selecting a good measure, adapted to his point of view.

The following section presents the properties that have been retained to describe the different measures.

5. Positioning of the problem

We have shown that the search for the best rules among a vast set of rules generated by a KDD procedure is directly linked to the search and the use of a good interestingness measure. As measures can be described by properties, we will consider an MCDA framework. From the user's point of view, the problem can then be resumed as a search for finding the best measure(s) according to the context. This context is defined by many parameters like the nature of the data (what is their type, do they suffer from noise, how imbalanced is the distribution of each attribute?), the type of rule extraction algorithm (what are its biases?), the goals, and the preferences of the user. In this article we focus on the latter two points.

We define the problem by considering a sextuplet $<\mathcal{D}, \mathcal{R}, \mathcal{M}, \mathcal{A}, \mathcal{P}, \mathcal{F}>$ where:

- $\mathcal{D}$ is a dataset. The data are described by a list of attributes;
- $\mathcal{R}$ is a set of association rules $\mathtt{A} \rightarrow \mathtt{B}$ which can be applied to $\mathcal{D}$;
- $\mathcal{M}$ is a set of interestingness measures of the rules of $\mathcal{R}$ (see Section 3);
- $\mathcal{A}$ is a set of properties which describe the characteristics of the measures of $\mathcal{M}$ (see Section 6);
- $\mathcal{P}$ is a set of preferences expressed by the expert user (of the field of $\mathcal{D}$) on $\mathcal{A}$ in relation to his objectives. The major difficulty in the construction of $\mathcal{P}$ is the

formalisation of the user's objectives. They are often given in natural language and a non-trivial task is to keep their semantics;

- $\mathcal{F}$ is a set of evaluation criteria of the measures of $\mathcal{M}$. $\mathcal{F}$ is built on the basis of the sets $\mathcal{A}$ and $\mathcal{P}$. In brief, one can say that $\mathcal{F}$ corresponds to an evaluation of the quality measures of $\mathcal{M}$ on the properties of $\mathcal{A}$ by taking into account the preferences of $\mathcal{P}$;

The quality measures considered in this study evaluate only the individual quality of rules. We do not evaluate the quality of the whole set of rules $\mathcal{R}$.

Two actors take part in this analysis: the user, an expert in the data (expert of $\mathcal{D}$ and $\mathcal{R}$), who tries to select the *best* rules of $\mathcal{R}$ and the analyst, a specialist in MCDA procedures and in KDD, who tries to help the expert. We call the first one E_r and the second one E_a. Consequently, the main problem is to translate the properties of $\mathcal{A}$ into a set $\mathcal{F}$ of criteria by considering the preferences $\mathcal{P}$ in view of determining the *best* measures. Note that the sets $\mathcal{D}$, $\mathcal{R}$ and $\mathcal{P}$ mainly concern the expertise of E_r. On the other hand, the sets $\mathcal{M}$, $\mathcal{A}$ and $\mathcal{F}$ are related to the expertise of E_a.

The resolution of this problem implies a close collaboration and a permanent discussion between the two actors: the specialist E_a needs to know the preferences $\mathcal{P}$ and the objectives of the expert user E_r. These preferences can then be modelled and be used to build a family of criteria $\mathcal{F}$ to help in the selection of the *best* measure(s).

The following section presents the properties that have been retained to describe the different measures.

6. Evaluation criteria

In this section, we present a list of eligible properties to evaluate the previous list of measures.

For some of these properties, an order on the values they can take is straightforward. These properties can be considered as criteria by E_a (the analyst, expert in MCDA and KDD) without the intervention of E_r (the expert in the data). These properties, g_1, g_2, g_3, g_4 and g_7, will be called normative. In addition to these, the properties g_5, g_6 and g_8 need E_r to express his preferences on the values they can take (Lenca et al. (2004)).

Table 6 summarises the semantic and the modalities of the 8 properties and the results of the evaluations are presented in Table 7.

g_1: asymmetric processing of `A` and `B` (Freitas, 1999). Since the antecedent and the consequent of a rule may have very different significations, it is desirable to make a distinction between measures that evaluate rules $\texttt{A} \rightarrow \texttt{B}$ differently from rules $\texttt{B} \rightarrow \texttt{A}$ and those which do not. We note `sym` if the measure is symmetric, `asym` otherwise.

g_2: decrease with n_b (Piatetsky-Shapiro, 1991). Given n_{ab}, $n_{a\overline{b}}$ and $n_{\overline{a}\overline{b}}$, it is of interest to relate the interestingness of a rule to the size of `B`. In this situation, if the number of records verifying `B` but not `A` increases, the interestingness of the rule should decrease. We note `dec(`n_b`)` if the measure is a decreasing function with n_b, `no-dec(`n_b`)` otherwise.

g_3: reference situations, independence (Piatetsky-Shapiro, 1991). To avoid keeping rules that contain no information, it is necessary to eliminate the $\texttt{A} \rightarrow \texttt{B}$ rule

when A and B are independent, which means that the probability of obtaining B does not depend of the fact that A is true or not. A comfortable way of dealing with this is to require that a measure's value at independence should be constant (independent of the marginal frequencies). We note `cst` if the measure's value at independence is constant and `var` otherwise.

g_4**: reference situations, logical rule (Lenca et al., 2003a).** Similarly, the second reference situation we consider is related to the value of the measure when there is no counter-example. Depending on the co-domain (see Table 3), three cases arise. First, the measure takes a value independent of the marginal frequencies and thus takes a constant and maximal value[1]. A second case occurs when the measure tends to infinity when $n_{a\bar{b}} \to 0$. Finally, a third and more uncomfortable case arises when the value taken by the measure depends on the marginal frequencies when $n_{a\bar{b}} = 0$.

