Association Rule Interestingness Measures: Experimental and Theoretical Studies

Philippe Lenca[1], Benoît Vaillant[1], Patrick Meyer[2], and Stéphane Lallich[3]

[1] GET/ENST Bretagne, LUSSI Department, TAMCIC, UMR CNRS 2872, France
philippe.lenca@enst-bretagne.fr, benoit.vaillant@enst-bretagne.fr
[2] University of Luxemburg, Applied Mathematics Unit, Luxembourg
patrick.meyer@uni.lu
[3] University of Lyon 2, ERIC Laboratory, France
stephane.lallich@univ-lyon2.fr

Summary. It is a common problem that KDD processes may generate a large number of patterns depending on the algorithm used, and its parameters. It is hence impossible for an expert to assess these patterns. This is the case with the well-known APRIORI algorithm. One of the methods used to cope with such an amount of output depends on using association rule interestingness measures. Stating that selecting interesting rules also means using an adapted measure, we present a formal and an experimental study of 20 measures. The experimental studies carried out on 10 data sets lead to an experimental classification of the measures. This study is compared to an analysis of the formal and meaningful properties of the measures. Finally, the properties are used in a multi-criteria decision analysis in order to select amongst the available measures the one or those that best take into account the user's needs. These approaches seem to be complementary and could be useful in solving the problem of a user's choice of measure.

Key words: association rule, interestingness measure, interestingness criteria, measure classification, measure selection.

Introduction

One of the main objectives of Knowledge Discovery in Databases (KDD) is to produce interesting patterns. This notion of interest highly depends on the user's goals. 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 pattern is difficult to evaluate objectively. Indeed, this estimation greatly depends on the expert user's interests [48], [37]. Ideally, a pattern should be *valid, new and comprehensive* [24], but these generic terms cover a large number of situations when examined in a precise context. It is

Philippe Lenca et al.: *Association Rule Interestingness Measures: Experimental and Theoretical Studies*, Studies in Computational Intelligence (SCI) **43**, 51–76 (2007)
www.springerlink.com

a common problem that data mining algorithms produce a huge amount of output, and that the end user is then unable to analyse it individually. What is more, a large part of this output is uninteresting [72]. Thus, when dealing with pattern selection one has to face two problems: the quantity and the quality of rules. This is particularly true when mining association rules with the well-known algorithms of the Apriori family, within a support-confidence framework [2], and this is the issue that we will assess.

In this context, different solutions, more or less involving the user [83], can been considered. Visual data mining uses human visual capabilities to explore the data and patterns discovered (e.g. [78], [84], [47], [79], [43]). Human centred approaches emphasize the cooperation between the user and learning algorithms (e.g. [67], [54], [56], [9]).

Finally, interestingness measures can be used in order to filter and/or sort discovered rules (e.g. [37], [88], [38], [39]). Generally, one distinguishes between objective and subjective interestingness measures. Objective measures are said to be *data-driven* and only take into account the data cardinalities. Subjective measures are *user-driven* in the sense that they take into account the user's *a priori* knowledge and goals. For a discussion about subjective aspects of rule interestingness measures, the reader can refer to [82], [65] and [66].

It should be noted that, in practice, both objective and subjective approaches should be used to select interesting rules [26], the objective ones serving as a kind of first filter to select potentially interesting rules, while the subjective ones can be used as a final filter to retain only the truly interesting rules, depending on the applicative context.

We will focus on objective interestingness measures and take into account both user preferences or goals for association rule discovery and the nature of the data being mined. Such rules were defined in [2]: 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* $\mathtt{A}$*, then he/she probably also buys item* $\mathtt{B}$. It is of importance to make the distinction between the association rule $\mathtt{A} \rightarrow \mathtt{B}$, which focuses on cooccurrence and gives asymmetric meaning to $\mathtt{A}$ and $\mathtt{B}$, and logical implication $\mathtt{A} \Rightarrow \mathtt{B}$ or equivalence $\mathtt{A} \Leftrightarrow \mathtt{B}$ [51].

Interestingness measures play an essential role, reducing the number of discovered rules and retaining only the *best ones*, in a post-processing step.

In order to improve the selection of rules, many classical measures have been used, like the Chi-square test for independency, or the correlation coefficient. Due to specific needs, additional measures have been proposed, such as the lift [17], the M_{GK} measure [33], relative interestingness [41], general measure [44], the entropic intensity of implication [31], the probabilistic discriminant index [63], the maximal participation index [40], or the h-confidence [94], information theoretic based measures [12], parametrised measures [52]. As a consequence, a large number of measures are available (see for example [34] for an extensive list of classical measures).

Depending on the user's goals, data mining experts may propose the use of an appropriate interestingness measure, but this selection task cannot be done by the expert user, if left on his own.

This choice is hard, since rule interestingness measures have many different qualities or flaws, since there is no *optimal* measure. One way to solve this problem is to try to find good compromises [59]. A well-known example of such a controversial measure is the support. On the one hand, it is heavily used for filtering purposes in APRIORI algorithms [2], [73], as its anti-monotonicity 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 the value for a logical rule [75]. Finally, one should be very careful when using the support-confidence framework in defining the interestingness of a rule [76], [16]. To bypass this difficulty different works look for highly correlated items, like as in the CORCLASS algorithm [96] and in the algorithms presented in [21].

It is then relevant to study interestingness measures, so that rules are selected according to the user's needs and context [59]. Interestingness measures have to support KDD process through system-human interaction [71], [1]. Many works (for instance [6], [53], [36], [37], [87], [85], [51], [17], [60], [89], [86], [70]) have formally extracted and studied several specificities of various measures, and the importance of objective evaluation criteria of interestingness measures has already been focused on by [75] and [26].

In this chapter, we will assess the issue of selecting an adapted interestingness measure faced with an applicative context and user's aims.

First, we introduce a set of 20 classical measures which seem applicable in an association rule mining context [57]. In the second section, these measures are analyzed through eight formal properties that make sense from an end user's point of view. In order to highlight the wide variety of measures and have a case based overview of their behaviour, the third section focuses on a tool we have developed, HERBS [90], and an empirical classification of the measures is built out of experimental campaigns [92]. This classification is then compared to another clustering of the measures, based on their theoretical studies. Out of theoretical properties, we finally propose a multi-criteria decision aid (MCDA) approach assessing the issue of selecting an measure adapted to the user's context (aims, goals, nature of the data, etc.) [59]. Finally, we conclude and outline some perspectives that are to be studied.

1 Interestingness measures

In this section, we present the 20 objective association rules interestingness measures that we studied. These measures are usually defined using the 2×2 contingency table presented in figure 1, and is a classical way of measuring association in the case of paired attributes [23], such as in the GUHA method [21], in the 4FT-MINER tool [80] and in the APRIORI algorithm [2].

Given a rule A → B, we note:

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

For any X, we note p_x instead of n_x/n when we consider relative frequencies rather than absolute frequencies on the data set E. It is clear that, given n, n_a and n_b, or p_a and p_b, knowing one cell of the contingency table in figure 1 is enough to deduce the other ones.

