A Graph-based Clustering Approach to Evaluate Interestingness Measures: A Tool and a Comparative Study

Xuan-Hiep Huynh, Fabrice Guillet, Julien Blanchard, Pascale Kuntz, Henri Briand, Régis Gras

▶ **To cite this version:**

Xuan-Hiep Huynh, Fabrice Guillet, Julien Blanchard, Pascale Kuntz, Henri Briand, et al.. A Graph-based Clustering Approach to Evaluate Interestingness Measures: A Tool and a Comparative Study. Fabrice Guillet, Howard J. Hamilton. Quality Measures in Data Mining, Springer, pp.25-50, 2007, Studies in Computational Intelligence, 10.1007/978-3-540-44918-8_2 . hal-00420991

HAL Id: hal-00420991
https://hal.science/hal-00420991

Submitted on 30 Sep 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A graph-based clustering approach to evaluate interestingness measures : a tool and a comparative study

Xuan-Hiep Huynh, Fabrice Guillet, Julien Blanchard, Pascale Kuntz, Henri Briand, and Régis Gras

LINA CNRS 2729 - Polytechnic School of Nantes University, La Chantrerie BP 50609 44306 Nantes cedex 3, France {1stname.name}@polytech.univ-nantes.fr

Summary. Finding interestingness measures to evaluate association rules has become an important knowledge quality issue in KDD. Many interestingness measures may be found in the literature, and many authors have discussed and compared interestingness properties in order to improve the choice of the most suitable measures for a given application. As interestingness depends both on the data structure and on the decision-maker's goals, some measures may be relevant in some context, but not in others. Therefore, it is necessary to design new contextual approaches in order to help the decision-maker select the most suitable interestingness measures. In this paper, we present a new approach implemented by a new tool, ARQAT, for making comparisons. The approach is based on the analysis of a correlation graph presenting the clustering of objective interestingness measures and reflecting the post-processing of association rules. This graph-based clustering approach is used to compare and discuss the behavior of thirty-six interestingness measures on two prototypical and opposite datasets: a highly correlated one and a lowly correlated one. We focus on the discovery of the stable clusters obtained from the data analyzed between these thirty-six measures.

1 Introduction

As the number of discovered rules increases, end-users, such as data analysts and decision makers, are frequently confronted with a major challenge: how to validate and select the most interesting of those rules. Over the last decade the Knowledge Discovery in Databases (KDD) community has recognized this challenge – often referred to as interestingness – as an important and difficult component of the KDD process (Klemettinen et al. [15], Tan et al. [30]). To tackle this problem, the most commonly used approach is based on the construction of Interestingness Measures (IM).

In defining association rules, Agrawal et al. [1] [2] [3], introduced two IMs : support and confidence. These are well adapted to Apriori algorithm con-

straints, but are not sufficient to capture the whole aspects of the rule interestingness. To push back this limit, many complementary IMs have been then proposed in the literature (see [5] [14] [30] for a survey). They can be classified in two categories [10]: subjective and objective. Subjective measures explicitly depend on the user's goals and his/her knowledge or beliefs. They are combined with specific supervised algorithms in order to compare the extracted rules with the user's expectations [29] [24] [21]. Consequently, subjective measures allow the capture of rule novelty and unexpectedness in relation to the user's knowledge or beliefs. Objective measures are numerical indexes that only rely on the data distribution.

In this paper, we present a new approach and a dedicated tool ARQAT (Association Rule Quality Analysis Tool) to study the specific behavior of a set of 36 IMs in the context of a specific dataset and in an exploratory analysis perspective, reflecting the post-processing of association rules. More precisely, ARQAT is a toolbox designed to help a data-analyst to capture the most suitable IMs and consequently, the most interesting rules within a specific ruleset.

We focus our study on the objective IMs studied in surveys [5] [14] [30]. The list of IMs is added with four complementary IMs (Appendix A): Implication Intensity (II), Entropic Implication Intensity (EII), TIC (information ratio modulated by contra-positive), and IPEE (probabilistic index of deviation from equilibrium). Furthermore, we present a new approach based on the analysis of a correlation graph (CG) for clustering objective IMs.

This approach is applied to compare and discuss the behavior of 36 IMs on two prototypical and opposite datasets: a strongly correlated one (mushroom dataset [23]) and a lowly correlated one (synthetic dataset). Our objective is to discover the stable clusters and to better understand the differences between IMs.

The paper is structured as follows. In Section 2, we present related works on objective IMs for association rules. Section 3 presents a taxonomy of the IMs based on two criteria: the "subject" (deviation from independence or equilibrium) of the IMs and the "nature" of the IMs (descriptive or statistical). In Section 4, we introduce the new tool ARQAT for evaluating the behavior of IMs. In Section 5, we detail the correlation graph clustering approach. And, Section 6 is dedicated to a specific study on two prototypical and opposite datasets in order to extract the stable behaviors.

