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

Association rule interestingness measures: an experimental study

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

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

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

LUSSI-TR-2004-02-EN
July 2004

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

Contents

Association rule interestingness measures: an experimental study

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

*: GET ENST Bretagne / Département LUSSI – CNRS TAMCIC, France
**: Université Lumière, Lyon 2 – Laboratoire ERIC, France

Abstract

It is an accepted fact that KDD processes generate a large number of patterns. It is hence impossible for an expert in the field being mined to sustain these patterns. This is a frequent issue with the well-known Apriori algorithm. One of the classical methods used to cope with such an amount of output depends on the use of interestingness measures. Stating that extracting interesting rules also means using an adapted measure, we present an experimental study of the behaviour of 20 measures on 10 datasets. This study is compared to a previous analysis of formal and meaningful properties of the measures, by means of two clusterings. One of the goals of this experimental study is to enhance our previous approach. Both approaches seem to be complementary and could be profitable for the problem of a user's choice of a measure.
Keywords: association rule interestingness measure, clustering, experimental study.

Association rule interestingness measures: an experimental study

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

*: GET ENST Bretagne / Département LUSSI – CNRS TAMCIC, France
**: Université Lumière - Lyon 2 – Laboratoire ERIC, France

1 Introduction

One of the main objectives of Knowledge Discovery in Databases (KDD) is the production of rules, interesting from the point of view of a user – who usually is an expert in the field being mined. A rule can be labeled as interesting providing that it is "valid, novel, potentially useful, and ultimately understandable" [Fayyad et al., 1996]. These generic terms cover a wide range of aspects, and the quality of a rule is information which is quite hard to grasp. Moreover, the large number of rules generated by the algorithms commonly used makes it impossible for an expert to take all the rules into consideration without some automated assistance. A common way of reducing the number of rules, and hence enabling the expert to focus on what should be of interest to him, is to pre-filter the output of KDD algorithms according to interestingness measures. By enabling the selection of a restricted subset of "best rules" out of a larger set of potentially valuable ones, interestingness measures play a major role within a KDD process. We show in Lenca et al. [2003b] that the selection of the best rules implies the use of an adapted interestingness measure, as the measures may generate different rankings (see for exemple experimental studies of Hilderman and Hamilton [2001]). Different criteria, depending mostly on the user's goals, may give hints, and a multi-criteria decision approach can be applied in order to recommend a restricted number of candidate measures. However, the choice of a measure responding to a user's needs is not easy. This can be explained by the fact that the measures may have various and sometimes incompatible properties (Tan et al. [2002], Lallich and Teytaud [2004]). As there is no *optimal* measure, a way of solving this problem is to try to find good compromises. In Lenca et al. [2003a], we focused on helping the user select an adapted interestingness measure with regards to his goals and preferences, and with respect to formal properties of the measures.

The aim of the experiments developed in this paper is to complete our formal approach with an analysis of the behaviour of measures on concrete data. This two-fold characterisation is useful since some properties of the measures are hard to evaluate in a formal way, such as the robustness to noise [Azé and Kodratoff, 2002].

In our paper, the survey based on formal evaluations of measures done in [Lenca et al., 2003a] is confronted with experimental results (see [Hilderman and Hamilton, 2001] for different experimental studies on sixteen measures). In order to carry out the experiments we developped HERBS, a rule dataset and interestingness measure analysis tool. In section 2 we present an experimentation scheme using HERBS. In section 3, we show a detailed analysis for one dataset, and more results are summarized in section 4. The experimental clustering of the measures is compared to a hierarchical ascendant clustering of the measures in section 5. We conclude in section 6.

2 A short description of the analysis tool, HERBS

The aim of HERBS [Vaillant et al., 2003] is to analyse and compare rule datasets or interestingness measures. It has been designed to be a *post-analysis* tool, and hence case datasets, rule datasets and interestingness measures are seen as inputs. HERBS is a graphical tool designed either for the expert in the data mined or for the KDD expert.