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

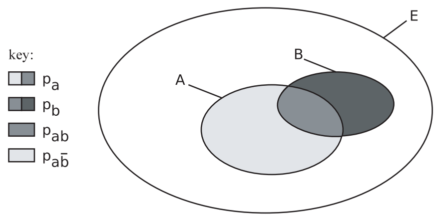


Fig. 1. Notations

We restricted the list of measures 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. Thus some measures like χ^2, Pearson's r^2, Goodman and Smyth's J-measure or Pearl's measure are not considered in this study. The selected measures are listed in table 1, which also includes bibliographical references. Their definition and co-domain, using absolute frequencies, is given in table 2. At first glance, table 2 shows important variations between the formulae. This is due to the fact that measures do not tell the same story. These variations are also noticeable since co-domains are quite different ([0, 1], [0, +∞[,]-∞, 1] and others with bounds depending on n_a, n_b and/or n_{ab}). For taking into account such variations one may use aggregation operators of valued relations [5] or normalized measures [25].

For a given decreasing monotonic measure μ (with respect to $n_{a\bar{b}}$ margins n_a and n_b being fixed), the selection of interesting rules is done by positioning a threshold α and keeping only the rules satisfying $\mu(A \rightarrow B) \geq \alpha$. The value of this threshold α has to be fixed by the expert, and the same threshold is considered for all the rules extracted during the data mining process. Thus, fixing α is an important issue [16].

Table 1. List of selected measures

	Name	References
BF	Bayes factor	[45]
CenConf	centred confidence	
Conf	confidence	[2]
Conv	conviction	[18]
ECR	examples and counter-examples rate	
EII	entropic intensity of implication	[31]
IG	information gain	[20]
- ImpInd	implication index	[64]
IntImp	intensity of implication	[29]
Kappa	Kappa coefficient	[22]
Lap	Laplace	[28]
LC	least contradiction	[3]
Lift	Lift	[17]
Loe	Loevinger	[36]
PDI	probabilistic discriminant index	[63]
PS	Piatetsky-Shapiro	[75]
r	Pearson's correlation coefficient	[74]
Seb	Sebag and Schoenauer	[81]
Sup	support	[2]
Zhang	Zhang	[95]

In our set of measures, we kept the well-known support and confidence: these are the two most frequently used measures in algorithms based on the selection of frequent itemsets for association rule extraction [2], [73].

Many other measures are linear transformations of the confidence, enhancing it, by enabling comparisons with p_b. This transformation is generally achieved by centering the 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 and Schoenauer's or the rate of examples and counter-examples, are monotonically increasing transformations of confidence, while the information gain is a monotonically increasing transformation of the lift. Thus, these measures will rank rules in the same order and differ, for example, from their semantic meaning [28].

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. These last two measures were adapted in order to let them have the desired property of being constant under a null hypothesis (this property is discussed in section 2). For the intensity of implication, the statistical law was

Table 2. Association rule quality measures

	Definition	Co-domain
BF	$\frac{n_{ab}n_{\bar{b}}}{n_b n_{a\bar{b}}}$	$[0, +\infty[$
CenConf	$\frac{nn_{ab}-n_a n_b}{nn_a}$	$[-\frac{n_b}{n}, \frac{n_{\bar{b}}}{n}]$
Conf	$\frac{n_{ab}}{n_a}$	$[0, 1]$
Conv	$\frac{n_a n_{\bar{b}}}{nn_{a\bar{b}}}$	$[\frac{n_{\bar{b}}}{n}, +\infty[$
ECR	$\frac{n_{ab}-n_{a\bar{b}}}{n_{ab}} = 1 - \frac{1}{\frac{n_a}{n_{a\bar{b}}}-1}$	$]-\infty, 1]$
EII	$\left\{[(1-h_1(\frac{n_{a\bar{b}}}{n})^2)(1-h_2(\frac{n_{a\bar{b}}}{n})^2)]^{1/4}\textsc{IntImp}\right\}^{1/2}$	$[0, 1]$
IG	$\log(\frac{nn_{ab}}{n_a n_b})$	$]-\infty, \log\frac{n}{n_b}]$
-ImpInd	$\frac{n_a n_b - nn_{ab}}{\sqrt{nn_a n_{\bar{b}}}}$	$[-\frac{\sqrt{n_a}n_b}{\sqrt{nn_{\bar{b}}}}, \sqrt{\frac{n_a n_{\bar{b}}}{n}}]$
IntImp	$P\left[N(0,1) \geq \textsc{ImpInd}\right]$	$[0, 1]$
Kappa	$2\frac{nn_{ab}-n_a n_b}{nn_a+nn_b-2n_a n_b}$	$[-2\frac{n_a n_b}{n_a n_{\bar{b}}+n_{\bar{a}}n_b}, 2\frac{n_a n_{\bar{b}}}{n_a n_{\bar{b}}+n_{\bar{a}}n_b}]$
Lap	$\frac{n_{ab}+1}{n_a+2}$	$[\frac{1}{n_a+2}, \frac{n_a+1}{n_a+2}]$
LC	$\frac{n_{ab}-n_{a\bar{b}}}{n_b}$	$[-\frac{n_a}{n_b}, \frac{n_a}{n_b}]$
Lift	$\frac{nn_{ab}}{n_a n_b}$	$[0, \frac{n}{n_b}]$
Loe	$\frac{nn_{ab}-n_a n_b}{n_a n_{\bar{b}}}$	$[-\frac{n_b}{n_{\bar{b}}}, 1]$
PDI	$P\left[\mathcal{N}(0,1) > \textsc{ImpInd}^{CR/\mathcal{B}}\right]$	$]0, 1[$
PS	$n_{ab} - \frac{n_a n_b}{n}$	$[-\frac{n_a n_b}{n}, \frac{n_a n_{\bar{b}}}{n}]$
R	$\frac{nn_{ab}-n_a n_b}{\sqrt{nn_a n_b n_{\bar{a}} \cdot n_{\bar{b}}}}$	$[-\sqrt{\frac{n_a n_b}{nn_{\bar{a}}n_{\bar{b}}}}, \sqrt{\frac{n_a n_{\bar{b}}}{nn_{\bar{a}}n_b}}]$
Seb	$\frac{n_{ab}}{n_{a\bar{b}}}$	$[0, +\infty[$
Sup	$\frac{n_{ab}}{n}$	$[0, \frac{n_a}{n}]$
Zhang	$\frac{nn_{ab}-n_a n_b}{\max\{n_{ab}n_{\bar{b}}, n_b n_{a\bar{b}}\}}$	$[-1, 1]$

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

$h_1(t) = -(1-\frac{n\cdot t}{n_a})\log_2(1-\frac{n\cdot t}{n_a}) - \frac{n\cdot t}{n_a}\log_2(\frac{n\cdot t}{n_a})$ if $t \in [0, n_a/(2n)[$; 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 $t \in [0, n_{\bar{b}}/(2n)[$; else $h_2(t) = 1$

$\mathcal{N}(0,1)$ stands for the centered and reduced normal repartition function approximated using the centred and reduced normal distribution function. The entropic intensity of implication was modified, according to the definition of the truncated entropic intensity of implication, TEII, as presented in [52].

The bayesian factor, also called sufficiency in [26] or odd-multiplier by [28], 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 $\overline{\mathtt{A}}$ on `B`. It has been thoroughly studied in [32].

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

2 Evaluation properties

In this section, we propose a list of eight meaningful properties to evaluate the previous list of measures. We present each property, explaining its interest and the modalities it can take.

Two actors take part in this analysis: the user who is an expert of the data mined, whose problem is to select the *best rules*, and the analyst, a specialist of MCDA and KDD, who tries to help the expert. We call the former E_r and the latter E_a.

For some properties, a preference order on the modalities they can take is straightforward. These properties can be considered as criteria by E_a without the intervention of E_r, namely g_1, g_2, g_3, g_4 and g_7, and 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, and will be called subjective [60].

