A multicriteria decision aid for interestingness measure selection

Philippe Lenca, Patrick Meyer, Benoît Vaillant, Stéphane Lallich

Département Logiques des Usages, Sciences Sociales
et Sciences de l'Information

Laboratoire Traitement Algorithmique et Matériel de la
Communication, de l'Information et de la Connaissance
CNRS FRE 2658

A Multicriteria Decision Aid for Interestingness Measure Selection

Philippe Lenca*, Patrick Meyer**,
Benoît Vaillant*, Stéphane Lallich***

*: GET ENST Bretagne / Département LUSSI – CNRS TAMCIC, France
philippe.lenca@enst-bretagne.fr, benoit.vaillant@enst-bretagne.fr

**: Faculty of Law, Economy and Finance,
University of Luxembourg, Luxembourg
patrick.meyer@letzebuerg.info

***: Université Lumière, Lyon 2 – Laboratoire ERIC, France
stephane.lallich@univ-lyon2.fr

LUSSI-TR-2004-01-EN
May 2004

GET ENST Bretagne – Département LUSSI
Technopôle de Brest Iroise CS 83818 – 29238 BREST Cedex, France

Contents

A Multicriteria Decision Aid for Interestingness Measure Selection

Philippe Lenca*, Patrick Meyer**,
Benoît Vaillant*, Stéphane Lallich***

*: GET ENST Bretagne / Département LUSSI – CNRS TAMCIC, France
**: Faculty of Law, Economy and Finance,
University of Luxembourg, Luxembourg
***: Université Lumière, Lyon 2 – Laboratoire ERIC, France

Abstract

Datamining algorithms, especially those used for unsupervised learning, generate a large quantity of rules. It is hence impossible for an expert in the field being mined to sustain these rules. To help carrying out the task, many rule interestingness measures have been developed in order to filter and sort automatically a set of rules with respect to given goals. Since measures give different results, and as experts have different understandings of what is a *good* rule, we propose in this article a new direction to select the *best* rules: a two-step solution to the choice problem of a user-adapted interestingness measure. First, a characterization of interestingness measures, based on meaningful classical properties, is provided. Second, a multicriteria decision aiding process is applied on this characterization and illustrates the benefit that a user, who is not a datamining expert, can achieve with such methods.
Keywords: interestingness measures, association rules, decision aiding.

A Multicriteria Decision Aid for Interestingness Measure Selection

Philippe Lenca*, Patrick Meyer**,
Benoît Vaillant*, Stéphane Lallich***

*: GET ENST Bretagne / Département LUSSI – CNRS TAMCIC, France
**: Faculty of Law, Economy and Finance,
University of Luxembourg, Luxembourg
***: Université Lumière, Lyon 2 – Laboratoire ERIC, France

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 supposed to be a data mining expert, but rather an expert of 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]. Ideally, a rule should be *valid, new and comprehensive* [Fayyad et al., 1996] but these generic terms cover a large number of various situations when examined in a precise context. It is also well known that data mining algorithms may produce huge amounts of rules and that the end user is then unable to analyze them manually [Hilderman and Hamilton, 2001a].

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 goals, data mining experts may choose the required interestingness measure, but this task cannot be taken care of by the expert user if left on his own. Indeed, this choice is hard since rule interestingness measures have many different qualities or flaws [Tan et al., 2002]. What is more, some of these properties are incompatible [Lenca et al., 2003]. Therefore there is no *optimal* measure, and a way of solving this problem is to try to find good compromises [Vaillant et al., 2004]. A well-known example of such a controversial measure is the support. On the one hand, it is heavily used for filtering purposes in KDD algorithms [Agrawal et al., 1993; Pasquier et al., 1999], for 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; Picouet and Lenca, 2001]. Bayardo and Agrawal [1999], Tan and Kumar [2000], Hilderman and Hamilton [2001b], Lallich and Teytaud [2004], for instance, have formally extracted several specificities of measures.

The importance of objective evaluation criteria of interestingness measures has already been focused on by Piatetsky-Shapiro [1991] and Freitas [1999]. However, the relevant aspects of these criteria to help the user to choose the right measure is still difficult to establish. In Tan et al. [2002], the authors provide a comparative study according to properties and provide an original approach to measure selection by an expert. However, this approach does not exploit the above-mentioned comparative table: from the set of rules resulting from a data mining algorithm, authors propose to extract a small subset of rules where measures give very different results. The authors experimentally establish that the diversity of results on the rule subset enable the user to efficiently select an adapted measure.

Our article might be seen as an alternative contribution to Tan et al. [2002]. We promote a two-step process. First, we provide a comparative description of measures thanks to a list of properties that are partially different from the properties evaluated in Tan et al. [2002], since some of them do not apply efficiently to association rule interestingness, or others do not make any distinction between the different kinds of interestingness measures we studied. In addition, we introduce and study new properties, such as the easiness to fix a threshold, for filtering purposes. We think that users are quite interested in this property. Second, we propose to use a MultiCriteria Decision Aiding (MCDA) method on the previous identified set of properties. MCDA [Roy, 1996] methods have already prooved their utility in different fields. We argue in this paper that an MCDA method could be profitable for the specific problem of a user's choice of a measure.

This paper is organized as follows. In section 2 we briefly remind the context of association 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 measure evaluation. 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. Finally, we conclude in section 8.

Table 1: Dataset and large itemsets of size 1

Dataset:

a	b	c	d	e
1	0	1	1	0
0	1	1	0	1
1	1	1	0	1
0	1	0	0	1

1-itemset	support
a	2
b	3
c	3
d	1
e	3

1-large itemset	support
a	2
b	3
c	3
e	3

2 Association rules mining

Given a database of transactions where each transaction is a list of items the problem of mining association rules consist of discovering all rules that correlate the presence of one set of items (or itemset) with that of another set of items under minimum support and minimum confidence conditions :

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

The well know APRIORI algorithm [Agrawal et al., 1993] proceedes in two steps within the support-confidence framework (minimum support and confidence thresholds have to be fixed by the user) :

- find frequent itemsets (the sets of items which occur above 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 $k + 1$ form frequent itemsets of size k
- generate rules from frequent itemsets and filter them with the minimum confidence threshold

Table 1 contains a small dataset of four transactions and illustrates the determination of the 1-large itemsets with a minimum support of 2 (*i.e.* 50%). Tables 2 and 3 show the determination of the 2 and 3-large itemsets. Then table 4 shows the following rules with minimum support 50% and minimum confidence 60%.