2.1 Analysis background, notations

We will use in this article the following notations. $\mathcal{C}$ is a case dataset containing n cases and $\mathcal{R}$ a rule dataset containing η rules. Each case is described by a fixed number of attributes (usually in $\{0,1\}$, $\mathbb{N}$, $\mathbb{R}$, or in a fixed set of strings). The dataset may have been cleaned and we assume that there are no missing values.

We consider association rules of the form $\mathtt{A} \rightarrow \mathtt{B}$ where $\mathtt{A}$ and $\mathtt{B}$ are conjunctions of tests on the values taken by the attributes. In the AIS algorithm initially proposed by Agrawal et al. [1993], the attributes are all binary, hence they can be tested only once in each rule; what is more, the conclusion is restricted to a single test. This kind of analysis is a particular case of the analysis of a contingency table, introduced by Hajek et al. [1966] within the GUHA method and developed later on by Rauch and Simunek [2001] in the 4FT-MINER tool.

We note n_a the number of cases of $\mathcal{C}$ matching $\mathtt{A}$, n_b the number of items $\mathtt{B}$, n_{ab} the number of items matching both $\mathtt{A}$ and $\mathtt{B}$ (the examples of the rule), and $n_{a\bar{b}}$ those matching $\mathtt{A}$ but not $\mathtt{B}$ (or counter-examples); hence $n_{a\bar{b}} = n_a - n_{ab}$.

For a given conjuction of tests X, we will refer either to its absolute frequency representation within the data, n_x, or to its relative frequency, $p_x = n_x/n$.

It is important to keep in mind the fact that we do not bind the rules with the dataset used to create them. Thus, n_a, n_b, n_{ab} and $n_{a\bar{b}}$ can potentially take any integer value from 0 to n, providing that $n_{ab} \leq \min(n_a, n_b)$. For a given $\mathcal{C}$, we can reject rules that do not make sense (rules having a null value for n_a, n_b or n_{ab}). This can be easily done by filtering the rules and retaining only those having a support strictly greater than zero for example.

The interestingness measures we will take into account are all functions of n, n_a, n_b and n_{ab}. They are defined in order to have a ranking of the rules, from the "best" to the "worst", though this interestingness relies highly on the users' aims.

We call t_c the coverage rate, which is the proportion of cases in $\mathcal{C}$ verifying at least one of the rules in $\mathcal{R}$. Such cases are said to be covered by $\mathcal{R}$. We note $\mathcal{C}_{|\mathcal{R}}$ the subset of $\mathcal{C}$ of all the cases covered by $\mathcal{R}$. We call t_r the recovering measure, defined as the average number of redundant rules on $\mathcal{C}_{|\mathcal{R}}$ (*i.e.* the average number of rules verifying each case of $\mathcal{C}_{|\mathcal{R}}$ minus one).

2.2 Experimentation scheme

As mentioned above, HERBS is a post-analysis tool. It does not mine the data for knowledge, case and rule sets are its inputs. The user's task is to select the appropriate sets, the interestingness measures for a given experimentation scheme, and specify the parameters of the latter. The results are presented to the user, who can modify the inputs of the scheme in order to tune it more precisely.

The analysis and the experimental comparison of the measures we propose rely on their application to a couple ($\mathcal{C}$, $\mathcal{R}$). As both sets are considered as inputs, it is interesting to start the analysis with a synthetic overview of how well they match together. For example, one may want to know how many "extreme" rules there are for the couple of sets. An extreme rule may be a rule with no examples, or no counter-examples, or rules that weaken the knowledge of `B` (*i.e.* rules such that the probability of `B` knowing `A` is lower than the plain probability of `B`).

The coverage rate and the recovering measure are good synthetic indicators of the structural adequation of the rules to the cases. The first one is an overview of how well the dataset and the rule set match, and the second one quantifies the classification redundancy of the cases by the rules. With the classical scheme *learning* and *generalisation* tests we used (section 3), it is useful to know such values.