For normative properties, we note `yes` if the measure has the desired property and `no` otherwise.

Table 3 recalls the semantics and the number of modalities of the 8 properties. The results of the evaluations are summarized in table 4.

Property g_1**: asymmetric processing of `A` and `B` [26].** Since the head and the body of a rule may have a very different signification, it is desirable to distinguish measures that give different evaluations of rules `A` $\rightarrow$ `B` and `B` $\rightarrow$ `A` from those that do not. We note `no` if the measure is symmetric, `yes` otherwise.

Property g_2**: decrease with n_b [75].** Given n_{ab}, $n_{a\bar{b}}$ and $n_{\bar{a}\bar{b}}$, 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` (i.e. verifying `B` but not `A`) increases, the interestingness of the rule should decrease. We note `yes` if the measure is a decreasing function with n_b, `no` otherwise.

Property g_3**: reference situations, independence [75].** To avoid keeping rules that contain no information, it is necessary to eliminate the `A` $\rightarrow$ `B` rule when `A` and `B` are independent, which means that the probability of obtaining `B` is independent 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. We note `yes` if the measure's value at independence is constant and `no` otherwise.

Property g_4**: reference situations, logical rule [57].** 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 2), three cases arise. First, the measure takes a value independent of the marginal frequencies (see table in figure 1) and thus takes a constant and maximal value[4]. A second case is considered when the measure takes an infinite value when $n_{a\bar{b}}$ is null. 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 or possibly infinite. We note `yes` in the cases of a constant or infinite value, `no` 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 `B` cannot be realized if `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 $\mathtt{A} \rightarrow \mathtt{B}$.

Property g_5**: linearity with $p_{a\bar{b}}$ around 0^+ [17].** Some users 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 of them. 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 `convex` if the measure is convex with $n_{a\bar{b}}$ near 0, `linear` if it is linear and `concave` if it is concave.

Property g_6**: sensitivity to n (total number of records) [51], [17].** Intuitively, if the rates of presence of `A`, $\mathtt{A} \rightarrow \mathtt{B}$, `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).

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. The preference of the user might be indifferent to having a measure which is invariant or not with the dilatation of data. We note `desc` (for descriptive measures) if the measure is invariant and `stat` (for statistical ones) if it increases with n.

Property g_7**: easiness to fix a threshold [57].** 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 [51]. In some cases, the measure is defined as such a probability. More

[4] 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.

generally, we can define such a threshold from one of the three types of models proposed by [62] to establish the law followed by $n_{a\bar{b}}$ under the hypothesis of link absence. We note `yes` if the measure easily supports such an evaluation, and `no` otherwise.

Property g_8**: intelligibility [57].** 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 of the data E_r[5]. We assign the value `yes` to this property if the measure can be expressed in that way, `avg` if the measure can be estimated with common quantities, and `no` if it seems impossible to give any simple concrete explanation of the measure.

Table 3. Properties of the measures

Property	Semantics	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 $n_{a\bar{b}}$ around 0^+	3
g_6	sensitivity to n	2
g_7	easiness to fix a threshold	2
g_8	intelligibility	3

The extension of this list is currently being studied, and in particular discrimination, antimonotonicity, and robustness to noise. 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 also is an interesting property from the computing point of view, both for Apriori algorithms and Galois lattice based methods [73]. Robustness to noise has been focused on in [4] and [61].

Finally, different alternatives could be proposed for property g_3 (independence). It could be interesting to replace the independence condition ($p_{b/a} = p_b$) by the equilibrium condition ($p_{b/a} = 0.5$) that corresponds to predictive purposes [10]. More generally, a confidence threshold θ ($p_{b/a} = \theta, p_b < \theta < 1$) could be taken into account, especially for targeting purposes [52].

---

[5] 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 4. Evaluation matrix

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
BF	yes	yes	yes	yes	convex	desc	yes	yes
CenConf	yes	yes	yes	no	linear	desc	yes	yes
Conf	yes	no	no	yes	linear	desc	yes	yes
Conv	yes	yes	yes	yes	convex	desc	yes	avg
ECR	yes	no	no	yes	concave	desc	yes	avg
TEII	yes	yes	yes	no	concave	stat	no	no
IG	no	yes	yes	no	concave	desc	yes	no
- ImpInd	yes	yes	yes	no	linear	stat	yes	no
IntImp	yes	yes	yes	no	concave	stat	yes	no
Kappa	no	yes	yes	no	linear	desc	yes	no
Lap	yes	no	no	no	linear	desc	yes	no
LC	yes	yes	no	no	linear	desc	yes	avg
Lift	no	yes	yes	no	linear	desc	yes	yes
Loe	yes	yes	yes	yes	linear	desc	yes	avg
PDI	yes	yes	yes	no	concave	stat	yes	no
PS	no	yes	yes	no	linear	stat	yes	avg
r	no	yes	yes	no	linear	desc	yes	avg
Seb	yes	no	no	yes	convex	desc	yes	avg
Sup	no	no	no	no	linear	desc	yes	yes
Zhang	yes	yes	yes	yes	concave	desc	no	no

3 Interestingness measure classifications

Beyond a formal analysis, based on meaningful properties, it is interesting to observe the behavior of the measures on data. We present an experimental classification based on preorder comparisons, these preorders being induced by interestingness measures on rule sets. This classification is carried out using our experimentation tool, Herbs. A formal classification based on the formal properties is proposed using a hierarchical ascendent clustering. Finally, we compare the two classifications.

3.1 An overview of Herbs, an experimentation tool

The aim of Herbs [90], [46] is to analyse rule sets and compare or investigate interestingness measures through concrete experiments. It has been designed as an interactive *post-analysis* tool, and hence data sets, rule sets and interestingness measures are considered as inputs. Various useful experimentation schemes are implemented in Herbs, from simple descriptive statistics about rule sets, to comparative overviews of the evaluation of a rule set by several measures.

We here propose an experimental analysis and comparison of measures, based on their application to 10 pairs of data sets and rule sets. A synthetic comparison of the rankings of a rule set by the measures is given by computing a preorder agreement coefficient, τ_1 which is derived from Kendall's τ (see [27]). This agreement compares a pair of preorders induced by two measures, and its value is in the range $[-1;1]$. The maximum value is obtained when the two pre-orders are equal, whereas the minimum value is obtained in various cases, and especially for reversed preorders.

From a computational point of view, using such a coefficient can be seen as complex since its evaluation is done in $\mathcal{O}(\eta^2)$, where η is the number of rules in the rule set considered, when a correlation analysis can be done in $\mathcal{O}(\eta)$ (the correlation index between interestingness measures is used in the ARQAT tool [42] for example). Still, from the numerous coefficients presented in [27], the τ_1 coefficient best suits our needs. What is more, HERBS uses a relational database in order to store the experimental results. Building an index on these values greatly optimizes the computation of this coefficient. Finally, only a slight modification of the formula is required in order to return to more classical agreement coefficients, such as Kendall's τ or Spearman's ρ.