Table 2: Large itemsets of size 2

2-itemset
a, b
a, c
a, e
b, c
b, e
c, e

2-large itemset	support
a, c	2
b, c	2
b, e	3
c, e	2

Table 3: Large itemsets of size 3

3-itemset
b, c, e

3-large itemset	support
b, c, e	2

Unfortunately APRIORI within the support and confidence framework tends to generate a large amount 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 method relies on the use of objective and subjective interestingness measures.

Strong rules (interesting rules within support and confidence framework) satisfy the minimum support and minimum confidence thresholds. Still, they are not necessarily interesting neither from an expert's point of view nor from a statistical one. For example, consider the well-known data: of the 10 000 transactions, 6 000 include computer games, while 7 500 include videos, and 4 000 include both computer games and videos. Let the minimum support be 30% and the minimum confidence be 60%. Thus the strong rule *buy computer games* $\rightarrow$ *buy videos* is discovered with support of 40% and confidence 66%. However this (strong) rule is misleading since the probability of purchasing videos is 75%.

Depending on the user's goals, data mining experts should apply the required interestingness measures: they have many different and contradictory qualities or flaws. Moreover, they may generate different rankings out of a given set of rules, and hence highlight different pieces of information. In the next section we present 20 measures that we studied.

3 Selected interestingness measures

The 20 measures we here list evaluate the interestingness of association rules as defined in Agrawal et al. [1993]: given a typical market-basket (transactional) database $\mathtt{E}$, the association rule $\mathtt{A} \rightarrow \mathtt{B}$ means *if someone buys the set of items*

Table 4: Rules with minimum support 50% and minimum confidence 60%

rule	support	confidence	rule	support	confidence
a → c	50.0%	100.0%	b → e	75.0%	100.0%
c → a	50.0%	66.7%	e → b	75.0%	100.0%
c → b	50.0%	66.7%	cb → e	50.0%	100.0%
b → c	50.0%	66.7%	ce → b	50.0%	100.0%
c → e	50.0%	66.7%	be → c	50.0%	66.7%
e → c	50.0%	66.7%			

A, *then he/she probably also buys item* B. It is very important to differentiate between the association rule $\mathtt{A} \rightarrow \mathtt{B}$, which focuses on cooccurence and gives asymmetric meaning to A and B, from logical implication $\mathtt{A} \Rightarrow \mathtt{B}$ or equivalence $\mathtt{A} \Leftrightarrow \mathtt{B}$ (see Lallich and Teytaud [2004]).

A\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

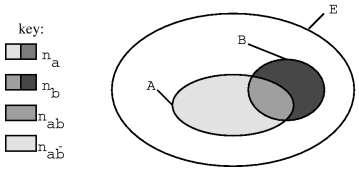


Figure 1: Notations

These measures are usually defined using frequency counts presented in figure 1. 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 A
- $n_b = |\mathtt{B}|$ the number of records satisfying B
- $n_{ab} = |\mathtt{A} \cap \mathtt{B}|$ the number of records satisfying both A and B (the examples of the rule)
- $n_{a\bar{b}} = |\mathtt{A} \cap \bar{\mathtt{B}}|$ the number of records satisfying A but not 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 in figure 1 is enough to deduce the other ones.

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 (A true and B false)

to the rule there are, the higher the interestingness of the rule is. For a given decreasing monotonic measure μ, it is then possible to select interesting rules by positioning 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 datamining 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].

Some measures like χ^2, Pearson's r^2, Goodman and Smyth's J-measure or Pearl's measure were excluded from this list, since they are not monotonically decreasing. From previous works, we selected a reasonable set of interestingness measures. They are listed in table 6. Bibliographical references are given in table 5.

Table 5: List of selected measures

	Name	References
SUP	support	Agrawal et al. [1993]
CONF	confidence	Agrawal et al. [1993]
R	Pearson's correlation coefficient	Pearson [1896]
CENCONF	centered confidence	
PS	Piatetsky-Shapiro	Piatetsky-Shapiro [1991]
LOE	Loevinger	Loevinger [1947]
ZHANG	Zhang	Terano et al. [2000]
- IMPIND	implication index	Lerman et al. [1981]
LIFT	Lift	Brin et al. [1997a]
LC	least contradiction	Azé and Kodratoff [2002]
SEB	Sebag and Schoenauer	Sebag and Schoenauer [1988]
OM	odd multiplier	Lallich and Teytaud [2004]
CONV	conviction	Brin et al. [1997b]
ECR	examples and counter-examples rate	
KAPPA	Kappa coefficient	Cohen [1960]
IG	information gain	Church and Hanks [1990]
INTIMP	intensity of implication	Gras et al. [1996]
EII	entropic intensity of implication	Gras et al. [2001]
PDI	probabilistic discriminant index	Lerman and Azé [2003]
LAP	Laplace	Good [1965]

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.

Many other measures are linear transformations of the confidence, enhancing it, as they enable comparisons with p_b. This transformation is generally achieved by centering of confidence on p_b, using different scale coefficients (centered confidence, Piatetsky-Shapiro's measure, Loevinger's measure, Zhang's measure, correlation, implication index, least contradiction). It is also possible to divide the confidence by p_b (lift).