It is desirable that the value should be constant and maximal, or possibly infinite. We note `cst` in the cases of a constant or infinite value, `var` otherwise.

Independence is the lower value in which we are interested and we do not take into account the value for the incompatibility situation. The latter reference situation is obtained when $\mathtt{A} \cap \mathtt{B} = \emptyset$, and expresses the fact that B cannot be realized if A already has been. Our choice is based on the fact that incompatibility is related to the rule $\mathtt{A} \to \overline{\mathtt{B}}$ and not $\mathtt{A} \to \mathtt{B}$.

g_5**: linearity with** $p_{a\bar{b}}$ **around** 0^+ **(Gras et al., 2001).** It is desirable to have a slow decrease in the neighbourhood of a logical rule rather than a fast or even linear decrease (as with confidence or its linear transformations). This reflects the fact that the user may tolerate a few counter-examples without significant loss of interest, but will definitely not tolerate too many of them. However, the opposite choice could also be preferred. In that case, a convex decrease with $n_{a\bar{b}}$ around the logic rule increases the sensitivity to a false positive. We hence note `convex` if the measure is convex with $n_{a\bar{b}}$ near 0, `linear` if it is linear and `concave` if it is concave.

g_6**: sensitivity to** n **(total number of records) (Lallich, 2002; Gras et al., 2004).** Intuitively, if the rates of presence of A, $\mathtt{A} \to \mathtt{B}$, B are constant, it may be interesting to see how the measure reacts to a global extension of the database (with no evolution of rates). Measures that are sensitive to n are called statistical measures while those not sensitive are called descriptive measures.

The user can prefer to have a measure which is invariant or not with the dilatation of data. Note that, if the measure increases with n and has a maximum value, then there is a risk that all the evaluations might come close to this maximum. The measure would then lose its discrimination power. We note `stat` if it increases with n and `desc` if the measure is invariant.

g_7**: easiness to fix a threshold (Lenca et al., 2003a).** Even if properties g_3 and g_4 are valid, it is still difficult to decide on the best threshold value that separates interesting from uninteresting rules. This property allows us to identify measures whose threshold is more or less difficult to locate.

To establish this property, we propose to proceed in the following (and very conven-

[1]Recall that due to our eligibility criterion, we restrict our study to decreasing measures with respect to $n_{a\bar{b}}$, all marginal frequencies being fixed.

tional) way by providing a sense of the strength of the evidence against the null hypothesis[2] H_0 (absence of a link between A and B), that is the p-value[3]. Due to the high number of tests, this probability should not be interpreted as a statistical risk, but rather as a control parameter (Lallich and Teytaud, 2004).

In some cases, the measure is defined as the complement to 1 of such a probability, for example PDI or IntImp. It is then justified to set the threshold to at least 0.95 or 0.99. More generally, it is possible to set the threshold by the same way for all the measures whose distribution under the hypothesis of link absence (H_0) can be established. The threshold is the observed value of the measure for which the complementary p-value is at least 0.95 or 0.99. The distribution of $N_{a\bar{b}}$ (or N_{ab}) under H_0 can be established from one of the three types of models proposed by Lerman (1970). The margins n_a and n_b being fixed (hypergeometric model), it is possible to determine the distribution of the confidence under H_0. The same can be done for all the measures which are a monotone transformation of the confidence, the margins being fixed. For example, the expectation of the Lift under H_0 is 1 and its variance is approximately $(1/n)\frac{p_{\bar{a}}p_{\bar{b}}}{p_a p_b}$. Under the condition of normal approximation ($np_ap_b \geq 3$), the threshold of the Lift must be at least $1+1.645\sqrt{[(1/n)\frac{p_{\bar{a}}p_{\bar{b}}}{p_a p_b}]}$. The only measures which do not comply with this framework are Zhang (because of the max) and TEII (because of the inclusion index).

g_8**: intelligibility (Lenca et al., 2003a).** Intelligibility denotes the ability of the measure to express a comprehensive idea of the interestingness of a rule. We will consider that a measure is intelligible if its semantics is easily understandable by the expert in the data E_r[4].

We define this criterion according to three factors. First, the definition of the measure integrates only simple arithmetic operations on the frequencies. Second, the variations of the values taken by the measure are easily interpretable. And third, the definition of the measure is intelligible for the user.

We affect the value a to this property if the measure has the three preceding characteristics, b if the measure only verifies two of them, and c if it seems impossible to give any concrete explanation of the measure.

We evaluate the measures described in the previous section with respect to these criteria and we obtain the evaluation matrix of Table 7.

7. Evaluation of the interestingness measures

In this Section, we analyse and evaluate the measures described earlier and summarised in Table 2. This analysis has been done using a few MCDA procedures, in particular the Tomaso method for sorting (Marichal et al. (2005)), a ranking procedure based on kernels of digraphs by Bisdorff (1999) and the Promethee method (Brans and Vincke

[2]The null hypothesis is presumed true until statistical evidence in the form of a hypothesis test indicates otherwise. The alternative hypothesis is chosen if the observed data values are sufficiently improbable under the null hypothesis.

[3]The p-value is the probability of getting a value of the test statistic "at least as extreme" as that observed strictly by chance, given the assumption that the null hypothesis is true. The null hypothesis is rejected if the p-value is less than the significance level.

[4]It is obvious that this property is subjective. The evaluations of the measures on this property given hereafter can be commonly accepted. Nevertheless, depending on E_r, our evaluations could be revised.