3.2 Experimental classification

Experiments were carried out on databases retrieved from the UCI Repository (`ftp://ftp.ics.uci.edu/` [8]). When there is no ambiguity, we will refer indifferently to the pair formed by a data set and a rule set, or to the single data set or rule set, using their names in the Repository. We denote by *BCW* the *breast-cancer-wisconsin* database. The parameters of the APRIORI algorithm [9] were fixed experimentally in order to obtain rule sets of an acceptable size in terms of computational cost (see table 5). The great differences in size of the rule sets is related to the number of modalities of the different attributes of the case databases. A particular option was used in order to compute *Cmc*: APRIORI, which usually explores a restricted number of nodes of the lattice formed by the different modalities of the attributes, was forced to explore the entire lattice. *Cmc2* was obtained by filtering *Cmc*, with a minimum lift of 1.2. The *Solarflare* database is divided into two case sets, $\mathcal{SF}_1$ and $\mathcal{SF}_2$, described by the same attributes. $\mathcal{R}_1$ (resp. $\mathcal{R}_2$) is the rule set coming from $\mathcal{SF}_1$ (resp. $\mathcal{SF}_2$). We filtered $\mathcal{R}_1$, with the method exposed in [91] following the results of [50] in order to keep only rules that are significant from a statistical point of view. Using $\mathcal{SF}_1$ (resp. $\mathcal{SF}_2$), we obtained the rule set $\mathcal{R}_1^1$ (resp. $\mathcal{R}_1^2$). The characteristics of the sets are summarized in table 5.

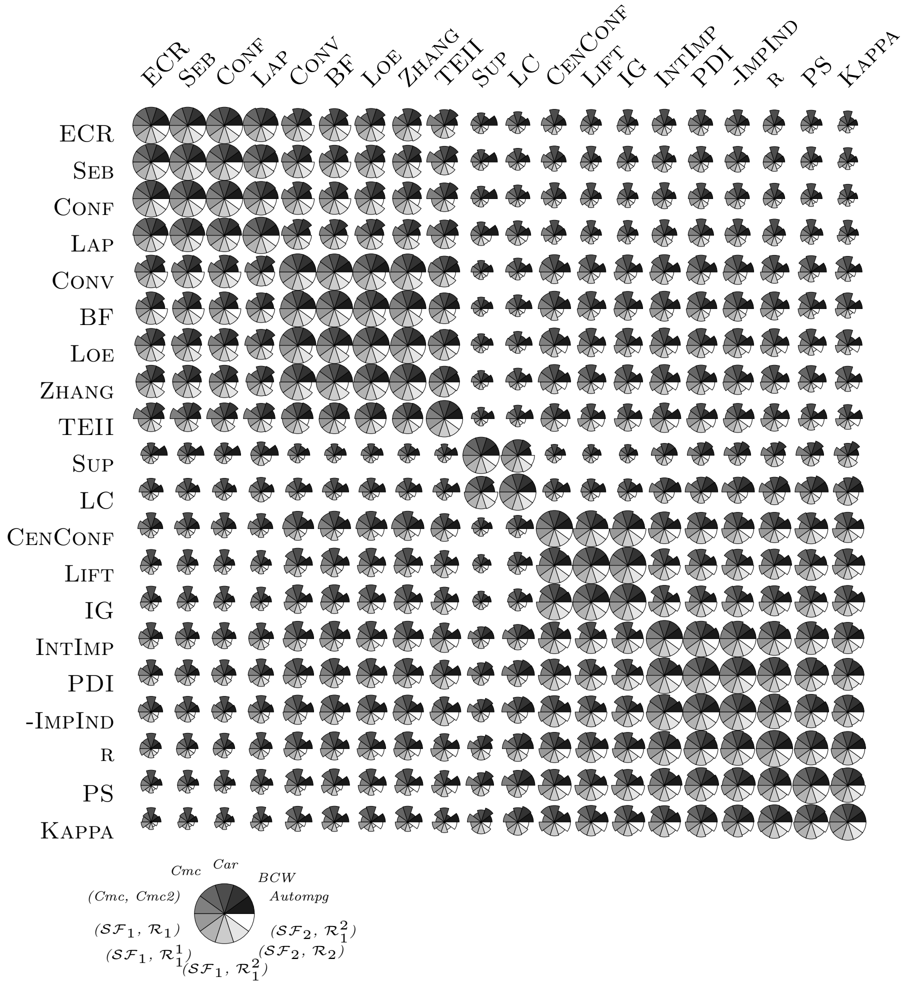
We generated 10 preorder comparison matrices, which are presented in table 6 (the value of τ_1 is proportional to the radius of the corresponding portion of disc, a radius null corresponding to an agreement of -1, and a radius of 1 corresponding to an agreement value of 1). The AMADO method [19] was applied to the average matrix of the results in order to reorganize the rows and the columns of this matrix, and highlight the block structures. The results are quite in agreement, and we can make out 3 main groups of measures, and in two of these groups we can distinguish two subgroups (see tables 6 and 7).

Table 5. Summary of the different sets used, and APRIORI parameters

name	n	sup_{min}	$conf_{min}$	η
Autompg	392	5	50	49
BCW	683	10	70	3095
Car	1728	5	60	145
Cmc	1473	5	60	2878
Cmc2	n/a	n/a	n/a	766

name	n	sup_{min}	$conf_{min}$	η
$(\mathcal{SF}_1, \mathcal{R}_1)$	323	20	85	5402
$(\mathcal{SF}_2, \mathcal{R}_2)$	1066	20	85	6312
$\mathcal{R}_1^1$	n/a	n/a	n/a	4130
$\mathcal{R}_1^2$	n/a	n/a	n/a	2994

Table 6. Preorder comparisons of 20 measures on 10 experiments.



The first group consists of {ECR, SEB,CONF, LAP, CONV, BF, LOE, ZHANG, TEII} and can be sub-categorized into two subgroups: E_1 ={ECR, SEB,CONF, LAP} and E_2 ={CONV, BF, LOE, ZHANG, TEII}. The second main group consists of E_3 ={SUP, LC}, behaving very differently from the previous measures. The third group, {CenConf, Lift, IG, IntImp, PDI, -ImpInd, r, PS, Kappa}, can be split into two, as was the first one, and leads to the two following subgroups: E_4 ={CenConf, Lift, IG} and E_5 ={IntImp, PDI, -ImpInd, r, PS, Kappa}.

3.3 Formal classification

The formal approach can be synthetized with a 20×8 matrix, containing the evaluation of the 20 measures on the 8 properties. We kept only 6 of the properties for the comparison between experimental and formal approaches, as two of them – namely g_7 (easiness to fix a threshold) and g_8 (intelligibility) – do not influence the experimental results at all.

All these properties are bivaluate except g_5 which is trivaluate. The 20×6 matrix formally obtained was re-encoded in a 20×6 matrix composed of real values, 0 or 1 in the binary cases, and 0, 0.5 or 1 for g_5. These values do not represent any judgement on the measures, but only list the properties shared by the different measures.

The typology in 5 classes, F_i, $i = 1 \dots 5$ (see table 7) coming from this matrix is obtained with a hierarchical ascendant clustering, using the average linkage, applied to the Manhattan distance.

3.4 Comparison of the two classifications

Table 7 shows that both approaches globally lead to similar clusterings, but some shifts are interesting. The main differences concern {Sup, LC} and TEII.

The experimental classification leads to two main classes, $E_1 \cup E_2$ and $E_4 \cup E_5$. The coherence between the two classifications is underlined by the fact that apart from the three above-mentioned measures, $E_1 = F_1 \cup F_2$, $F_3 \subset E_2$ and $E_4 \cup E_5 \subset F_4 \cup F_5$.