Other measures, like Sebag-Schoenauer's or the examples and counter-examples

Table 6: List of selected measures

	Absolute definitions	Relative definitions
SUP	$\frac{n_a - n_{a\bar{b}}}{n}$	p_{ab}
CONF	$1 - \frac{n_{a\bar{b}}}{n_a}$	$p_{b/a}$
R	$\frac{nn_{ab} - n_a n_b}{\sqrt{nn_a n_b n_{\bar{a}} \cdot n_{\bar{b}}}}$	$\frac{p_{ab} - p_a p_b}{\sqrt{p_a p_{\bar{a}} p_b p_{\bar{b}}}}$
CENCONF	$\frac{nn_{ab} - n_a n_b}{nn_a}$	$p_{b/a} - p_b$
PS	$\frac{1}{n}(\frac{n_a n_{\bar{b}}}{n} - n_{a\bar{b}})$	$np_a\left(p_{b/a} - p_b\right) = np_a p_b\,(\text{Lift} - 1)$
LOE	$1 - \frac{nn_{a\bar{b}}}{n_a n_{\bar{b}}}$	$\frac{p_{b/a} - p_b}{p_{\bar{b}}} = \frac{1}{p_{\bar{b}}}\text{CenConf} = 1 - \frac{1}{\text{Conv}}$
ZHANG	$\frac{nn_{ab} - n_a n_b}{\max\{n_{ab}n_{\bar{b}}, n_b n_{a\bar{b}}\}}$	$\frac{p_{ab} - p_a p_b}{\max\{p_{ab}p_{\bar{b}}; p_b(p_a - p_{ab}\}}$
- IMPIND	$\frac{nn_{a\bar{b}} - n_a n_{\bar{b}}}{\sqrt{nn_a n_{\bar{b}}}}$	$\sqrt{n}\frac{p_{a\bar{b}} - p_a p_{\bar{b}}}{\sqrt{p_a p_{\bar{b}}}}$
LIFT	$\frac{nn_{ab}}{n_a n_b}$	$\frac{p_{ab}}{p_a p_b}$
LC	$\frac{n_{ab} - n_{a\bar{b}}}{n_b}$	$\frac{p_{ab} - p_{a\bar{b}}}{p_b} = 2\frac{p_a}{p_b}\,(\text{Conf} - 0.5)$
SEB	$\frac{n_a - n_{a\bar{b}}}{n_{a\bar{b}}}$	$\frac{p_{ab}}{p_{a\bar{b}}} = \frac{\text{Conf}}{1-\text{Conf}}$
OM	$\frac{(n_a - n_{a\bar{b}})n_{\bar{b}}}{n_b n_{a\bar{b}}}$	$\frac{\frac{p_{b/a}}{p_{\bar{b}/a}}}{\frac{p_b}{p_{\bar{b}}}} = \frac{p_{ab}}{p_b}\frac{p_{\bar{b}}}{p_{a\bar{b}}} = \text{Lift} \cdot \text{Conv}$
CONV	$\frac{n_a n_{\bar{b}}}{nn_{a\bar{b}}}$	$\frac{p_a p_{\bar{b}}}{p_{a\bar{b}}}$
ECR	$\frac{n_a - 2n_{a\bar{b}}}{n_a - n_{a\bar{b}}} = 1 - \frac{1}{\frac{n_a}{n_{a\bar{b}}} - 1}$	$1 - \frac{p_{a\bar{b}}}{p_{ab}} = 1 - \frac{1}{Seb}$
KAPPA	$2\frac{nn_a - nn_{a\bar{b}} - n_a n_b}{nn_a + nn_b - 2n_a n_b}$	$2\frac{p_{ab} - p_a p_b}{p_a + p_b - 2p_a p_b}$
IG	$\log(\frac{nn_{ab}}{n_a n_b})$	$\log\frac{p_{ab}}{p_a p_b} = \log(\text{Lift})$
INTIMP	$\varphi = P\Big[poisson(\frac{n_a n_{\bar{b}}}{n}) \geq n_{a\bar{b}}\Big]$	$P\left[Poisson\left(np_a p_{\bar{b}}\right) \geq np_{a\bar{b}}\right]$
EII	$\left\{[(1 - h_1(\frac{n_{a\bar{b}}}{n})^2)(1 - h_2(\frac{n_{a\bar{b}}}{n})^2)]^{1/4}\varphi\right\}^{1/2}$	$\left\{[(1 - h_1(p_{ab})^2)(1 - h_2(p_{ab})^2)]^{1/4}\varphi\right\}^{1/2}$
PDI	$P\Big[\mathcal{N}(0,1) > \text{IMPIND}^{RC/\mathcal{B}}\Big]$	
LAP	$\frac{n_{ab}+1}{n_a+2}$	$\frac{p_{b/a} + \frac{n}{p_a}}{1+\frac{2n}{p_a}}$

- $h_1(t) = -(1 - \frac{n\cdot t}{n_a})\log_2(1 - \frac{n\cdot t}{n_a}) - \frac{n\cdot t}{n_a}\log_2(\frac{n\cdot t}{n_a})$ if $t \in [0, n_a/2\,n[$; else $h_1(t) = 1$
- $h_2(t) = -(1 - \frac{n\cdot t}{n_{\bar{b}}})\log_2(1 - \frac{n\cdot t}{n_{\bar{b}}}) - \frac{n\cdot t}{n_{\bar{b}}}\log_2(\frac{n\cdot t}{n_{\bar{b}}})$ if $t \in [0, n_{\bar{b}}/2\,n[$; else $h_2(t) = 1$
- *poisson* stands for the Poisson distribution
- $\mathcal{N}(0,1)$ stands for the centred and reduced normal repartition function
- IMPIND $^{CR/\mathcal{B}}$ corresponds to IMPIND, centred reduced (CR) for a rule set $\mathcal{B}$

rate, are monotonically increasing transformations of confidence, while the information gain is a monotonically increasing transformation of the lift. Some measures focus on counter examples, like the conviction or the above cited implication index. 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. Finally, the odd multiplier is a kind of odd-ratio, based on the comparison of the odd of A and B on B rather than the odd of A and $\overline{\text{A}}$ on B, and Laplace's measure is a variant of the confidence, taking the total number of records n into account.

4 Total pre-order comparison

In order to get an idea of the difficulty of selecting the subset of the n best rules, we studied the total pre-orders induced by the measures' values on rule sets. A pre-order is a poset with a binary relation $\geq$ where $\geq$ is only known to be reflexive and transitive; this allows equality between two elements of the poset. We calculate an objective coefficient which tells us how two pre-orders are different.