Table 6
Properties of the measures

Property	Semantic	Modalities
g_1	asymmetric processing of `A` and `B`	`asym`, `sym`
g_2	decrease with n_b	`dec`(n_b), `no-dec`(n_b)
g_3	reference situations: independence	`cst`, `var`
g_4	reference situations: logical rule	`cst`, `var`
g_5	linearity with $n_{a\bar{b}}$ around 0^+	`convex`, `linear`, `concave`
g_6	sensitivity to n	`desc`, `stat`
g_7	easiness to fix a threshold	`easy`, `hard`
g_8	intelligibility	`a`, `b`, `c`

Table 7
Evaluation matrix

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
BF	asym	dec(n_b)	cst	cst	convex	desc	easy	a
CenConf	asym	dec(n_b)	cst	var	linear	desc	easy	a
Conf	asym	no-dec(n_b)	var	cst	linear	desc	easy	a
Conv	asym	dec(n_b)	cst	cst	convex	desc	easy	b
ECR	asym	no-dec(n_b)	var	cst	concave	desc	easy	b
IG	sym	dec(n_b)	cst	var	concave	desc	easy	c
- ImpInd	asym	dec(n_b)	cst	var	linear	stat	easy	c
IntImp	asym	dec(n_b)	cst	var	concave	stat	easy	c
Kappa	sym	dec(n_b)	cst	var	linear	desc	easy	c
Lap	asym	no-dec(n_b)	var	var	linear	desc	easy	c
LC	asym	dec(n_b)	var	var	linear	desc	easy	b
Lift	sym	dec(n_b)	cst	var	linear	desc	easy	a
Loe	asym	dec(n_b)	cst	cst	linear	desc	easy	b
PDI	asym	dec(n_b)	cst	var	concave	stat	easy	c
PS	sym	dec(n_b)	cst	var	linear	stat	easy	b
R	sym	dec(n_b)	cst	var	linear	desc	easy	b
Seb	asym	no-dec(n_b)	var	cst	convex	desc	easy	b
Sup	sym	no-dec(n_b)	var	var	linear	desc	easy	a
TEII	asym	dec(n_b)	cst	var	concave	stat	hard	c
Zhang	asym	dec(n_b)	cst	cst	concave	desc	hard	c

(1985)). These three methods produced very similar results. In this paper, we focus on the analysis by the PROMETHEE method to obtain a ranking. Its general objectives are to build partial and complete rankings on so-called alternatives (in this case, the measures) and to visualise the structure of the problem in a plane called the GAIA plane, similarly to a principal component analysis. The PROMETHEE method requires information about the importance of the criteria to be given by a set of weights. Several tools allow these weights to be fixed in order to represent the decision maker's preferences (E_r in our context). The first step of the method is to make pairwise comparisons on the measures within each criterion. This means that for small (large) deviations, E_r will allocate a small (large) preference to the best measure. This is done through the concept of preference functions. Then, each measure is confronted with the others in order to define outranking flows. The positive (negative) outranking flow expresses to what degree a measure outranks (is outranked by) all the others. Finally, partial and complete rankings are generated from these outrankings. The GAIA plane provides information about the conflicting character of the criteria and about the impact of the weights on the final decision. It is a projection, based on a net flow derived from the outranking flows, of the measures and the criteria in a common plane. The GAIA plane shows useful information and allows intuitive interaction with the decision maker. This is one of the reasons why we choose the PROMETHEE method.

For a better understanding of our future discourse, we briefly present the main concepts of the PROMETHEE method. For a more detailed description, the reader can refer to Brans and Mareschal (2002), Brans and Mareschal (2005), for example.

7.1. A glance at the PROMETHEE method

Let $A = \{a_1, \ldots, a_m\}$ be a set of possible alternatives. In the present discourse, the alternatives are the quality measures. Let $\{g_j(\cdot), j = 1, \ldots, k\}$ be a set of evaluation criteria to be maximised or minimised. Each of the possible alternatives of A is evaluated on each of the criteria.

Pairwise comparison

The method is based on pairwise comparisons of the alternatives. First, it formalises the degree of preference of one alternative over another for each criterion. For two alternatives a_i and a_j of A and for a criterion g_k, this is done by transforming the difference of the evaluations of a_i and a_j on g_k by a so-called preference function. The result is a degree of preference of a_i over a_j on g_k which is between 0 and 1.

Aggregated preference index

The next step involves aggregating the preference degrees on each criterion in an aggregated preference index $\pi(a_l, a_m)$ which expresses the degree to which a_l is preferred to a_m on the whole set of criteria. This is done by a weighted sum, where the weights are associated with the importance of the criteria.

Outranking flows

The aggregated preference index is used to build outranking flows for each alternative. The positive outranking flow expresses the overall power (its outranking character) of the

considered alternative, whereas the negative outranking flow gives an indication about its overall weakness (its outranked character). The net outranking flow ϕ is the difference between the positive and the negative outranking flows.

The rankings

A partial and a complete ranking are obtained on the basis of the positive and negative outranking flows.

The GAIA plane

The PROMETHEE method allows the alternatives and the criteria to be visualised in a common plane called the GAIA plane. This useful representation gives a synthetic clear view of the conflicting characteristics of certain criteria and of the impact of the weights on the final rankings. It is a projection of the data which is quite similar to what is done in principal components analysis. The alternatives are represented by points and the criteria by segments (or axes) in this plane. In addition, the GAIA plane contains a so-called decision axis π, which roughly indicates the direction of the best alternatives for a given weight system.