In order to analyse with greater precision the rankings of the rules of $\mathcal{R}$ by

the various measures, one may choose to observe the subset of rules that are selected at least k times by these measures as belonging to the subset of N best rules. For $k = 1$, we hence obtain the union of all these subsets, and for k equal to the number of measures taken into account, we obtain the intersection of these subsets.

For a more synthetic comparison of the rankings of the rules by two given measures, we compute a preorder agreement coefficient, τ_1 which is derived from Kendall's τ and was selected amongst different possibilities listed in Giakoumakis and Monjardet [1987].

We intentionally limited our study to measures related to the interestingness of association rules, such as defined in Agrawal et al. [1993]. We detail the selection criterion of the measures in Lenca et al. [2003a]. The 20 measures presented in this paper are implemented in HERBS. They are listed in table 1 and their definitions are reminded in table 2 ($ImpInd^{CR/B}$ corresponds to IMPIND, centred and reduced $-CR$ notation– for a given rule set B).

Table 1: Studied measures

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

Table 2: Studied measures

Measure (Abreviation)	Definition
SUP	$\frac{n_a - n_{a\bar{b}}}{n}$
CONF	$1 - \frac{n_{a\bar{b}}}{n_a}$
R	$\frac{nn_{ab} - n_a n_b}{\sqrt{nn_a n_b n_{\bar{a}} . n_{\bar{b}}}}$
CENCONF	$\frac{nn_{ab} - n_a n_b}{nn_a}$
CONV	$\frac{n_a n_{\bar{b}}}{nn_{a\bar{b}}}$
PS	$\frac{1}{n}(\frac{n_a n_{\bar{b}}}{n} - n_{a\bar{b}})$
LOE	$1 - \frac{nn_{a\bar{b}}}{n_a n_{\bar{b}}}$
IG	$\log(\frac{nn_{ab}}{n_a n_b})$
SEB	$\frac{n_a - n_{a\bar{b}}}{n_{a\bar{b}}}$
LIFT	$\frac{nn_{ab}}{n_a n_b}$
LAP	$\frac{n_{ab}+1}{n_a+2}$
LC	$\frac{n_{ab} - n_{a\bar{b}}}{n_b}$
OM	$\frac{(n_a - n_{a\bar{b}})n_{\bar{b}}}{n_b n_{a\bar{b}}}$
ECR	$\frac{n_a - 2n_{a\bar{b}}}{n_a - n_{a\bar{b}}} = 1 - \frac{1}{\frac{n_a}{n_{a\bar{b}}} - 1}$
KAPPA	$2\frac{nn_a - nn_{a\bar{b}} - n_a n_b}{nn_a + nn_b - 2n_a n_b}$
ZHANG	$\frac{nn_{ab} - n_a n_b}{\max\{n_{ab}n_{\bar{b}}, n_b n_{a\bar{b}}\}}$
-IMPIND	$\frac{nn_{a\bar{b}} - n_a n_{\bar{b}}}{\sqrt{nn_a n_{\bar{b}}}}$
INTIMP	$P\left[poisson(\frac{n_a n_{\bar{b}}}{n}) \geq n_{a\bar{b}}\right]$
EII	$\{[(1 - h_1(\frac{n_{a\bar{b}}}{n})^2) \times (1 - h_2(\frac{n_{a\bar{b}}}{n})^2)]^{1/4}$ INTIMP $\}^{1/2}$
PDI	$P\left[\mathcal{N}(0,1) >$ IMPIND $^{CR/B}\right]$

Where:

$$
\begin{array}{ll}
h_1(t) = -(1 - \frac{t}{p_a})\log_2(1 - \frac{t}{p_a}) - \frac{t}{p_a}\log_2(\frac{t}{p_a}) & \text{for } t \in [0, p_a/2[, \\
h_1(t) = 1 & \text{otherwise;} \\
h_2(t) = -(1 - \frac{t}{p_{\bar{b}}})\log_2(1 - \frac{t}{p_{\bar{b}}}) - \frac{t}{p_{\bar{b}}}\log_2(\frac{t}{p_{\bar{b}}}) & \text{for } t \in [0, p_{\bar{b}}/2[, \\
h_2(t) = 1 & \text{otherwise.}
\end{array}
$$

3 Detailed analysis

We here present a comparative study of 20 measures, carried out with HERBS.