This comparison is based on counts over all the possible couples of rules. For a couple of rules (r_1, r_2)[1], and two measures μ_1 and μ_2, there is strict agreement when $\mu_1(r_1) < \mu_1(r_2)$ and $\mu_2(r_1) < \mu_2(r_2)$, co-agreement when $\mu_1(r_1) > \mu_1(r_2)$ and $\mu_2(r_1) > \mu_2(r_2)$, large agreement if the two rules are equivalent for both measures. There is semi-agreement if for one of the measures, the value taken for r_2 is greater than the value taken for r_1, the values being equal for the other measure, and semi-disagreement if there is semi-agreement for (r_2, r_1). Finally, there is strict disagreement if the value taken by one of the measures is greater for r_1 than for r_2, the opposite being true for the other measure. These situations are summarized in table 7.

Table 7: Summary of the 9 possible rankings of two rules by two measures

	$\mu_1(r_1) < \mu_1(r_2)$	$\mu_1(r_1) = \mu_1(r_2)$	$\mu_1(r_1) > \mu_1(r_2)$
$\mu_2(r_1) < \mu_2(r_2)$	strict agreement	semi-agreement	strict disagreement
$\mu_2(r_1) = \mu_2(r_2)$	semi-agreement	large agreement	semi-disagreement
$\mu_2(r_1) > \mu_2(r_2)$	strict disagreement	semi-disagreement	co-agreement

Lingoes [1979] defined 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 pre-orders are equal. In this case, there are only strict agreements, co-agreements or large agreements. The minimum value is obtained if for any couple of different rules, there is either strict disagreement, semi-disagreement or semi-agreement.

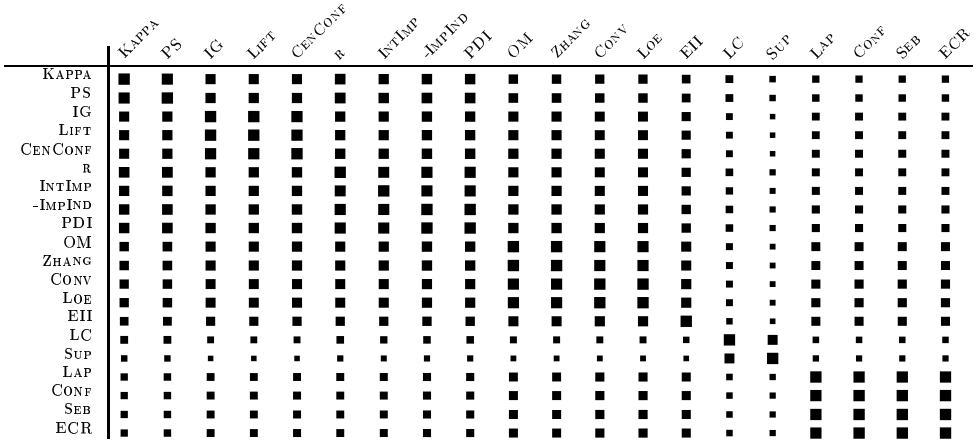
In the first case, both measures sort the rules in the same way and the subset of the n best rules is the same. On the contrary, in the second case, the ordering of the rules is reversed, and no couple of distinct rules that are equivalent for one of the measures can be equivalent for the other. Hence, no rule can belong to the first half of both rankings.

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*

[1]The order of the rules is important, and r_1 and r_2 may represent the same rule.

choice, Lim et al. [2000], a subset of the 1987 National Indonesia Contraceptive Prevalence Survey). The rule set is composed of 444 rules, generated by the APRIORI algorithm of Borgelt and Kruse [2002] with a support threshold of 10% and a confidence threshold of 80%. The results are presented in table 8. 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 columns have been reorganized 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].

Table 8: Comparison of total pre-orders between 20 measures



We have only 10 negative values, the lowest being -0.14 for (LC, OM). The average value of τ_1 is 0.60, and the variance is 0.08. 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 is confirmed by the formulas given in table 6. This means that although some measures do generate the same rankings, there are some significant differences, and the subset of n best rules will differ, depending on the measure used. This is illustrated in table 9, presenting the proportion of rules in common within the subsets of n best rules for 4 measures. Moreover, as some measures do generate the same rankings, the user may freely pick out among these the one that best fits his preferences, without loss of interesting rules.

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

Table 9: Proportion of best rules in common in the subset of n best rules

n	20	50	70	100	150	200	400
CenConf & Conv	50%	50%	47.14%	49%	74%	81%	97.75%
CenConf & Loe	50%	50%	47.14%	49%	74%	81%	97.75%
CenConf & OM	60%	60%	62.86%	56%	77.33%	82%	98.5%
Conv & Loe	100%	100%	100%	100%	100%	100%	100%
Conv & OM	90%	90%	84.29%	93%	96.67%	99%	99.25%
Loe & OM	90%	90%	84.29%	93%	96.67%	99%	99.25%

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 a MCDA framework. From the user's point of view, the problem can then be resumed to 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 of noise?), 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 two final 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 rules of the type $\mathtt{A} \rightarrow \mathtt{B}$ which can be applied to $\mathcal{D}$. We call $\mathtt{A}$ the antecedent and $\mathtt{B}$ the conclusion of the rule. $\mathtt{A}$ and $\mathtt{B}$ are logical forms on the attributes. In this study we are interested in the particular case of association rules as defined by Agrawal et al. [1993] ;
- $\mathcal{M}$ is a set of interestingness measures of the rules of $\mathcal{R}$ (see Section 3, table 6) ;
- $\mathcal{A}$ 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 with his objectives. The major difficulty in the construction of $\mathcal{P}$ is the formalization of the user's objectives. They are often given in a 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 basis of the sets $\mathcal{A}$ and $\mathcal{P}$. To make it short, 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 don't evaluate the quality of the whole set of rules $\mathcal{R}$.

Two actors take part in this analysis: the user who is an expert of the data (expert of $\mathcal{D}$ and $\mathcal{R}$), who tries to select the *best* rules of $\mathcal{R}$ and the analyst, specialist of MCDA procedures and of KDD, who tries to help the expert. We call E_r the first one and E_a the second one. Consequently the main problem is to translate the properties of $\mathcal{A}$ in 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 side, the sets $\mathcal{M}$, $\mathcal{A}$ and $\mathcal{F}$ are related to the expertise of E_a.