2 Related works on objective IMs

The surveys on the objective IMs mainly address two related research issues : *(1)* defining a set of principles or properties that lead to the design of a good IM, *(2)* comparing the IM behavior from a data-analysis point of view. The results yielded can be useful to help the user select the suitable ones.

Considering the principles of a good IM issue, Piatetsky-Shapiro [25] introduced the Rule-Interest, and proposed three underlying principles for a good IM on a rule $a \rightarrow b$ between two itemsets a and b : 0 value when a and b are independent, monotonically increasing with a *and* b, monotonically decreasing with a or b. Hilderman and Hamilton [14] proposed five principles: minimum value, maximum value, skewness, permutation invariance, transfer. Tan et al. [30] defined five interestingness principles: symmetry under variable permutation, row/column scaling invariance, anti-symmetry under row/column permutation, inversion invariance, null invariance. Freitas [10] proposed an "attribute surprisingness" principle. Bayardo and Agrawal [5] concluded that the most interesting rules according to some selected IMs must reside along a support/confidence border. The work allows for improved insight into the data and supports more user-interaction in the optimized rule-mining process. Kononenko [19] analyzed the biases of eleven IMs for estimating the quality of multi-valued attributes. The values of information gain, J-measure, Gini-index, and relevance tend to linearly increase with the number of values of an attribute. Zhao and Karypis [33] used seven different criterion functions with clustering algorithms to maximize or minimize a particular one. Gavrilov et al. [11] studied the similarity measures for the clustering of similar stocks. Gras et al. [12] discussed a set of ten criteria: increase, decrease with respect to certain expected semantics, constraints for semantics reasons, decrease with trivial observations, flexible and general analysis, discriminative residence with the increment of data volume, quasi-inclusion, analytical properties that must be countable, two characteristics of formulation and algorithms.

Some of these surveys also address the related issue of the IM comparison by adopting a data-analysis point of view. Hilderman and Hamilton [14] used the five proposed principles to rank summaries generated from databases and used sixteen diversity measures to show that: six measures matched five proposed principles, and nine remaining measures matched at least one proposed principle. By studying twenty-one IMs, Tan et al. [30] showed that an IM cannot be adapted to all cases and use both a support-based pruning and standardization methods to select the best IMs; they found that, in some cases many IMs are highly correlated with each other. Eventually, the decision-maker will select the most suitable measure by matching the five proposed properties. Vaillant et al. [31] evaluated twenty IMs to choose a user-adapted IM with eight properties: asymmetric processing of a and b for an association rule $a \rightarrow b$, decrease with n_b, independence, logical rule, linearity with $n_{a\bar{b}}$ around 0^+, sensitivity to n, easiness to fix a threshold, intelligibility. Finally, Huynh et al. [16] introduced the first result of a new clustering approach for classifying thirty-four IMs with a correlation analysis.

3 A taxonomy of objective IMs

In this section, we propose a taxonomy of the objective IMs (details in Appendixes A and B) according to two criteria: the "subject" (deviation from independence or equilibrium), and the "nature" (descriptive or statistical). The conjunction of these criteria seems to us essential to grasp the meaning of the IMs, and therefore to help the user choose the ones he/she wants to apply.

In the following, we consider a finite set T of transactions. We denote an association rule by $a \rightarrow b$ where a and b are two disjoint itemsets. The itemset a (respectively b) is associated with a transaction subset $A = T(a) = \{t \in T, a \subseteq t\}$ (respectively $B = T(b)$). The itemset $\overline{a}$ (respectively $\overline{b}$) is associated with $\overline{A} = T(\overline{a}) = T - T(a) = \{t \in T, a \not\subseteq t\}$ (respectively $\overline{B} = T(\overline{b})$). In order to accept or reject the general trend to have b when a is present, it is quite common to consider the number $n_{a\overline{b}}$ of negative examples (contra-examples, counter-examples) of the rule $a \rightarrow b$. However, to quantify the "surprisingness" of this rule, consider some definitions are functions of $n = |T|$, $n_a = |A|$, $n_b = |B|$, $n_{\overline{a}} = |\overline{A}|$, $n_{\overline{b}} = |\overline{B}|$.

Let us denote that, for clarity, we also keep the probabilistic notations $p(a)$ (respectively $p(b)$, $p(a \textit{ and } b)$, $p(a \textit{ and } \overline{b})$) as the probability of a (respectively b, a *and* b, a *and* $\overline{b}$). This probability is estimated by the frequency of a: $p(a) = \frac{n_a}{n}$ (respectively $p(b) = \frac{n_b}{n}$, $p(a \textit{ and } b) = \frac{n_{ab}}{n}$, $p(a \textit{ and } \overline{b}) = \frac{n_{a\overline{b}}}{n}$).