3.1 Analysis of the database used

We studied the *Solarflare* database (UCI Repository; see `ftp.ics.uci.edu/`, Blake and Merz [1998]). It is divided into two case sets described by the same attributes, $\mathcal{SF}_1$ (323 cases) and $\mathcal{SF}_2$ (1066 cases). We used APRIORI [Borgelt and Kruse, 2002] in order to generate a rule set $\mathcal{R}_1$ (5402 rules) from $\mathcal{SF}_1$, and a rule set $\mathcal{R}_2$ (6312 rules) from $\mathcal{SF}_2$, with a minimal support of 20% and a minimal confidence of 85% (see table 3 for descriptive statistics). Many rules tend to lessen the knowledge of `B`. What is more, when a test case database is used on a rule set, many rules no longer have any example at all. We will therefore pre-treat our rule set in order to eliminate such problematic rules.

Table 3: Descriptive statistics of the Solarflare rule sets

couple		number of rules such that:				
$(\mathcal{C}, \mathcal{R})$	η	$n_{a\bar{b}} = 0$	$n_{ab} = 0$	$p_{B/A} \leq p_B$	t_c	t_r
$(\mathcal{SF}_1, \mathcal{R}_1)$	5402	1022	0	1016	100%	1828,6
$(\mathcal{SF}_2, \mathcal{R}_1)$	5402	2195	1378	2192	100%	1536,5
$(\mathcal{SF}_1, \mathcal{R}_2)$	6312	226	118	1080	100%	1427,1
$(\mathcal{SF}_2, \mathcal{R}_2)$	6312	2431	0	1409	100%	2276,6

3.2 Pre-treatment of a rule set

Given a rule set extracted by an APRIORI like algorithm for example, it seems logical to filter this set in order to retain only the rules that strenghten the knowledge of `B` by knowing `A`, *i.e.* rules such that $p_{B/A} > p_B$. We consider that it is wise to go even further in the filtering process of the rules, and keep only those for which the probability of `B` knowing `A` is significantly greater than the probability of `B`, that is to say the rules that are significant from a statistical point of view. It is hence required to compute a critical value for the test of independence of `A` and `B` hypothesis (H_0) against a positive dependence hypothesis (H_1).

Using a hypergeometric model, we test the independency hypothesis of the *itemsets* `A` and `B`, thus n_a and n_b are constants: n_{ab} follows a hypergeometric law, equivalently either $H(n, np_a, p_b)$ or $H(n, p_a, np_b)$ under H_0 (see Lallich [2002]).

We can consider a normal approximation of this hypergeometric law under light restrictions, namely $t_{ab} = \frac{n_a n_b}{n} \geq 5$ and $t_{a\bar{b}} = \frac{n_a n_{\bar{b}}}{n} \geq 5$. Denoting by Φ the centred and reduced normal repartition function $N(0,1)$, $n_{ab\,\text{obs}}$ being the value observed on the data and n_{ab} the theoretical (expected) one, we get:
$\frac{n_{ab} - t_{ab}}{\sqrt{\frac{n_{\bar{a}}}{n-1} t_{ab} p_{\bar{b}}}} \approxeq N(0,1)$, and so $\frac{n_{ab} - t_{ab}}{\sqrt{n p_a p_b p_{\bar{a}} p_{\bar{b}}}} \approxeq N(0,1)$

hence $P(n_{ab} \geq n_{ab\,\text{obs}}) = 1 - \Phi\left(\frac{n_{ab\text{obs}} - t_{ab}}{\sqrt{n p_a p_b p_{\bar{a}} p_{\bar{b}}}}\right) \approxeq 1 - \Phi\left(\sqrt{n}\frac{p_{ab} - p_a p_b}{\sqrt{p_a p_b p_{\bar{a}} p_{\bar{b}}}}\right)$