The resolution of this problem implies a narrow 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 modeled and be used to build a family of criteria $\mathcal{F}$ for aiding in the selection of the *best* measure(s).

6 Evaluation criteria

In this section, we propose a list of eligible properties to evaluate the previous list of measures. The 5 last properties arise from discussions within the CNRS group GafoQualité. We present each property, explaining its interest and its possible values on an ordinal scale. The results of the evaluations[2] are then presented in table 11. Table 10 summarises the semantic and the number of modalities of our 8 criteria.

g_1**: asymmetric processing of** `A` **and** `B` **[Freitas, 1999].** Since the head and body of a rule may have a very different signification, it is desirable to distinguish measures that give different evaluations of rules $\mathtt{A} \rightarrow \mathtt{B}$ and $\mathtt{B} \rightarrow \mathtt{A}$ from those which do not. We note 0 if the measure is symmetric, 1 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 1 if the measure is a decreasing function with n_b, 0 otherwise.

[2]Theoretically, for each property, a user may express any preference, and our particular assignment of the values (0, 1 and sometimes 2) is not to be considered as representative of users' preferences.

g_3**: reference situations, independence [Piatetsky-Shapiro, 1991].** To avoid keeping rules that contain no information, it is necessary to eliminate the $\mathtt{A} \rightarrow \mathtt{B}$ rule when $\mathtt{A}$ and $\mathtt{B}$ are independent, which means when the probability of obtaining $\mathtt{B}$ is independent of the fact that $\mathtt{A}$ is true or not. A comfortable way of dealing with this is to require that a measure's value at independence should be constant. We note 1 if the measure value is constant at independence and 0 otherwise.

g_4**: reference situations, logical rule.** In the same way, the second reference situation we consider is related to the value of the measure when there is no counter example. It is desirable that the value should be constant or possibly infinite. We note 1 in the case of a constant or infinite value, 0 otherwise.

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 $\mathtt{B}$ cannot be realized if $\mathtt{A}$ already is. Our choice is based on the fact that incompatibility is related to the rule $\mathtt{A} \rightarrow \overline{\mathtt{B}}$ and not $\mathtt{A} \rightarrow \mathtt{B}$.

g_5**: linearity with $p_{a\bar{b}}$ around 0^+.** Some authors [Gras et al., 2002] express the desire to have a weak decrease in the neighborhood of a logic 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. However, the opposite choice may be preferred as a convex decrease with $n_{a\bar{b}}$ around the logic rule increases the sensitivity to a false positive. We hence note 0 if the measure is convex with $n_{a\bar{b}}$ near 0, 1 if it is linear and 2 if it is concave.

g_6**: sensitivity to n (total number of records).** Intuitively, if the rates of presence of $\mathtt{A}$, $\mathtt{A} \rightarrow \mathtt{B}$, $\mathtt{B}$ are constant, it may be interesting to see how the measure reacts to a global extension of the database (with no evolution of rates). The preference of the user might be indifferent to having a measure which is invariant or not with the dilatation of data. 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 0 if the measure is invariant and 1 if it increases with n.

g_7**: easiness to fix a threshold.** Even if properties g_3 and g_4 are valid, it is still difficult to decide 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 conventional) way by providing a sense of the strength of the evidence against the null hypothesis, that is the p-value. 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 such a probability.

More generally, we can define such a threshold from one of the three types of models proposed by Lerman [1970] to establish the law followed by n_{ab} under the hypothesis of link absence (H_0). We note 1 if the measure easily supports such an evaluation, and 0 otherwise.

g_8**: intelligibility.** 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 can be expressed in one concrete sentence. We affect the value 2 to this property if the measure can be expressed in that way, 1 if the measure can be estimated with common quantities, and 0 if it seems impossible to give any concrete explanation of the measure.

Table 10: Properties of the measures

Property	Semantic	Modalities
g_1	asymmetric processing of `A` and `B`	2
g_2	decrease with n_b	2
g_3	reference situations: independence	2
g_4	reference situations: logical rule	2
g_5	linearity with $n_{a\overline{b}}$ around 0^+	3
g_6	sensitivity to n	2
g_7	easiness to fix a threshold	2
g_8	intelligibility	3

We evaluated the measures described in the previous section with respect to these criteria and we obtain the following decision matrix (table 11).

The extension of this list is currently being studied, and in particular the discrimination and antimonotonicity characters of a measure. Discrimination is quite interesting since it might be related to criteria g_6 (sensitivity to the cardinality of the total space), which generally occurs simultaneously with a loss of discrimination. Antimonotonicity is a very interesting property from the computing point of view, both for Apriori algorithms and Galois lattice based methods [Pasquier et al., 1999].

7 Evaluation of the interestingness measures

In this section, we will analyze and evaluate the measures described above and resumed in table 6. This analysis is done by a MCDA procedure called Promethee [Brans and Mareschal, 1994, 2002]. Its general objectives are to build partial and complete rankings on alternatives (in this case, the measures) and to visualize the structure of the problem in a plane called the Gaia plane,

Table 11: Decision matrix

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
Sup	0	0	0	0	1	0	1	2
Conf	1	0	0	1	1	0	1	2
r	0	1	1	0	1	0	1	1
CenConf	1	1	1	0	1	0	1	2
PS	0	1	1	0	1	1	1	1
Loe	1	1	1	1	1	0	1	1
Zhang	1	1	1	1	2	0	0	0
- ImpInd	1	1	1	0	1	1	1	0
Lift	0	1	1	0	1	0	1	1
LC	1	1	0	0	1	0	1	1
Seb	1	0	0	1	0	0	1	1
OM	1	1	1	1	0	0	1	2
Conv	1	1	1	1	0	0	1	1
ECR	1	0	0	1	2	0	1	1
Kappa	0	1	1	0	1	0	1	0
IG	0	1	1	0	2	0	1	0
IntImp	1	1	1	1	2	1	1	0
EII	1	1	1	1	2	1	0	0
PDI	1	1	1	0	1	1	1	0
Lap	1	0	0	0	1	0	1	0

similarly to a principal component analysis. The Promethee method requires information about the criteria 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 alternatives within each criterion. This means that for small (resp. large) deviations, E_r will allocate a small (resp. large) preference to the best index. This is done through preference functions. Then, each alternative is confronted with the other alternatives in order to define outranking flows. The positive (resp. negative) outranking flow expresses how an alternative a is outranking (resp. outranked by) the others. Finally, partial and complete rankings are built out of these outrankings. The Gaia plane provides information on the conflicting character of the criteria and on 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 alternatives and the criteria in a common plane.