Let us point out a few features of the GAIA plane for a useful analysis of the problem:

- a long axis for a criterion in the GAIA plane stands for a discriminating criterion;
- criteria representing similar (opposite) preferences on the set of alternatives are represented by axes which have a similar direction (opposing directions);
- independent criteria are represented by orthogonal axes;
- alternatives which have *good* evaluations on a given criterion are represented by points which are close to the axis of this criterion;
- similar alternatives are close in the GAIA plane;
- if the π axis is long, it has a strong decision power, and the decision maker should choose alternatives which lie in the direction and the sense of the axis;
- if the π axis is short, it has a weak decision power. This means that for this configuration of weights, the criteria are conflicting, and a good compromise can be found at the origin of the plane.

Stability intervals

The PROMETHEE method allows stability intervals to be computed for the weights of the criteria. They indicate to what degree the value of a weight can be modified without modifying the complete ranking.

7.2. Analysis of the quality measures

This section focuses on the analysis of the selected quality measures by the MCDA procedure PROMETHEE. We consider the following two realistic scenarios for the analysis:

Sc1: The expert E_r tolerates *the appearance of a certain number of counter-examples* to a decision rule. In this case, the rejection of a rule is postponed until enough counter-examples are found. The shape of the curve representing the value of the measure versus the number of counter-examples should ideally be concave (at least in the neighbourhood of the maximum); the order on the values of criterion g_5 (non-linearity with respect to the number of counter-examples) is therefore `concave` $\succ$ `linear` $\succ$ `convex`.

Sc2: The expert E_r refuses *the appearance of too many counter-examples* to a decision rule. The rejection of the rule must be done rapidly with respect to the number of counter-examples. The shape of the curve is therefore ideally convex (in the neighbourhood of the maximum at least) and the order on the values of criterion g_5 is `convex` $\succ$ `linear` $\succ$ `concave`.

We recall that the evaluation matrix of the measures is presented in Table 7. For both scenarios, for criterion g_6 we suppose that the expert prefers a measure which increases with n, the size of the data. The order on the values of criterion g_6 is `stat` $\succ$ `desc`. We suppose that the expert agrees with the analysis on the intelligibility of the measures. Therefore the order on the values for g_8 is `a` $\succ$ `b` $\succ$ `c`.

For the other criteria which are assumed to be normative, the expert has no influence on the order of the values. The orders on the values for criteria g_1, g_2, g_3, g_4 and g_7 are given in Table 8.

Table 8
The order on the values of the normative criteria

criterion	order
g_1	`asym` $\succ$ `sym`
g_2	`dec`(n_b) $\succ$ `no-dec`(n_b)
g_3	`cst` $\succ$ `var`
g_4	`cst` $\succ$ `var`
g_7	`easy` $\succ$ `hard`

We start by analysing the problem with equal weights for the criteria in order to get a first idea about the structure of the problem. The total rankings for the two scenarios are given in Table 9.

First, we notice that both scenarios reflect the preferences of E_r concerning the shape of the curve. We can see that for **Sc1** the two leading measures are IntImp and PDI which are both concave. Similarly, for **Sc2**, the two leading measures are BF and Conv which are both convex. This is quite interesting because in both scenarios the weights of the criteria are all equal. This means that E_r has not expressed any particular preferences on the criteria. A small experiment shows that it is important to distinguish the two scenarios. If we give criterion g_5 a high weight (33%), the first positions of the ranking for **Sc1** (**Sc2**) are held by IntImp, PDI, Zhang, TEII and ECR (BF, Conv, Seb, and Loe) which are mostly concave (convex). This first analysis with equal weights also shows that the linear measure Loe is a very interesting measure as it is well-placed in both scenarios. It represents a good compromise.

Table 9
Total rankings for scenarios **Sc1** and **Sc2**.

Rank:	1	2	3	4	5	6	7
Sc1:	IntImp, PDI		Loe	BF	CenConf	Conv	-ImpInd
Sc2:	BF	Conv	Loe	CenConf	-ImpInd	PS	Seb
Rank:	8	9	10	11	12	13	14
Sc1:	Zhang, TEII		PS	ECR	Lift	Conf	IG
Sc2:	Lift	Conf	IntImp, PDI		r, LC		Zhang
Rank:	15	16	17	18	19	20	
Sc1:	r, LC		Seb	Kappa	Sup	Lap	
Sc2:	TEII	Kappa	ECR	Sup	IG	Lap	

Table 10
Net flows for **Sc1** and **Sc2**.

Sc1:	IntImp, PDI	Loe	BF	CenConf	Conv	-ImpInd	...	Lap
ϕ	.19	.18	.16	.13	.09	.08	...	-.32
Sc2:	BF	Conv	Loe	CenConf	-ImpInd	PS	...	Lap
ϕ	.39	.31	.22	.16	.12	.09	...	-.28

Sensitivity analyses on the weight systems show that small changes in the weights affect the rankings. Nevertheless, a closer look shows that these modifications only occur locally and that the first positions in the ranking remain stable. This is confirmed by the values of the net flows ϕ of the leading elements of each of the rankings presented in Table 10. This table shows that the $\phi(a), a \in \mathcal{M}$ are spread more or less uniformly between their minimum and their maximum values for both scenarios. In particular, we can see that the leading positions will vary only for very significative changes in the weight system. Therefore, one can say that for an expert who has no particular opinion on the importance of the different criteria, or who considers that the criteria are equally important, the rankings of Table 9 are valid.

An analysis of the Gaia planes gives us further indications about the measures (see Fig. 2 and Fig. 3).