finally, $P(n_{ab} \geq n_{ab\,\text{obs}}) = 1 - \Phi\left(r\sqrt{n}\right)$

Written this way, it is easy to see that a simple way of filtering the statistically significant rules is to compute the linear correlation coefficient r of the boolean variables A and B, leading to the *p-value* of $n_{a\bar{b}}$, which is $1 - \Phi\left(r\sqrt{n}\right)$. For a classical risk threshold of 0.05, a rule will be statistically significant if $r \geq \frac{1{,}645}{\sqrt{n}}$.

We used the classical *generalisation and test* framework on $\mathcal{R}_1$, and filtered it with the method exposed previously using $\mathcal{SF}_2$ as a test database. We obtained a new rule set $\mathcal{R}_1^2$ composed of 2994 remaining rules coming from $\mathcal{R}_1$. Table 4 summarises the caracteristics of this rule set on the two case databases.

Table 4: Description of $\mathcal{R}_1^2$, evaluated on $\mathcal{SF}_1$ on $\mathcal{SF}_2$

	number of rules such that:				
couple $(\mathcal{C}, \mathcal{R})$	$n_{a\bar{b}} = 0$	$n_{ab} = 0$	$p_{B/A} \leq p_B$	t_c	t_r
$(\mathcal{SF}_1, \mathcal{R}_1^2)$	803	0	155	100%	1043,6
$(\mathcal{SF}_2, \mathcal{R}_1^2)$	951	0	0	100%	1181,4

There are no more rules without examples in the two cases, which is logical. Note that, this was previously not the case in the learning database. What is more, the filtering process rejected all such rules when evaluated on the test database. But some rules still weaken the knowledge of B for the learning database, although their number has greatly disminished. We also see that the number of rules with no counter-examples (logical rules) has disminished. Last but not least, the coverage rate remains at 100%, and the recovering rate is still high too. Indeed, there are many rules with a high support value.

3.3 Experimental comparison of the behaviour of quality measures

We used this filtered rule set in order to study the behaviour of quality measures on real data. The aim of this study is to confront previous results obtained from a formal study of the measures (see Lenca et al. [2003a]) with experimental ones. The behaviour differences between the measures are presented in tables 5 and 6. Each block is proportional to the agreement coefficient τ_1 of the preorders induced by the measures on $(\mathcal{SF}_1, \mathcal{R}_1^2)$ and $(\mathcal{SF}_2, \mathcal{R}_1^2)$. The rows and the columns of the two matrices have been reorganised in order to highlight the block structures, using the AMADO method [?]. There seem to be 4 well separated groups of measures. Within these groups, more subtle agreements appear, some of which can be easily explained since measures are monotonic decreasing

transformations of one another (for example CONF, SEB, ECR and really close to them is LAP which does not differ much). The surprising proximity between SUP and LC is mostly related to a bias introduced by the APRIORI algorithm. Indeed, the support and confidence thresholds used confine the possible values of n_a and n_b to a small range of values, and the two measures differ mostly when this range is wide.

Table 5: Preorder comparison for $(\mathcal{SF}_1, \mathcal{R}_1^2)$

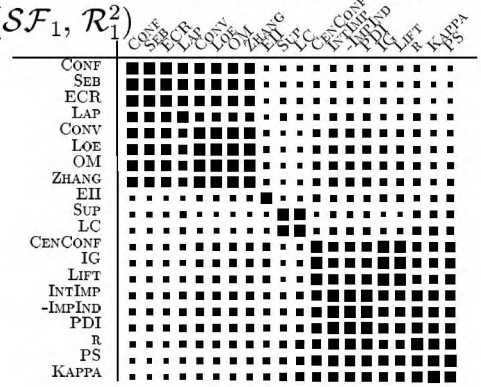
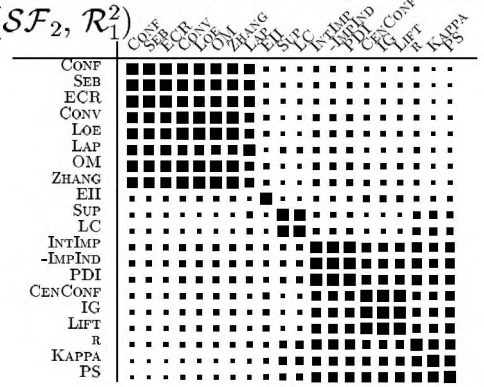