From a formal point of view, Sup and LC are quite close, forming class F_2 together with Lap. There also is a strong link between the classes F_1 and F_2. Apart from Sup and Lap, the measures belonging to these classes are those sharing the property of making reference to indetermination when evaluating the quality of a rule (i.e. measures having a constant value when $n_{ab} = n_{a\bar{b}} = n_a/2$, [11], [10]), although this property was not taken into account in our formal classification.

The formal class F_5 is made out of the measures built on the implication index, namely -ImpInd itself, IntImp which is derived from the former through the use of the normal distribution, and the two discriminent measures, TEII and PDI. In our formal approach no dinstiction can be made between IntImp, TEII and PDI, since none of the criteria g_1 to g_6 take into account the discriminating power of the measures. We are currently working on such a criterion. Apart from TEII, these measures make up the same experimental class, which also includes r, Kappa and PS. The altered behavior of TEII is

Table 7. Cross-classification of the measures

Formal \ Experimental	Class E_1	Class E_2	Class E_3	Class E_4	Class E_5
Class F_1	Conf, Seb, ECR				
Class F_2	Lap		Sup, LC		
Class F_3		Conv, BF, Loe, Zhang			
Class F_4				Lift, IG, CenConf	r, Kappa, PS
Class F_5		TEII			IntImp, -ImpInd, PDI

due to the fact that it is derived from IntImp through the use of an inclusion index. This inclusion index plays a major role in the evaluation of the quality of a rule and thus accounts for the experimental differences. Experimentally, TEII thus shifts to Loe, Zhang, BF and Conv (class E_2).

Formally, Lap shifts to LC and Sup (class F_2). A reason for this shift is that although it is really close to Sup in our formal study, Lap can differ from Conf experimentally only for values of n_a close to 0 (nuggets). The minimum thresholds of the Apriori algorithms make this impossible, and this can be seen as an algorithmic bias [92].

Property g_4 has an important impact on experimental results. When it is verified, all the logical rules are evaluated with a maximal value, no matter what the conclusion is. BF, Conv, Loe, Zhang, and ECR, Seb, Conf, *i.e.* the measures for which $g_4 = \texttt{yes}$, make the experimental group $E_1 \cup E_2$. Only TEII and Lap, also belonging to these classes, do not share this property.

4 A multi-criteria decision approach towards measure selection

In this section, we will analyze and evaluate the measures described earlier and summarized in table 2. This analysis was done by a few Mcda procedures, in particular the Tomaso method for sorting [69], a ranking procedure based on kernels of digraphs [7] and the Promethee method [15]. These three methods have produced very similar results. In this chapter, we focus on the analysis by the Promethee method to obtain a ranking. A formalization of the decision problem is discussed in [58]. This approach has been used in a real context by [77].

4.1 A few words on the Promethee method

Its 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, similarly to a principal component analysis. The Promethee method requires information about the importance of the criteria (a criteria is a property on which a preference modeling is known) 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 other ones in order to define outranking flows. The positive (negative) outranking flow expresses to what degree a measure a is outranking (outranked by) the others. Finally, partial and complete rankings are generated from these outranking flows. 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 ϕ derived from the outranking flows, of the measures and the criteria in a common plane. For a more detailed description of this method, the reader can refer to [14], for example.

4.2 Analysis of the quality measures

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`, where $\succ$ means "is preferred to".

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`.

For both scenarios, for criterion g_6 we assume that the expert prefers a measure which increases with n, the size of the data. Thus, the order on the values of criterion g_6 is `stat` $\succ$ `desc`. For the other criteria which are assumed to be normative, the expert has no influence on the order of the values.

We start by analysing the problem with equal weights for the criteria to get a first view of the structure of the problem. The total rankings for the two scenarios are given in table 8.

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 PDI which are both concave. Similarly, for **Sc2**, the two leading measures are BF and CONV which are both convex. This first analysis also 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.

Sensitivity analyses on the weights systems show that small changes in the weights affect the rankings. Nevertheless a closer look shows that these

Table 8. 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	

modifications only occur locally and that the first positions of the rankings remain stable.

Therefore one can say that for an expert E_r 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 8 are quite acceptable.

An analysis of the Gaia planes gives us further indications about the measures. Figure 2 shows the Gaia planes for **Sc1** and **Sc2**.

Let us first note that the percentage of cumulated variance for the first two factors represented by the Gaia plane is 60.20%. The information taken from the Gaia plane should therefore be considered as approximative and conclusions be drawn with great care. First we observe that the measures (triangles in the figure) are distributed homogeneously in the plane. Then we can see that the Gaia plane is well covered by the set of criteria (axes with squares in the figure). We conclude that the description of the measures selected by the criteria is discriminant and only slightly redundant.

The Gaia plane furthermore helps to detect independent and conflicting criteria. The decision axis π (axis with a circle) indicates in what direction the best alternatives are situated for a given weights system.

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)[6]. 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 on the behaviour and the structure of the problem. For example, measures which are 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 is conflicting with g_6.

For **Sc1**, the decision axis π is moderately long and heads in the opposite direction of g_7 and g_8. This means that measures which allow us to fix the threshold 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

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

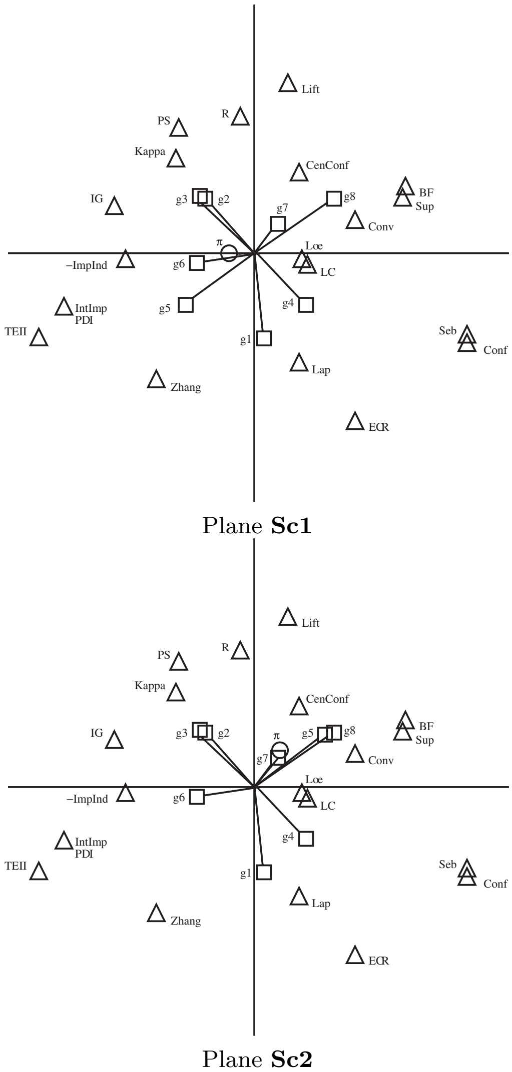


Fig. 2. Gaia planes for **Sc1** and **Sc2**

neglected compared to a criterion like g_8 (intelligibility). Thus, if the expert is aware of the impact of his weights 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 direction of g_7, g_5 and g_8. This partly explains the ranking of table 8.

The positions of the measures in the Gaia plane (for **Sc1** and **Sc2**) show that many alternatives have similar behaviors with respect to weight variations. This is confirmed by their similar profiles in the decision matrix. Thus Seb and Conf, or -ImpInd and PDI are close in the Gaia plane and have similar profiles. These couples of measures will tend to appear in neighbour positions in the rankings. An important comment should be made 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 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 the case of weight 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 observations above 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 properties. 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 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 analyzed, especially for the leading positions. If a stable ranking is obtained, the Gaia plane, the value of the net flows and the profile visualization tool allow a finer analysis of the leading measures. The values of the net flows give a hint about the *distance* between two alternatives in the ranking. Two measures with similar values for the net flows can be considered as similar.