7.1 A glance at the Promethee method

Let $A = \{a_1, \ldots, a_m\}$ be a set of possible alternatives. Let $\{g_j(\cdot), j = 1, \ldots, k\}$ be a set of evaluation criteria to be maximized or minimized. Each of the possible alternatives of A are evaluated on each of the criteria.

Pairwise comparison As the method is based on pairwise comparisons of the alternatives, it is necessary to represent and to formalize the degree of

preference of one alternative on another. For a pair of alternatives $(a_l, a_m) \in A \times A$, for one particular criterion, a small (respectively a large) difference of the evaluations should be represented by a weak (respectively a strong) degree of preference. The authors of the method suggest to represent these preferences by real numbers of the unit interval $[0, 1]$. The concept of preference function P_j helps to calculate this degree of preference based on the differences of evaluations of the alternatives for a given criterion g_j:

$$P_j((a_l, a_m)) = P_j(g_j(a_l) - g_j(a_m)), \text{ where } 0 \leq P_j((a_l, a_m)) \leq 1.$$

Of course there is no preference of a_l over a_m when $g_j(a_m)$ is better than $g_j(a_l)$.

In case of a criterion to be maximized, the preference function's general shape is given on figure 2. It is an increasing and continuous function. We use in our study the usual preference function (figure 3).

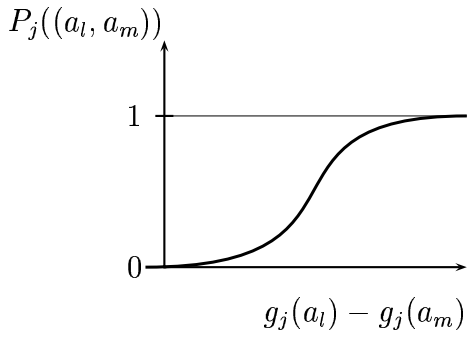


Figure 2: Typical preference function

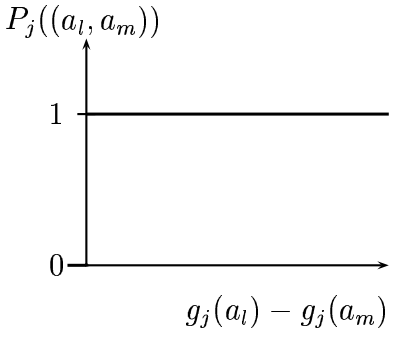


Figure 3: Usual preference function

The PROMETHEE approach proposes six different types of preference functions, each of them needing 0, 1 or 2 parameters to be defined by the user (an indifference threshold, a threshold for strict preference and a parameter in between).

In our problem the retained criteria are purely ordinal. We therefore consider that a non zero difference of evaluations on alternatives a_l and a_m should be reflected in the preference of a_l over a_m, and choose the usual preference function, which equals 1 in $]0, +\infty[$ and 0 elsewhere.

Aggregated preference index The next step consists in 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:

$$\pi(a_l, a_m) = \sum_{j=1}^{k} P_j((a_l, a_m)) \cdot w_j$$

where w_j is the weight associated with criterion j, such that $\sum_{j=1}^{k} w_j = 1$. We observe that $\pi(a_l, a_l) = 0$ and $0 \leq \pi(a_l, a_m) \leq 1, \forall a_l, a_m \in A$. Furthermore, for a given pair of alternatives (a_l, a_m), a preference index close to 0 (respectively close to 1) implies a weak (respectively a strong) global preference of a_l over a_m.

Outranking flows In order to build a ranking on A, each alternative has to be compared with the $n-1$ other ones of A. This is done with the computation of two outranking flows, the positive outranking flow $\phi^+(a) = \frac{1}{n-1} \sum_{x \in A} \pi(a, x)$ and the negative outranking flow $\phi^-(a) = \frac{1}{n-1} \sum_{x \in A} \pi(x, a)$.

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 rankings Two possible rankings can be obtained in the PROMETHEE approach: a partial and a complete ranking. The partial ranking is an intersection of the rankings which can be deduced from both the positive and the negative outranking flows. It is built as follows:

$$\begin{cases} a_l P_p a_m \iff \begin{cases} \phi^+(a_l) > \phi^+(a_m) \text{ and } \phi^-(a_l) < \phi^-(a_m) \\ \phi^+(a_l) = \phi^+(a_m) \text{ and } \phi^-(a_l) < \phi^-(a_m) \\ \phi^+(a_l) > \phi^+(a_m) \text{ and } \phi^-(a_l) = \phi^-(a_m) \end{cases} \\ a_l I_p a_m \iff \phi^+(a_l) = \phi^+(a_m) \text{ and } \phi^-(a_l) = \phi^-(a_m) \\ a_l R_p a_m \text{ otherwise.} \end{cases}$$

P_p, I_p and R_p stand for preference, indifference and incomparability relations respectively.

The complete ranking is built out of the net outranking flow (balance of flow) $\phi(a_l) = \phi^+(a_l) - \phi^-(a_l)$ as follows:

$$\begin{cases} a_l P_c a_m \iff \phi(a_l) > \phi(a_m) \\ a_l I_c a_m \iff \phi(a_l) = \phi(a_m) \end{cases}$$

The GAIA plane The PROMETHEE method allows to visualize the alternatives and the criteria in a common plane called the GAIA plane. It gives a synthetic and 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. Nevertheless, the GAIA plane allows to visualize the behavior of the final ranking with respect to different configurations of the weights.