Table 6: Preorder comparison for $(\mathcal{SF}_2, \mathcal{R}_1^2)$



4 More results

Similar experiments were carried out on other databases retrieved from the same repository, without however splitting the databases. It was threfore not possible to first learn the rules and then undertake a generalisation test on previously unseen data. Moreover, the statistical filtering of the rule sets was not done.

The parameters of the APRIORI algorithm were fixed experimentally in order to obtain rule sets of an acceptable size in terms of computational cost. The caracteristics of these sets are summarized in table 7. 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* and *Cmc2* (*cmc2* was obtained by filtering *cmc* and retaining only the rules having a lift greater or equal to $1,2$): 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.

We also computed the values of τ_1 for the *solarflare* couples $(\mathcal{SF}_1, \mathcal{R}_1)$, $(\mathcal{SF}_2, \mathcal{R}_2)$, $(\mathcal{SF}_1, \mathcal{R}_1^1)$, $(\mathcal{SF}_1, \mathcal{R}_1^2)$ and $(\mathcal{SF}_2, \mathcal{R}_1^2)$. Finally, we generated 10 preorder comparison matrices. Table 8 summarizes all these results (the value of τ_1 is proportional to the radius of the corresponding portion of disc). The AMADO method was applied to the average matrix of the results. The results are in fairly good agreement: we can still make out 4 groups of measures, although

some slight differences appear depending on which database is considered. This justifies once again the need for empirical studies of the properties and behaviour of the measures.

Table 7: Summary of the different rule sets used

name	n	sup_{min}	$conf_{min}$	η	t_c	t_r
Autompg	392	5	50	49	81,37%	0,79
Breastcancerwisconsin	683	10	70	3095	96,19%	646
Car	1728	5	60	145	80,09%	13,75
Cmc	1473	5	60	2878	100%	259
Cmc2	1473	5	60	825	92,93%	84

5 Experimental *vs.* formal approach

In order to have a better understanding of the behaviour of the measures, we compared the typology in 4 classes coming from our experiments with the formal typology developed in Lenca et al. [2003a]. The formal approach can be synthesised with a 20×8 decision matrix, containing the evaluation of the 20 measures on the 8 criteria (see table 9). We kept only 6 of the criteria for the comparison, as two of them – namely g_7 and g_8 – do not influence the experimental results at all:

g_1: **asymmetric processing of A and B [Freitas, 1999].** Since A and B 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 which do not.

g_2: **decrease with n_b [Piatetsky-Shapiro, 1991].** 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 but not A increases, the interestingness of the rule should decrease.

g_3: **reference situations: independence [Piatetsky-Shapiro, 1991].** To avoid keeping rules that contain no information, it is necessary to eliminate the $A \rightarrow B$ rule when A and B are independent, meaning when 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.

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.

Table 8: Preorder comparisons for 20 measures for 10 couples $(\mathcal{C}, \mathcal{R})$.