The second strategy involves a first step in an exploration of the Gaia plane. This procedure helps the expert to understand the structure of the problem and to detect similar and different measures. Furthermore, the visualization of the criteria in the same plane as the alternatives make it possible 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 a preorder on the importance of the criteria by detecting some well known measures in the Gaia plane. By using this first approximate weights 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.

5 Conclusion and perspectives

Association rule quality measures play a major role within a KDD process, but they have a large diversity of properties, which have to be studied both on formal aspects and on real data in order to use a measure adapted to the user's context. In this chapter, we have studied 20 association rule interestingness measures evaluated on 8 properties, and 10 data sets.

The experimental results we present come from a tool we developed, HERBS briefly presented. We were then able to identify 3 main groups of measures in the two approaches, which may be refined in 5 smaller classes. The resulting clusterings are globally in agreement, and the discordancies discussed. The experimental approach seems to be an important addition to the formal approach. Indeed, it first confirmed the validity of the list of formal properties we thought were worth studying. What is more, it has also led to a new reflection on the importance of these properties. For example, requiring that a rule quality measure should have a fixed value for a logical rule has the bias of favouring logical rules with a large conclusion. From the formal study, we proposed a multicriteria decision aid approach illustrating how to help expert users choose an adapted interestingness measure in the context of association rule mining. We present the use of the PROMETHEE decision aid method.

Our approach is a first step to improving the quality of a set of rules that will effectively be presented to the user. Other factors, beyond interestingness measures, can be used. Among them, attribute costs and misclassification costs [26], and cognitive constraints [55].

In addition to the interest of having such a list of properties 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).

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 and the robustness).

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 to have this property.

Acknowledgment

Benot Vaillant would like to thanks Brest Métropole Océane for financial support of his Ph.D. thesis. The authors would like to thank the members of the CNRS working group GAFOQUALITÉ for interesting discussions about evaluation properties.

References

1. H. Abe, S. Tsumoto, M. Ohsaki, and T. Yamaguchi. Evaluating model construction methods with objective rule evaluation indices to support human experts. In V. Torra, Y. Narukawa, A. Valls, and J. Domingo-Ferrer, editors, *Modeling Decisions for Artificial Intelligence*, volume 3885 of *Lecture Notes in Computer Science*, pages 93–104, Tarragona, Spain, 2006. Springer-Verlag.
2. R. Agrawal, T. Imielinski, and A.N. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, *ACM SIGMOD International Conference on Management of Data*, pages 207–216, 1993.
3. J. Azé and Y. Kodratoff. Evaluation de la résistance au bruit de quelques mesures d'extraction de règles d'assocation. In D. Hérin and D.A. Zighed, editors, *Extraction des connaissances et apprentissage*, volume 1, pages 143–154. Hermes, 2002.
4. J. Azé and Y. Kodratoff. A study of the effect of noisy data in rule extraction systems. In *The Sixteenth European Meeting on Cybernetics and Systems Research*, volume 2, pages 781–786, 2002.
5. J. P. Barthélemy, A. Legrain, P. Lenca, and B. Vaillant. Aggregation of valued relations applied to association rule interestingness measures. In V. Torra, Y. Narukawa, A. Valls, and J. Domingo-Ferrer, editors, *Modeling Decisions for Artificial Intelligence*, volume 3885 of *Lecture Notes in Computer Science*, pages 203–214, Tarrogona, Spain, 2006. Springer-Verlag.
6. R. J. Bayardo and R. Agrawal. Mining the most interesting rules. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 145–154, 1999.
7. R. Bisdorff. Bipolar ranking from pairwise fuzzy outrankings. *Belgian Journal of Operations Research, Statistics and Computer Science*, 37 (4) 97:379–387, 1999.
8. C.L. Blake and C.J. Merz. UCI repository of machine learning databases. http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998.
9. J. Blanchard, F. Guillet, and H. Briand. A virtual reality environment for knowledge mining. In R. Bisdorff, editor, *Human Centered Processes*, pages 175–179, Luxembourg, 2003.
10. J. Blanchard, F. Guillet, H. Briand, and R. Gras. Assessing the interestingness of rules with a probabilistic measure of deviation from equilibrium. In J. Janssen and P. Lenca, editors, *The XIth International Symposium on Applied Stochastic Models and Data Analysis*, pages 191–200, Brest, France, 2005.
11. J. Blanchard, F. Guillet, H. Briand, and R. Gras. IPEE : Indice probabiliste d'écart à l'équilibre pour l'évaluation de la qualité des règles. In *Atelier Qualité des Données et des Connaissances (EGC 2005)*, pages 26–34, 2005.

12. J. Blanchard, F. Guillet, R. Gras, and H. Briand. Using information-theoretic measures to assess association rule interestingness. In *The 5th IEEE International Conference on Data Mining*, pages 66–73, Houston, Texas, USA, 2005. IEEE Computer Society Press.
13. C. Borgelt and R. Kruse. Induction of association rules: APRIORI implementation. In *Compstat'02*, pages 395–400, Berlin, Germany, 2002. Physica Verlag.
14. J.P. Brans and B. Mareschal. PROMETHEE-GAIA – *Une méthode d'aide à la décision en présence de critères multiples*. Ellipses, 2002.
15. J.P. Brans and P. Vincke. A preference ranking organization method. *Management Science*, 31(6):647–656, 1985.
16. T. Brijs, K. Vanhoof, and G. Wets. Defining interestingness for association rules. *International journal of information theories and applications*, 10(4):370–376, 2003.
17. S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: generalizing association rules to correlations. In *ACM SIGMOD/PODS'97 Joint Conference*, pages 265–276, 1997.
18. S. Brin, R. Motwani, J.D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In J. Peckham, editor, *ACM SIGMOD International Conference on Management of Data*, pages 255–264, Tucson, Arizona, USA, 1997. ACM Press.
19. J.-H. Chauchat and A. Risson. *Visualization of Categorical Data*, chapter 3, pages 37–45. Blasius J. & Greenacre M. ed., 1998. New York: Academic Press.
20. K.W. Church and P. Hanks. Word association norms, mutual information an lexicography. *Computational Linguistics*, 16(1):22–29, 1990.
21. E. Cohen, M. Datar, S. Fujiwara, A. Gionis, P. Indyk, R. Motwani, J. Ullman, and C. Yang. Finding interesting associations without support pruning. In *The 16th International conference on Data engineering*, 2000.
22. J. Cohen. A coefficient of agreement for nominal scale. *Educational and Psychological Measurement*, 20:37–46, 1960.
23. A.W.F. Edwards. The measure of association in a 2 x 2 table. *Journal of the Royal Statistical Society, Series A*, 126(1):109–114, 1963.
24. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors. *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 1996.
25. D. Feno, J. Diatta, and A. Totohasina. Normalisée d'une mesure probabiliste de la qualité des règles d'association : étude de cas. In *Atelier Qualité des Données et des Connaissances (EGC 2006)*, pages 25–30, 2006.
26. A. Freitas. On rule interestingness measures. *Knowledge-Based Systems journal*, pages 309–315, 1999.
27. V. Giakoumakis and B. Monjardet. Coefficients d'accord entre deux préordres totaux. *Statistique et Analyse des Données*, 12(1 et 2):46–99, 1987.
28. I.J. Good. The estimation of probabilities: An essay on modern bayesian methods. The MIT Press, Cambridge, MA, 1965.
29. R. Gras, S. Ag. Almouloud, M. Bailleuil, A. Larher, M. Polo, H. Ratsimba-Rajohn, and A. Totohasina. *L'implication Statistique, Nouvelle Méthode Exploratoire de Données. Application à la Didactique, Travaux et Thèses*. La Pensée Sauvage, 1996.
30. R. Gras, R. Couturier, J. Blanchard, H. Briand, P. Kuntz, and P. Peter. 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, 2004.