The building of the GAIA plane is based on the analysis of particular net flows, relatively to each of the criteria, by decomposing the net outranking flow. We have:

$$\phi(a_l) = \phi^+(a_l) - \phi^-(a_l) = \frac{1}{n-1} \sum_{j=1}^{k} \sum_{a \in A} [P_j((a_l, a)) - P_j((a, a_l))] w_j.$$

We then can see that:

$$\phi(a_l) = \sum_{j=1}^{k} \phi_j(a_l) w_j, \tag{1}$$

where $\phi_j(a_l) = \frac{1}{n-1} \sum_{a \in A} [P_j((a_l, a)) - P_j((a, a_l))]$ is the unicriterion net flow. Each alternative can therefore be characterized by its k unicriterion net flows $\alpha(a_l) = (\phi_1(a_l), \ldots, \phi_k(a_l))$ and can be represented in a k-dimensional space whose axes correspond to the different criteria.

The corresponding cloud of points is projected into a 2-dimensional space in order to represent the information in a more synthetical way. Let us observe that in the k-dimensional space, the set of points is centered around the origin, as $\sum_{a \in A} \phi_j(a) = 0$. The projection is done along the first two factor axes of the principal components analysis. Furthermore the initial unitary vectors of the axes of the k-dimensional space are also projected in the plane. These projections represent the criteria in the new space.

According to the decomposition 1, the net flow of an alternative is the scalar product between the vector which represents its unicriterion net flow and the vector of the weights. This also means that the net flow of the alternative a is also the projection of $\alpha(a)$ on the weights vector w in the k-dimensional space.

Therefore, the projection of the $\alpha(a), \forall a \in A$ on w leads to the total ranking. The w vector is though a decision axis. Its projection in the GAIA plane is called the decision axis PROMETHEE π.

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 (respectively opposite) preferences on the set of alternatives are represented by axes which have a similar direction (respectively opposing directions).
- independent criteria are represented by orthogonal axes.
- alternatives which have *good* values 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. In this case, the w vector is nearly orthogonal to the GAIA plane. This means that for this configuration of weights, the criteria are conflicting, and a good compromise can be found at the origin.

Stability intervals on the weights can easily be calculated so that no modification of the complete ranking takes place.

7.2 Analysis of the quality measures

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

Sc1: The expert 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 neighborhood 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 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 neighborhood of the maximum at least) and the order on the values of criterion g_5 is convex $\succ$ linear $\succ$ concave.

We first analyze the problem with equal weights for the criteria. The total rankings for both scenarios are given in table 12.

Table 12: Total rankings for scenarios **Sc1** and **Sc2**.

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

First, we notice that both scenarios reflect the preferences of E_r on the shape of the curve. We can see that for **Sc1** the two leading measures are IntImp and EII which are both concave. Similarly, for **Sc2**, the two leading measures are OM 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 experience shows that it is important to distinguish both scenarios. If we give criterion g_5 an important weight (33%), the first positions of the ranking for **Sc1** (respectively for **Sc2**) are held by IntImp, EII, Zhang, ECR, Loe and IG (respectively by OM, Conv, Seb, and Loe) which are mostly concave (respectively convex). This analysis furthermore shows that the linear measure Loe is a very interesting measure as it is well placed in both scenarios. It stands for a good compromise.

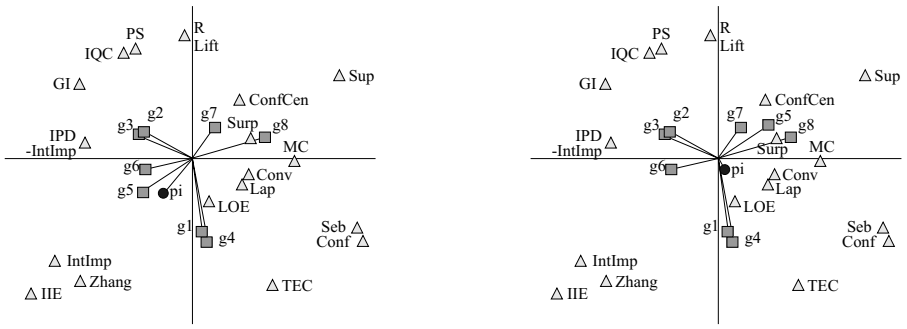
Table 13: Net flows for **Sc1** and **Sc2**.

Sc1	IntImp	EII	Loe	OM	CenConf	Conv	...	Lap
ϕ	.32	.18	.18	.16	.13	.08	...	-.32
Sc2	OM	Conv	Loe	CenConf	IntImp	-ImpInd	...	Sup
ϕ	.38	.30	.20	.15	.12	.10	...	-.30

A sensitivity analysis on the weights system shows that small changes in the weights affect the ranking. A closer look shows that these modifications only occur locally and that the first positions of the ranking remain stable. This is confirmed by the values of the net flows ϕ of the 5 leading elements of each of

the rankings presented in table 13. This table shows that the $\phi(a), a \in \mathcal{M}$ are spread uniformly between their minimum and their maximum values for both scenarios. In particular, we can see that the leading positions vary only for very significative changes in the weights 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 12 are good hints.

An analysis of the GAIA planes gives us further indications on the measures. Figure 4 shows the GAIA planes for **Sc1** and **Sc2**.



Plan **Sc1** Plan **Sc2**

Figure 4: Plans GAIA pour **Sc1** et **Sc2**

Let us first note that the percentage of cumulated variance for the first two factors represented by the GAIA plane is 58.8%. The information taken from the GAIA plane should therefore be considered as approximative and conclusions be drawn with much care. Firstly we observe that the measures (triangles on the figure) are distributed homogenously in the plane. Secondly we can see that the GAIA plane is well covered by the set of criteria (axes with squares on the figure). We conclude that the description of the selected measures by the criteria is discriminant and only little redundant.