Let us first note that the percentage of cumulated variance for the first two factors represented by the Gaia plane is 60.27% for both scenarios. The information taken from the Gaia plane should therefore be considered as approximate and conclusions be drawn with great care. First, we observe that the measures (triangles in the figures) are distributed homogeneously in the plane. Second, we can see that the Gaia plane is well covered by the set of criteria (axes with squares in the figures). We conclude that the description of the selected measures by the criteria is discriminant and only slightly redundant.

For **Sc1** we can see that several couples of criteria are independent: (g_4, g_5), (g_4, g_8), (g_5, g_3), (g_5, g_2), (g_8, g_3), (g_1, g_6) and (g_8, g_2)[5]. There are similar criteria, for example g_2 and g_3. We can also observe conflicting criteria. For example g_4 conflicts with g_3 and g_2; and criteria g_5 and g_6 conflict with g_7 and g_8. This type of information gives hints about the behaviour and the structure of the problem. For example, measures of $\mathcal{M}$ which are

[5]If g_i and g_j are independent, we write that the couple (g_i, g_j) is independent.

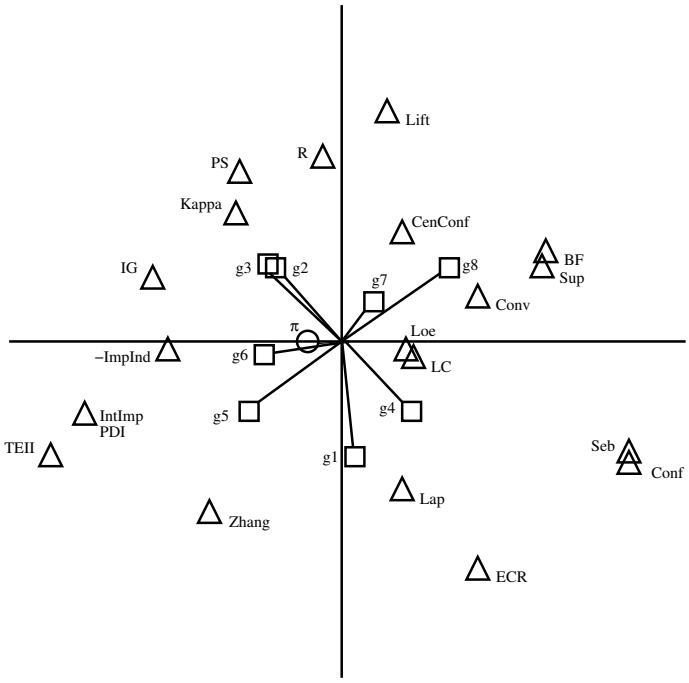


Figure 2. Gaia plane for scenario **Sc1**

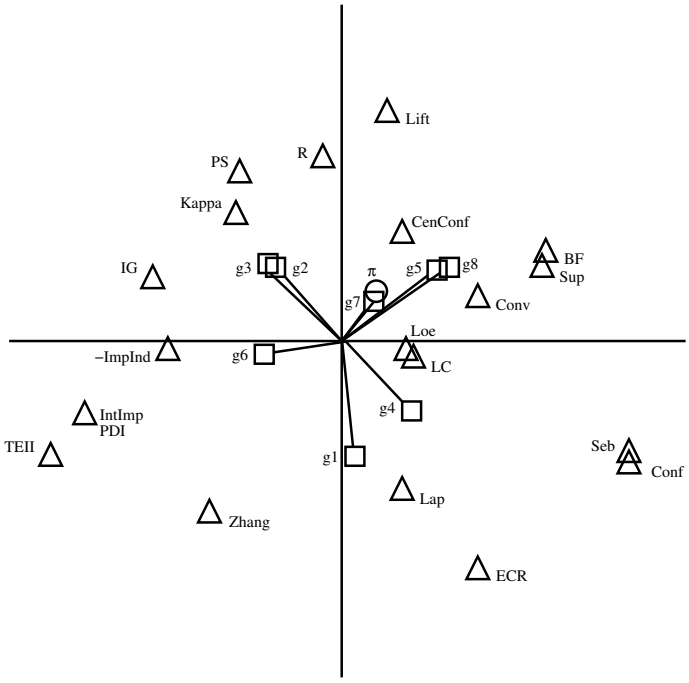


Figure 3. Gaia plane for scenario **Sc2**

good for criterion g_5 (concave) will tend to be bad for criterion g_8 (unintelligible).

For **Sc2** similar observations can be made. The major difference lies in criterion g_5 which represents similar preferences to criteria g_7 and g_8 but conflicts with g_6.

The decision axis π is moderately long in **Sc1** and heads in the opposite direction to g_7 and g_8. This means that measures which allow the threshold to be fixed easily and which are easily understandable (and which are quite bad on the remaining criteria) can appear in the leading positions of the ranking only if the relative weights of g_7 and g_8 are very high. However, we think that the importance of criterion g_3 (independence hypothesis) should not be neglected compared to a criterion like g_8 (intelligibility). Thus, if the expert is aware of the impact of his weight system on the result, we can suppose that a measure like Sup, exclusively good on g_7 and g_8, will never appear in the leading positions of the ranking.

For **Sc2** the decision axis π is also moderately long. It points in the opposite direction to g_7, g_5 and g_8. This partly explains the ranking of Table 9.