31. R. Gras, P. Kuntz, R. Couturier, and F. Guillet. Une version entropique de l'intensité d'implication pour les corpus volumineux. In H. Briand and F. Guillet, editors, *Extraction des connaissances et apprentissage*, volume 1, pages 69–80. Hermes, 2001.
32. S. Greco, Z. Pawlak, and R. Slowinski. Can bayesian confirmation measures be useful for rough set decision rules? *Engineering Applications of Artificial Intelligence*, 17(4):345–361, 2004.
33. S. Guillaume. *Traitement des données volumineuses, Mesures et algorithmes d'extraction de règles d'association et règles ordinales*. PhD thesis, Université de Nantes, 2000.
34. F. Guillet. Mesures de la qualité des connaissances en ECD. Atelier, Extraction et gestion des connaissances, 2004.
35. P. Hajek, I. Havel, and M. Chytil. The GUHA method of automatic hypotheses determination. *Computing*, (1):293–308, 1966.
36. R.J. Hilderman and H.J. Hamilton. Applying objective interestingness measures in data mining systems. In *Fourth European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 432–439. Springer Verlag, 2000.
37. R.J. Hilderman and H.J. Hamilton. Evaluation of interestingness measures for ranking discovered knowledge. *Lecture Notes in Computer Science*, 2035:247–259, 2001.
38. R.J. Hilderman and H.J. Hamilton. *Knowledge Discovery and Measures of Interest*. Kluwer Academic Publishers, 2001.
39. R.J. Hilderman and H.J. Hamilton. Measuring the interestingness of discovered knowledge: A principled approach. *Intelligent Data Analysis*, 7(4):347–382, 2003.
40. Y. Huang, H. Xiong, S. Shekhar, and J. Pei. Mining confident co-location rules without a support threshold. In *The 18th Annual ACM Symposium on Applied Computing*. ACM, 2003.
41. F. Hussain, H. Liu, E. Suzuki, and H. Lu. Exception rule mining with a relative interestingness measure. In T. Terano, H. Liu, and A.L.P. Chen, editors, *The Fourth Pacific-Asia Conference on Knowledge Discovery and Data Mining*, volume 1805 of *Lecture Notes in Artificial Intelligence*, pages 86–97. Springer-Verlag, 2000.
42. X-H. Huynh, F. Guillet, and H. Briand. ARQAT: An exploratory analysis tool for interestingness measures. In J. Janssen and P. Lenca, editors, *The XIth International Symposium on Applied Stochastic Models and Data Analysis*, pages 334–344, Brest, France, 2005.
43. A. Iodice D'Enza, F. Palumbo, and M. Greenacre. Exploratory data analysis leading towards the most interesting binary association rules. In J. Janssen and P. Lenca, editors, *The XIth International Symposium on Applied Stochastic Models and Data Analysis*, pages 256–265, Brest, France, 2005.
44. S. Jaroszewicz and D.A. Simovici. A general measure of rule interestingness. In *The 5th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 253–265, London, UK, 2001. Springer-Verlag.
45. H.J. Jeffreys. Some tests of significance treated by the theory of probability. In *Proceedings of the Cambridge Philosophical Society*, number 31, pages 203–222, 1935.
46. M. Kamber and R. Shingal. Evaluating the interestingness of characteristic rules. In *The Second International Conference on Knowledge Discovery and Data Mining*, pages 263–266, Portland, Oregon, August 1996.

47. D. A. Keim. Information visualization and visual data mining. *IEEE Transactions On Visualization And Computer Graphics*, 7(1):100–107, 2002.
48. M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A.I. Verkamo. Finding interesting rules from large sets of discovered association rules. In N.R. Adam, B.K. Bhargava, and Y. Yesha, editors, *Third International Conference on Information and Knowledge Management*, pages 401–407. ACM Press, 1994.
49. S. Lallich. Mesure et validation en extraction des connaissances à partir des données. Habilitation à Diriger des Recherches – Université Lyon 2, 2002.
50. S. Lallich, E. Prudhomme, and O. Teytaud. Contrôle du risque multiple en sélection de règles d'association significatives. In G. Hébrail, L. Lebart, and J.-M. Petit, editors, *Extraction et gestion des connaissances*, volume 1-2, pages 305–316. Cépaduès Editions, 2004.
51. S. Lallich and O. Teytaud. 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, 2004.
52. S. Lallich, B. Vaillant, and P. Lenca. Parametrised measures for the evaluation of association rule interestingness. In J. Janssen and P. Lenca, editors, *The XIth International Symposium on Applied Stochastic Models and Data Analysis*, pages 220–229, Brest, France, 2005.
53. N. Lavrac, P. Flach, and B. Zupan. Rule evaluation measures: A unifying view. In S. Dzeroski and P. Flach, editors, *Ninth International Workshop on Inductive Logic Programming*, volume 1634 of *Lecture Notes in Computer Science*, pages 174–185. Springer-Verlag, 1999.
54. E. Le Saux, P. Lenca, J-P. Barthélemy, and P. Picouet. Updating a rule basis under cognitive constraints: the COMAPS tool. In *The Seventeenth European Annual Conference on Human Decision Making and Manual Control*, pages 3–9, December 1998.
55. E. Le Saux, P. Lenca, and P. Picouet. Dynamic adaptation of rules bases under cognitive constraints. *European Journal of Operational Research*, 136(2):299–309, 2002.
56. R. Lehn, F. Guillet, P. Kuntz, H. Briand, and J. Philippé. Felix: An interactive rule mining interface in a KDD process. In P. Lenca, editor, *Human Centered Processes*, pages 169–174, Brest, France, 1999.
57. P. Lenca, P. Meyer, P. Picouet, B. Vaillant, and S. Lallich. Critères d'évaluation des mesures de qualité en ECD. *Revue des Nouvelles Technologies de l'Information (Entreposage et Fouille de Données)*, (1):123–134, 2003.
58. P. Lenca, P. Meyer, B. Vaillant, and S. Lallich. A multicriteria decision aid for interestingness measure selection. Technical Report LUSSI-TR-2004-01-EN, Département LUSSI, ENST Bretagne, 2004.
59. P. Lenca, P. Meyer, B. Vaillant, and P. Picouet. Aide multicritère à la décision pour évaluer les indices de qualité des connaissances – modélisation des préférences de l'utilisateur. In M.-S. Hacid, Y. Kodratoff, and D. Boulanger, editors, *Extraction et gestion des connaissances*, volume 17 of *RSTI-RIA*, pages 271–282. Lavoisier, 2003.
60. P. Lenca, P. Meyer, B. Vaillant, P. Picouet, and S. Lallich. 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, 2004.