For **Sc1** we can see that several couples of criteria are independent: (g_4, g_5), (g_4, g_6), (g_1, g_5), (g_1, g_6), (g_4, g_8) and (g_1, g_8). We can also observe conflicting sets of criteria. For example $\{g_1, g_4\}$ and $\{g_2, g_3\}$ are conflicting. Similar observations can be done for the two sets $\{g_5, g_6\}$ and $\{g_7, g_8\}$. This type of information gives hints on the behavior and the structure of the problem. For example, measures of $\mathcal{M}$ which are good for criterion g_5 (concave) will tend to be bad for criterion g_8 (unintelligible).

For **Sc2** similar observations can be done. The major difference lies in criterion g_5 which represents similar preferences than criteria g_7 and g_8 but is conflicting with g_6.

The decision axis π is quite long in **Sc1** and heads in the opposite direction

of g_7 and g_8. This means that measures which allow to fix the threshold easily and which are very 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 weights choice 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 rather short. This indicates that the vector w is almost orthogonal to the GAIA plane in the k-dimensional space. As indicated in earlier in Section 7.1 the good compromise is situated close to the origin. This explains the ranking of table 12.

The positions of the measures in the GAIA plane (for **Sc1** and **Sc2**) show that many alternatives have similar behaviors with respect to weights variations. This is confirmed by their similar profiles in the decision matrix. Thus PS and KAPPA, or SEB and CONF, or LIFT and R, or -IMPIND and PDI are close in the GAIA plane and have similar profiles. This couples of measures will tend to appear in neighbor positions in the ranking. An important comment should be done at this point of 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 basis of other techniques. An erroneous conclusion would be to consider CONV and LAP as similar measures due to their proximity in the GAIA plane. In fact, their profiles are very different and consequently their behavior in case of weights variations will not be similar.

This quite detailed study of the problem shows the utility of an analysis by means of a MCDA tool like PROMETHEE. On the basis of the previously made observations we can suggest two strategies.

The first strategy consists in checking first that the expert 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 weights to guarantee a certain coherence. Indeed a measure which has not 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 of 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 usiong this initial set of weights for the criteria. The stability of the resulting ranking should then be analyzed, especially for the leading positions. If a stable ranking is obtained, the GAIA plane, the value of the net flows and the profiles visualization tool allow a finer analysis of the leading measures. The values of the net flows gives

a hint on the *distance* between two alternatives in the ranking. Two measures with similar values for the flows can be considered as similar.

The second strategy consists in a first step in an exploration of the GAIA plane. This procedure helps the expert understand the structure of the problem and detect similar and different measures. Furthermore, the visualization of the criteria in the same plane as the alternatives allows to visualize the influence of the modification of the weights on the final ranking. This exploratory strategy should be applied with an expert who has an a priori knowledge on certain measures. He will be able to determine his preferences on the importance of the criteria by detecting some well known measures in the GAIA plane. Using this weights system as, the first strategy can then 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 simulation of the behavior 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 readable and that the thresholds are constant. The weights 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 14.

Table 14: Ranking and net flow for the preferences of E_r

1	2	4	5	...
OM (.30)	CENCONF, LOE (.22)	CONF (.17)	INTIMP (.16)	...

We can clearly see that OM is the best measure for this weights 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 weights system. Furthermore, OM stays in its leading position with a net flow which is significantly greater than the one of the second measure. The remaining desires of E_r are also satisfied. Indeed the measure is good on g_3, g_4, g_7 and g_8. Besides, OM is also competitive on g_1 and g_2. Its weaknesses are in the shape of the curve (convex) and its sensitivity to n, but these criteria are considered as less important by E_r.

8 Conclusion

In this article, we have proposed an initial array of 20 eligible measures evaluated on 8 properties. Given this array, we have shown how to use an MCDA method,

and help expert users choose an adapted interestingness measure in the context of association rules. Our approach is a first step to improve 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]...

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 user's preferences (the GAIA plane makes it easy to identify different clusters of criteria and alternatives).

Of course, the set of criteria has to be extended. Our set of criteria covers a large range of the user's preferences, but it is clearly not exhaustive. New criteria could also lead to a better distinction between measures which are similar at the present time. We are confident that some important criteria may also arise from experimental evaluation (such as the discrimination strength).

Finally, we would like to point out that even if SUP is poorly rated in both scenarios it is a mandatory measure in algorithms like APRIORI since its antimonotonicity property drives and simplifies the exploration of the lattice of itemsets. In our set of 20 measures, SUP is the only one having this property.

Acknowledgments

Benoît Vaillant would like to thanks the CUB (Urban Communauty of Brest) for financial support of his Ph.D. thesis. The authors would like to thank members's of the CNRS group GAFOQUALITÉ for productive discussions about *interestingness measure*.

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.

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.

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., 1994. The PROMETHEE-GAIA decision support system for multicriteria investigations. Investigation Operativa 4 (2), 102–117.

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

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., march 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.

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

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 la Didactique, Travaux et Thèses. La Pensée Sauvage.

Gras, R., Couturier, R., Bernadet, M., Blanchard, J., Briand, H., Guillet, F., Kuntz, P., Lehn, R., Peter, P., 2002. Quelques critères pour une mesure de qualités de règles d'association. Rapport de recherche pour le groupe de travail GAFOQUALITÉ de l'action spécifique STIC fouille de bases de données, Ecole Polytechnique de l'Université de Nantes.

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.

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

Hilderman, R. J., Hamilton, H. J., 2001b. Knowledge Discovery and Measures of Interest. Kluwer Academic Publishers.

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., Teytaud, O., 2004. Évaluation et validation de l'intérêt des règles d'association. RNTI-E-1, 193–217.

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., Meyer, P., Vaillant, B., Picouet, P., 2003. 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.

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.

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

Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L., 1999. Discovering frequent closed itemsets for association rules. In: Beeri, C., Buneman, P. (Eds.), Database Theory - ICDT '99, 7th International Conference, Jerusalem, Israel, January 10-12, 1999, Proceedings. 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.

Picouet, P., Lenca, P., 2001. Bases de données et internet. Hermes science, Ch. Extraction de connaissances à partir des données, pp. 395–420.

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

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.

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.

Terano, T., Liu, H., Chen, A. L. P. (Eds.), April 2000. Association Rules. Vol. 1805 of Lecture Notes in Computer Science. Springer.

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.

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.