The positions of the measures in the Gaia plane (for **Sc1** and **Sc2**) show that many measures have similar behaviours with respect to weight variations. This is confirmed by their similar profiles in the evaluation matrix. Thus, Seb and Conf, are close in the Gaia plane and have similar profiles. IntImp and PDI have an equal profile and therefore have the same representation in the Gaia plane. These couples of measures will tend to appear in neighbouring (or equal) positions in the rankings. An important comment should be made at this point about the analysis of the Gaia plane. As it represents only a part of the information of the original cloud of points, each observation must be verified in the data or on the basis of other techniques. An erroneous conclusion would be to consider BF and Sup as similar measures due to their proximity in the Gaia plane. In fact, their profiles are very different and consequently their behaviour in case of weight variations will not be similar.

This quite detailed study of the problem shows the usefulness of an analysis by means of an mcda tool like Promethee. On the basis of the previously made observations we can suggest two strategies.

The first strategy involves checking first that the expert E_r has well understood the meaning of each of the criteria and their influence on the final result. Then, by means of a set of questions, he must express the relative importance of the weights of each criterion. Criteria like g_3, g_4 and g_7 will necessarily have high relative weights to guarantee a certain coherence. Indeed a measure which does not have fixed values at independence and in the situation of a logical rule and, what is more, a threshold which is hard to fix is quite useless in an efficient search for interesting rules. According to the preferences of the expert the relative importance of criteria like g_1 and g_8 can vary. The analysis should be started by using an initial set of weights coherent with these considerations. The stability of the resulting ranking should then be analysed, especially for the leading positions. If a stable ranking is obtained, the Gaia plane, the value of the net flows and the profile visualisation tool allow a finer analysis of the leading measures. The values of the net flows give a hint about the *distance* between two measures in the ranking. Two measures with similar values for the flows can be considered as similar.

The second strategy involves in a first step in an exploration of the Gaia plane. This procedure helps the expert to understand the structure of the problem and detect similar

and different measures. Furthermore, the visualisation of the criteria in the same plane as the measures allows us to detect the influence of the modification of the weights on the final ranking. This exploratory strategy should be applied with an expert E_r who has a priori knowledge about certain measures. He will be able to determine his ranking on the importance of the criteria by detecting some well-known measures in the GAIA plane. By using this first approximate weight system, the first strategy can be applied. An a posteriori validation can be done by determining the positions of the well-known measures in the final ranking.

To show the utility and the usefulness of the method, we finish this section by a small simulation of the behaviour of an expert E_r. We suppose that E_r is searching for a measure which can be easily used. Thus, he would like the measure to be easily understandable and the thresholds to be constant (i.e. independent of the marginal frequencies). Ideally the shape of the selected measure would be convex. The weight system he suggests is given as follows: g_1 (10%), g_2 (5%), g_3 (15%), g_4 (15%), g_5 (10%), g_6 (5%), g_7 (15%) and g_8 (25%). The leading positions of the complete ranking are given in Table 11.

Table 11
A simulation: Ranking and net flow for the preferences of E_r

1	2	3	4	5	...
BF (.48)	CONV (.33)	CENCONF, LOE (.25)		CONF (.20)	...

We can clearly see that BF is the best measure for this weight system. It is an easily interpretable measure which is E_r's main objective. A stability analysis shows that the leading positions remain stable for variations in the weight system. The remaining desires of E_r are also satisfied. Indeed the measure is good on g_3, g_4, g_7 and g_8. In addition, BF is also competitive on g_1 and g_2. Its weakness is its sensitivity to n, but this criterion g_6 is not normative.

8. Conclusion

In this article, we have proposed an initial array of 20 eligible measures evaluated on 8 properties which cover a large range of potential users' preferences. Given this array, we have shown how to use an MCDA method, and help expert users to choose an adapted interestingness measure in the context of association rules. Our approach is a first step to improving the quality of a set of rules that will effectively be presented to the user. Of course several other factors could be used, like attribute costs and misclassification costs (Freitas, 1999), cognitive constraints (Le Saux et al., 2002), etc.

In addition to the interest of having such a list of evaluation criteria for a large number of measures, the use of the PROMETHEE method has confirmed the fact that the expert's preferences have some influence on the ordering of the interestingness measures, and that there are similarities between different measures. Moreover, the PROMETHEE method allows us to make a better analysis of the user's preferences (the GAIA plane makes it easy to identify different clusters of criteria and measures).

As already mentioned, we have shown here how an MCDA method can be used to help select an appropriate measure for filtering of a set of association rules. Another possibility would be to aggregate the outputs of all (or a subset of) the quality measures in order to obtain a global evaluation of a rule which takes into account the various properties of the measures. Such an approach is currently being explored and has already been applied in (Barthélemy et al., 2006).

Acknowledgments

Benoît Vaillant would like to thank the BMO (*Brest Métropole Océane*, the Urban Community of Brest) for financial support of his Ph.D. thesis. The authors would like to thank the members of the CNRS group GafoQualité for productive discussions about *interestingness measures*. Finally, the authors greatly appreciate the assistance of the referees and the editor. Their comments have improved both the content and the presentation of the original paper.

References

Agrawal, R., Imielinski, T., Swami, A., 1993. Mining association rules between sets of items in large databases. In: Buneman, P., Jajodia, S. (Eds.), Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. Washington, D.C., pp. 207–216.

Azé, J., Kodratoff, Y., 2002. Evaluation de la résistance au bruit de quelques mesures d'extraction de règles d'assocation. Extraction des connaissances et apprentissage (EGC 2002) 1 (4), 143–154.

Barthélemy, J. P., Legrain, A., Lenca, P., Vaillant, B., 2006. Aggregation of valued relations applied to association rule interestingness measures. In: Modeling Decisions for Artificial Intelligence. Lecture Notes in Artificial Intelligence. Springer-Verlag, pp. 203–214.

Bayardo, R. J., Agrawal, R., august 1999. Mining the most interesting rules. In: KDD 1999, Proceedings ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 145–154.