61. P. Lenca, B. Vaillant, and S. Lallich. On the robustness of association rules. In *IEEE International Conference on Cybernetics and Intelligent Systems*, Bangkok, Thailand, 2006.
62. I.C. Lerman. *Classification et analyse ordinale des données*. Dunod, 1970.
63. I.C. Lerman and J. Azé. Une mesure probabiliste contextuelle discriminante de qualité des règles d'association. In M.-S. Hacid, Y. Kodratoff, and D. Boulanger, editors, *Extraction et gestion des connaissances*, volume 17 of *RSTI-RIA*, pages 247–262. Lavoisier, 2003.
64. I.C. Lerman, R. Gras, and H. Rostam. Elaboration d'un indice d'implication pour les données binaires, i et ii. *Mathématiques et Sciences Humaines*, (74, 75):5–35, 5–47, 1981.
65. B. Liu, W. Hsu, and S. Chen. Using general impressions to analyze discovered classification rules. In *Third International Conference on Knowledge Discovery and Data Mining*, pages 31–36, 1997.
66. B. Liu, W. Hsu, S. Chen, and Y. Ma. Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems*, 15(5):47–55, 2000.
67. B. Liu, W. Hsu, K. Wang, and S. Chen. Visually aided exploration of interesting association rules. In *Third Pacific-Asia Conference on Methodologies for Knowledge Discovery and Data Mining*, pages 380–389. Springer Verlag, 1999.
68. J. Loevinger. A systemic approach to the construction and evaluation of tests of ability. *Psychological monographs*, 61(4), 1947.
69. J.-L. Marichal, P. Meyer, and M. Roubens. Sorting multi-attribute alternatives: The TOMASO method. *Computers & Operations Research*, (32):861–877, 2005.
70. K. McGarry. A survey of interestingness measures for knowledge discovery. *Knowledge Engineering Review Journal*, 20(1):39–61, 2005.
71. M. Ohsaki, Y. Sato, S. Kitaguchi, H. Yokoi, and T. Yamaguchi. Comparison between objective interestingness measures and real human interest in medical data mining. In R. Orchard, C. Yang, and M. Ali, editors, *The 17th international conference on Innovations in Applied Artificial Intelligence*, volume 3029 of *Lecture Notes in Artificial Intelligence*, pages 1072–1081. Springer-Verlag, 2004.
72. B. Padmanabhan. The interestingness paradox in pattern discovery. *Journal of Applied Statistics*, 31(8):1019–1035, 2004.
73. N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. In C. Beeri and P. Buneman, editors, *The 7th International Conference on Database Theory*, volume 1540 of *Lecture Notes in Computer Science*, pages 398–416, Jerusalem, Israel, 1999. Springer.
74. K. Pearson. Mathematical contributions to the theory of evolution. iii. regression, heredity and panmixia. *Philosophical Transactions of the Royal Society*, A, 1896.
75. G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In G. Piatetsky-Shapiro and W.J. Frawley, editors, *Knowledge Discovery in Databases*, pages 229–248. AAAI/MIT Press, 1991.
76. P. Picouet and P. Lenca. *Bases de données et internet*, chapter Extraction de connaissances à partir des données, pages 395–420. Hermes Science, 2001.
77. M. Plasse, N. Niang, G. Saporta, and L. Leblond. Une comparaison de certains indices de pertinence des règles d'association. In G. Ritschard and C. Djeraba, editors, *Extraction et gestion des connaissances*, volume 1-2, pages 561–568. Cépaduès-Éditions, 2006.

78. F. Poulet. Visualization in data-mining and knowledge discovery. In P. Lenca, editor, *Human Centered Processes*, pages 183–191, Brest, France, 1999.
79. F. Poulet. Towards visual data mining. In *6th International Conference on Enterprise Information Systems*, pages 349–356, 2004.
80. J. Rauch and M. Simunek. Mining for 4ft association rules by 4ft-miner. In *Proceeding of the International Conference On Applications of Prolog*, pages 285–294, Tokyo, Japan, 2001.
81. M. Sebag and M. Schoenauer. Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases. In J. Boose, B. Gaines, and M. Linster, editors, *The European Knowledge Acquisition Workshop*, pages 28–1–28–20. Gesellschaft für Mathematik und Datenverarbeitung mbH, 1988.
82. A. Silberschatz and A. Tuzhilin. On subjective measures of interestingness in knowledge discovery. In *Knowledge Discovery and Data Mining*, pages 275–281, 1995.
83. A. Silberschatz and A. Tuzhilin. User-assisted knowledge discovery: How much should the user be involved. In *ACM-SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, 1996.
84. S.J. Simoff. Towards the development of environments for designing visualisation support for visual data mining. In S.J. Simoff, M. Noirhomme-Fraiture, and M.H. Böhlen, editors, *International Workshop on Visual Data Mining in cunjunction with ECML/PKDD'01*, pages 93–106, 2001.
85. E. Suzuki. In pursuit of interesting patterns with undirected discovery of exception rules. In S. Arikawa and A. Shinohara, editors, *Progresses in Discovery Science*, volume 2281 of *Lecture Notes in Computer Science*, pages 504–517. Springer-Verlag, 2002.
86. E. Suzuki. Discovering interesting exception rules with rule pair. In *ECML/PKDD Workshop on Advances in Inductive Rule Learning*, pages 163–178, 2004.
87. P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *The Eighth ACM SIGKDD International Conference on KDD*, pages 32–41, 2002.
88. P-N. Tan, V. Kumar, and J. Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 4(29):293–313, 2004.
89. A. Totohasina, H. Ralambondrainy, and J. Diatta. Notes sur les mesures probabilistes de la qualité des règles d'association: un algorithme efficace d'extraction des règles d'association implicative. In *7ème Colloque Africain sur la Recherche en Informatique*, pages 511–518, 2004.
90. B. Vaillant. Evaluation de connaissances: le problème du choix d'une mesure de qualité en extraction de connaissances à partir des données. Master's thesis, Ecole Nationale Supérieure des Télécommunications de Bretagne, 2002.
91. B. Vaillant, P. Lenca, and S. Lallich. Association rule interestingness measures: an experimental study. Technical Report LUSSI-TR-2004-02-EN, Département LUSSI, ENST Bretagne, 2004.
92. B. Vaillant, P. Lenca, and S. Lallich. A clustering of interestingness measures. In E. Suzuki and S. Arikawa, editors, *Discovery Science*, volume 3245 of *Lecture Notes in Artificial Intelligence*, pages 290–297, Padova, Italy, 2004. Springer-Verlag.
93. B. Vaillant, P. Picouet, and P. Lenca. An extensible platform for rule quality measure benchmarking. In R. Bisdorff, editor, *Human Centered Processes*, pages 187–191, 2003.

94. H. Xiong, P. Tan, and V. Kumar. Mining strong affinity association patterns in data sets with skewed support distribution. In *Third IEEE International Conference on Data Mining*, pages 387–394, Melbourne, Florida, 2003.
95. T. Zhang. Association rules. In T. Terano, H. Liu, and A.L.P. Chen, editors, *4th Pacific-Asia Conference Knowledge Discovery and Data Mining, Current Issues and New Applications*, volume 1805 of *Lecture Notes in Computer Science*, Kyoto, Japan, 2000. Springer.
96. A. Zimmermann and L. De Raedt. CorClass: Correlated association rule mining for classification. In E. Suzuki and S. Arikawa, editors, *Discovery Science*, volume 3245 of *Lecture Notes in Artificial Intelligence*, pages 60–72, Padova, Italy, 2004. Springer-Verlag.