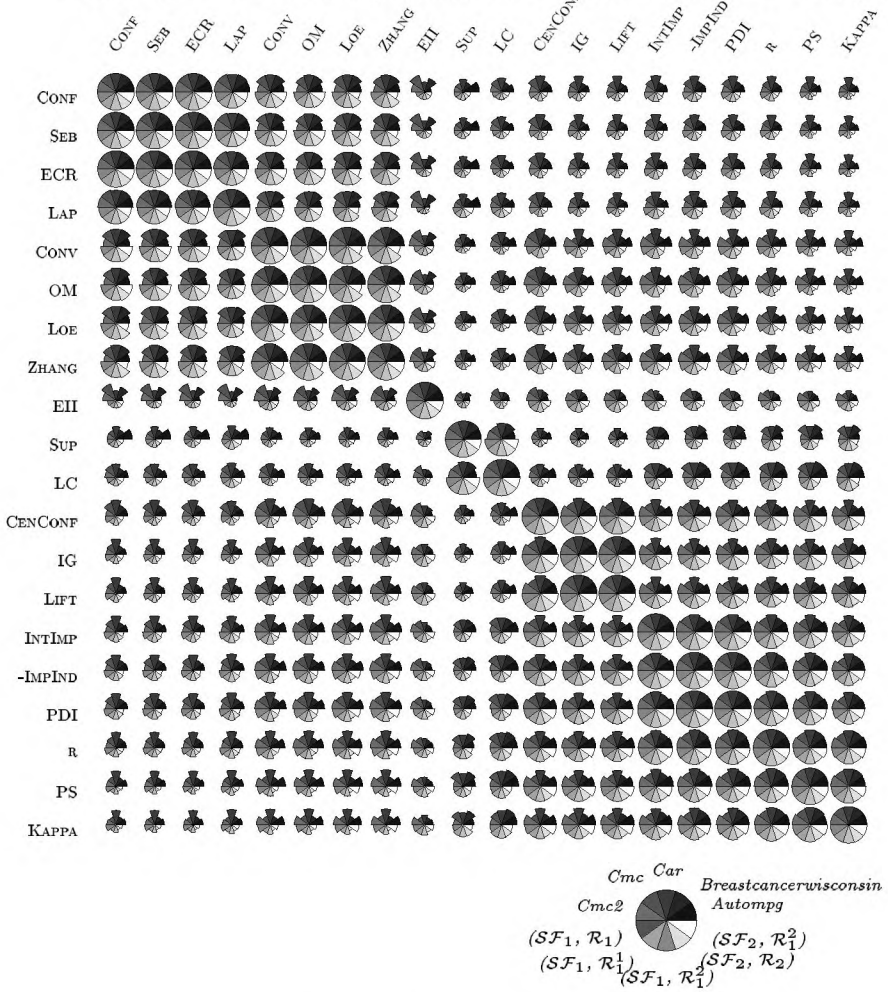


Table 9: Properties of the measures

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

g_5**: linearity with** $n_{a\bar{b}}$ **around** 0^+**.** Gras et al. [2002] express the desire to have a weak decrease in the neighbourhood 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.

g_6**: sensitivity to** n **(total number of records).** Intuitively, if the rates of presence of A, A $\rightarrow$ 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). 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.

The 20×6 matrix hence obtained was re-encoded in a 20×13 normal disjunctive matrix composed of boolean values. These values do not represent any judgement on the measures, but only list the properties shared by the different measures. The typology in 4 classes (see table 10) coming from this matrix is obtained with a hierarchical ascendent clustering, following the WARD criterion, applied to the square of the euclidian distance (that is, in our case, twice the number of differing properties).

Table 10 shows that both approaches globally lead to similar clusterings, but some shifts are interesting. The first class, {PS, KAPPA, IG, CENCONF, LIFT, R, -IMPIND, PDI} is the same in both cases. What is more, three groups of measures are stable: {LOE, ZHANG, OM, CONV}, {CONF, SEB, ECR} and {SUP, LC}. Within these groups the measures are similar both from the formal and experimental point of view.

The main difference is the presence of a third class in the experimental approach that spans over classes 2, 3 and 4 of the formal clustering. The behaviour of the measures LOE, ZHANG, OM and CONV (formal class 2) is close to that of CONF, SEB and ECR (formal class 3). However, INTIMP and EII that also belong to formal class 2 behave differently. INTIMP shifts to the experimental class 1 (with LIFT and CENCONF), and EII has an original behaviour. These differences strengthen our formal analysis since EII and INTIMP cannot be distinguished with our formal critera. EII is defined as the product of INTIMP with an inclusion index whose role is to make EII more discriminant. This explains the experimental differences.