Bertin, J., 1977. La graphique et le traitement graphique de l'information. Flammarion.

Bisdorff, R., 1999. Bipolar ranking from pairwise fuzzy outrankings. Belgian Journal of Operations Research, Statistics and Computer Science 37 (4) 97, 379–387.

Blanchard, J., Guillet, F., Briand, H., Gras, R., May 2005. Assessing the interestingness of rules with a probabilistic measure of deviation from equilibrium. In: The XIth International Symposium on Applied Stochastic Models and Data Analysis. Brest, France, pp. 191–200.

Borgelt, C., Kruse, R., 2002. Induction of association rules: Apriori implementation. In: Proceedings of the 15th Conference on Computational Statistics. Physika Verlag, Heidelberg, Germany.

Brans, J., Mareschal, B., 2002. PROMETHEE-GAIA – Une méthode d'aide à la décision en présence de critères multiples. Ellipses.

Brans, J., Vincke, P., 1985. A preference ranking organization method. Management Science 31 (6), 647–656.

Brans, J.-P., Mareschal, B., 2005. Multiple criteria decision analysis: state of the art surveys. Springer – International series in operations research and management science. Figueira, J. and Greco, S. and Ehrgott, M. eds., Ch. PROMETHEE Methods, pp. 163–195.

Brijs, T., Vanhoof, K., Wets, G., 2003. Defining interestingness for association rules. International journal of information theories and applications 10 (4), 370–376.

Brin, S., Motwani, R., Silverstein, C., 1997a. Beyond market baskets: generalizing association rules to correlations. In: ACM SIGMOD/PODS '97 Joint Conference. pp. 265–276.

Brin, S., Motwani, R., Ullman, J. D., Tsur, S., 05 1997b. Dynamic itemset counting and implication rules for market basket data. In: Peckham, J. (Ed.), SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, Tucson, Arizona, USA. ACM Press, pp. 255–264.

Chauchat, J.-H., Risson, A., 1998. Visualization of Categorical Data. Blasius J. & Greenacre M. ed., Ch. 3, pp. 37–45, new York : Academic Press.

Church, K. W., Hanks, P., 1990. Word association norms, mutual information an lexicography. Computational Linguistics 16 (1), 22–29.

Cohen, J., 1960. A coefficient of agreement for nominal scale. Educational and Psychological Measurement 20, 37–46.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.), 1996. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press.

Francisci, D., Brisson, L., Collard, M., Mai 2003. Extraction de règles selon des critères multiples : l'art du compromis. Tech. Rep. ISRN I3S/RR-2003-11-FR, Université de Nice.

Freitas, A., 1999. On rule interestingness measures. Knowledge-Based Systems journal, 309–315.

Giakoumakis, V., Monjardet, B., 1987. Coefficients d'accord entre deux préordres totaux. Statistique et Analyse des Données 12 (1 et 2), 46–99.

Good, I. J., 1965. The estimation of probabilities: An essay on modern bayesian methods. The MIT Press, Cambridge, MA.

Gras, R., Ag. Almouloud, S., Bailleuil, M., Larher, A., Polo, M., Ratsimba-Rajohn, H., Totohasina, A., 1996. L'implication Statistique, Nouvelle Méthode Exploratoire de Données. Application à la Didactique, Travaux et Thèses. La Pensée Sauvage.

Gras, R., Couturier, R., Blanchard, J., Briand, H., Kuntz, P., Peter, P., 2004. Quelques critères pour une mesure de qualité de règles d'association - un exemple : l'intensité d'implication. Revue des Nouvelles Technologies de l'Information (Mesures de Qualité pour la Fouille de Données) (RNTI-E-1), 3–31.

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

Greco, S., Pawlak, Z., Slowinski, R., 2004. Can bayesian confirmation measures be useful for rough set decision rules? Engineering Applications of Artificial Intelligence 17 (4), 345–361.

Hajek, P., Havel, I., Chytil, M., 1966. The GUHA method of automatic hypotheses determination. Computing (1), 293–308.

Hilderman, R., Hamilton, H., 2003. Measuring the interestingness of discovered knowledge: A principled approach. Intelligent Data Analysis 7 (4), 347–382.

Hilderman, R. J., Hamilton, H. J., 2001. Evaluation of interestingness measures for ranking discovered knowledge. Lecture Notes in Computer Science 2035, 247–259.

Jeffreys, H., 1935. Some tests of significance treated by the theory of probability. In: Proceedings of the Cambridge Philosophical Society. No. 31. pp. 203–222.

Kamber, M., Shingal, R., August 1996. Evaluating the interestingness of characteristic rules. In: Proceedings of the Second International Conference on Knowledge Discovery and Dta Mining (KDD96). Portland, Oregon, pp. 263–266.

Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., Verkamo, A. I., 1994. Finding interesting rules from large sets of discovered association rules. In: Adam, N. R., Bhargava, B. K., Yesha, Y. (Eds.), Third International Conference on Information and Knowledge Management (CIKM'94). ACM Press, pp. 401–407.

Lallich, S., 2002. Mesure et validation en extraction des connaissances à partir des données. Habilitation à Diriger des Recherches – Université Lyon 2.

Lallich, S., Teytaud, O., 2004. Évaluation et validation de l'intérêt des règles d'association. Revue des Nouvelles Technologies de l'Information (Mesures de Qualité pour la Fouille de Données) (RNTI-E-1), 193–217.

Lallich, S., Vaillant, B., Lenca, P., May 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.

Le Saux, E., Lenca, P., Picouet, P., 2002. Dynamic adaptation of rules bases under cognitive constraints. European Journal of Operational Research 136 (2), 299–309.

Lenca, P., Lallich, S., Vaillant, B., June 7-9 2006. On the robustness of association rules. In: The IEEE International Conference on Cybernetics and Intelligent Systems. Bangkok, Thailand, pp. 596–601.

Lenca, P., Meyer, P., Picouet, P., Vaillant, B., Lallich, S., 2003a. Critères d'évaluation des mesures de qualité en ECD. Revue des Nouvelles Technologies de l'Information (Entreposage et Fouille de données) (1), 123–134.

Lenca, P., Meyer, P., Vaillant, B., Picouet, P., 2003b. Aide multicritère à la décision pour évaluer les indices de qualité des connaissances – modélisation des préférences de l'utilisateur. RSTI-RIA (EGC 2003) 1 (17), 271–282.

Lenca, P., Meyer, P., Vaillant, B., Picouet, P., Lallich, S., 2004. Évaluation et analyse multicritère des mesures de qualité des règles d'association. Revue des Nouvelles Technologies de l'Information (Mesures de Qualité pour la Fouille de Données) (RNTI-E-1), 219–246.

Lerman, I., 1970. Classification et analyse ordinale des données. Dunod.

Lerman, I., Azé, J., 2003. Une mesure probabiliste contextuelle discriminante de qualité des règles d'association. RSTI-RIA (EGC 2003) 1 (17), 247–262.

Lerman, I., Gras, R., Rostam, H., 1981. Elaboration d'un indice d'implication pour les données binaires, i et ii. Mathématiques et Sciences Humaines (74, 75), 5–35, 5–47.

Lim, T., Loh, W., Shih, Y., 2000. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Machine Learning 40, 203–228.

Lingoes, J., 1979. Geometric Representations of Relational Data. Mathesis Press, Ch. Indices of configural similarity, pp. 675–679.

Liu, B., Hsu, W., Chen, S., 1997. Using general impressions to analyze discovered classification rules. In: Third International Conference on Knowledge Discovery and Data Mining. pp. 31–36.

Liu, B., Hsu, W., Chen, S., Ma, Y., 2000. Analyzing the subjective interestingness of association rules. IEEE Intelligent Systems 15 (5), 47–55.

Loevinger, J., 1947. A systemic approach to the construction and evaluation of tests of ability. Psychological monographs 61 (4).

Marichal, J.-L., Meyer, P., Roubens, M., 2005. Sorting multi-attribute alternatives: The TOMASO method. Computers & Operations Research (32), 861–877.

McGarry, K., 2005. A survey of interestingness measures for knowledge discovery. Knowledge Engineering Review Journal 20 (1), 39–61.

Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L., 1999. Discovering frequent closed itemsets for association rules. In: Beeri, C., Buneman, P. (Eds.), Database Theory, 7th International Conference. Vol. 1540 of Lecture Notes in Computer Science. Springer, pp. 398–416.

Pearson, K., 1896. Mathematical contributions to the theory of evolution. iii. regression, heredity and panmixia. Philosophical Transactions of the Royal Society A.

Piatetsky-Shapiro, G., 1991. Discovery, analysis and presentation of strong rules. In: Piatetsky-Shapiro, G., Frawley, W. (Eds.), Knowledge Discovery in Databases. AAAI/MIT Press, pp. 229–248.

Rauch, J., Simunek, M., 2001. Mining for 4ft association rules by 4ft-miner. In: Proceeding of the International Conference On Applications of Prolog. pp. 285–294.

Roy, B., 1996. Multicriteria Methodology for Decision Aiding. Kluwer Academic Publishers.

Roy, B., Bouyssou, D., 1993. Aide Multicritère à la Décision : méthodes et cas. Economica, Paris.

Sebag, M., Schoenauer, M., 1988. Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases. In: Boose, J., Gaines, B., Linster, M. (Eds.), Proc. of the European Knowledge Acquisition Workshop (EKAW'88). Gesellschaft für Mathematik und Datenverarbeitung mbH, pp. 28–1 – 28–20.

Silberschatz, A., Tuzhilin, A., 1995. On subjective measures of interestingness in knowledge discovery. In: Knowledge Discovery and Data Mining. pp. 275–281.

Suzuki, E., 2006. Data mining methods for discovering interesting exceptions from an unsupervised table. Journal of Universal Computer Science 12 (6), 627–653.

Tan, P., Kumar, V., 2000. Interestingness measures for association patterns: A perspective. Tech. Rep. TR00-036, University of Minnesota, Department of Computer Science.

Tan, P.-N., Kumar, V., Srivastava, J., 2002. Selecting the right interestingness measure for association patterns. In: Proceedings of the Eighth ACM SIGKDD International Conference on KDD. pp. 32–41.

Vaillant, B., Lenca, P., Lallich, S., 2004. A clustering of interestingness measures. In: Discovery Science. Vol. 3245 of Lecture Notes in Artificial Intelligence. Springer-Verlag, pp. 290–297.

Vaillant, B., Picouet, P., Lenca, P., Mai 2003. An extensible platform for rule quality measure benchmarking. In: Bisdorff, R. (Ed.), Human Centered Processes (HCP'2003). Luxembourg, pp. 187–191.

Zhang, T., April 2000. Association rules. In: Terano, T., Liu, H., Chen, A. L. P. (Eds.), Knowledge Discovery and Data Mining, Current Issues and New Applications, 4th Pacific-Asia Conference, PADKK 2000, Kyoto, Japan, Proceedings. Vol. 1805 of Lecture Notes in Computer Science. Springer.