LAP shifts to LC and SUP (class 2), LOE, ZHANG, OM and CONV shift to INTIMP and EII (class 4), whereas the core of class 3 consists of CONF, SEB and ECR. The reasons are that LOE, ZHANG, OM and CONV have many properties in common with INTIMP and EII (g_1, g_2, g_3, g_4), which is not the

case for CONF, SEB and ECR for g_2 and g_3. However, these 3 measures verify properties g_1 and g_4.

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 made up of. Another reason for these shifts is that LAP, really close to SUP in our formal study, can differ from CONF experimentally only for values of n_a close to 0 (nuggets). The minimum thresholds of the APRIORI algorithms used make this impossible, and this can be seen as an algorithmic bias.

Table 10: Cross-classification of the measures

Formal \ Experimental	Class 1	Class 2	Class 3	Class 4
Class 1	PS, KAPPA, IG CENCONF, LIFT, R -IMPIND, PDI			
Class 2	INTIMP	EII	LOE, ZHANG, OM, CONV	
Class 3			CONF, SEB, ECR	
Class 4			LAP	LC, SUP

6 Conclusion

Association rule quality measures play a major role within a KDD process, but they have a large diversity of properties, that have to be studied on real data in order to use a measure adapted to the experimental context. We have presented results coming from a tool we developed, HERBS, and compared the behaviour of 20 measures on 10 datasets. From these experimental results, we were able to identify 4 main groups of measures. This clustering was then compared to a clustering coming from a formal analysis we had done previously. The experimental approach seems to be a important addition to the formal approach. Indeed, it has 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. Although this may be of importance in some fields, such as medicine, it may be of no importance in others.

References

Agrawal, R., Imielinski, T., Swami, A., 1993. Mining association rules between sets of items in large databases. In: Buneman, P., Jajodia, S. (Eds.), ACM

SIGMOD Int. Conf. on Management of Data. pp. 207–216.

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

Azé, J., Kodratoff, Y., 2002. A study of the effect of noisy data in rule extraction systems. In: Sixteenth European Meeting on Cybernetics and Systems Research. Vol. 2. pp. 781–786.

Blake, C., Merz, C., 1998. Uci repository of machine learning databases. http://www.ics.uci.edu/~mlearn/MLRepository.html.

Borgelt, C., Kruse, R., 2002. Induction of association rules: APRIORI implementation. In: 15th Conf. on Computational Statistics.

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

Brin, S., Motwani, R., Ullman, J. D., Tsur, S., 1997b. Dynamic itemset counting and implication rules for market basket data. In: Peckham, J. (Ed.), ACM SIGMOD 1997 Int. Conf. on Management of Data. pp. 255–264.

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 KDD. AAAI/MIT Press.

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

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

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

Gras, R., Ag. Almouloud, S., Bailleuil, M., Larher, A., Polo, M., Ratsimba-Rajohn, H., Totohasina, A., 1996. L'implication Statistique, Nouvelle Méthode Exploratoire de Données. Application 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. EGC 2001 1 (1-2), 69–80.

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

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

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

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

Lenca, P., Meyer, P., Picouet, P., Vaillant, B., Lallich, S., 2003a. Critères d'évaluation des mesures de qualité en ECD. RNTI (1), 123–134.

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

Lerman, I., Azé, J., 2003. Une mesure probabiliste contextuelle discriminante de qualité des règles d'association. 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.

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

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

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

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

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.), EKAW'88. pp. 28–1 – 28–20.

Tan, P.-N., Kumar, V., Srivastava, J., 2002. Selecting the right interestingness measure for association patterns. In: Eighth ACM SIGKDD Int. Conf. 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., Picouet, P., Lenca, P., Mai 2003. An extensible platform for rule quality measure benchmarking. In: Bisdorff, R. (Ed.), HCP'2003. pp. 187–191.

Acknowledgments

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