3.1 Subject of an IM

Generally speaking, an association rule is more interesting when it is supported by lots of examples and few negative examples. Thus, given n_a, n_b and n, the interestingness of $a \rightarrow b$ is minimal when $n_{a\overline{b}} = \min(n_a, n_{\overline{b}})$ and maximal when $n_{a\overline{b}} = \max(0, n_a - n_b)$. Between these extreme situations, there exist two significant configurations in which the rules appear non-directed relations and therefore can be considered as neutral or non-existing: the independence and the equilibrium. In these configurations, the rules are to be discarded.

Independence

Two itemsets a and b are independent if $p(a \textit{ and } b) = p(a) \times p(b)$, i.e. $n.n_{a\overline{b}} = n_a n_{\overline{b}}$. In the independence case, each itemset gives no information about the other, since knowing the value taken by one of the itemsets does not alter the probability distribution of the other itemset: $p(b \backslash a) = p(b \backslash \overline{a}) = p(b)$ and $p(\overline{b} \backslash a) = p(\overline{b} \backslash \overline{a}) = p(\overline{b})$ (the same for the probabilities of a and $\overline{a}$ given b or $\overline{b}$). In other words, knowing the value taken by an itemset lets our uncertainty about the other itemset intact. There are two ways of deviating from the independent situation: either the itemsets a and b are positively correlated ($p(a \textit{ and } b) > p(a) \times p(b)$), or they are negatively correlated ($p(a \textit{ and } b) < p(a) \times p(b)$).

Equilibrium

We define the equilibrium of a rule $a \rightarrow b$ as the situation where examples and negative examples are equal in numbers: $n_{ab} = n_{a\overline{b}} = \frac{1}{2}n_a$ [7]. In this situation, the itemset a is as concomitant with b as with $\overline{b}$ in the data. So a rule $a \rightarrow b$ at equilibrium is as directed towards b as towards $\overline{b}$. There are two ways of deviating from this equilibrium situation: either a is more concomitant with b than with $\overline{b}$, or a is more concomitant with $\overline{b}$ than with b.

Deviation from independence and from equilibrium

As there exist two different notions of neutrality, the objective interestingness of association rules must be measured from (at least) two complementary points of view: the deviation from independence, and the deviation from equilibrium. These are what we call the two possible subjects for the rule IMs. These deviations are directed in favor of examples and in disfavor of negative examples.

Definition 1. An IM m evaluates a deviation from independence if the IM has a fixed value at the independence:

$$m(n, n_a, n_b, \frac{n_a n_{\overline{b}}}{n}) = constant$$

Definition 2. An IM m evaluates a deviation from equilibrium if the IM has a fixed value at the equilibrium:

$$m(n, n_a, n_b, \frac{n_a}{2}) = constant$$

Independence is a function of four parameters n, n_a, n_b and $n_{a\overline{b}}$[1], whereas equilibrium is a function of the two parameters n_a and $n_{a\overline{b}}$. Thus, all the IMs of deviation from independence depend on the four parameters, while the IMs of deviation from equilibrium do not depend on n_b and n generally. The only exceptions to this principle are IPEE [7] and the Least Contradiction [4]. IPEE (see the formula in Appendix A) measures the statistical significance of the deviation from equilibrium. It depends on n. The Least Contradiction depends on n_b (see the formula in Appendix B). This is a hybrid IM which has a fixed value at equilibrium – as the IMs of deviation from equilibrium – but decreases with n_b – as the IMs of deviation from independence.

Comparison of the filtering capacities

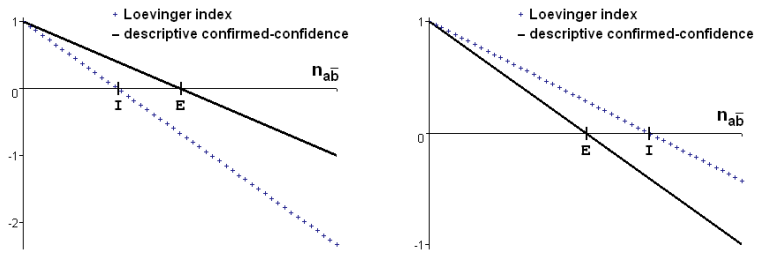
We aim at filtering the rules with a threshold on the IMs (by retaining only the high values of the IMs), and at comparing the numbers of rules that are

[1] Here we have chosen $n_{a\overline{b}}$ as a parameter, but we could have chosen another cardinality of the joint distribution of the itemsets a and b, such as n_{ab}.

rejected by the IMs of deviation from equilibrium and from independence. Let us consider a rule with the cardinalities n, n_a, n_b, and $n_{a\bar{b}}$. By varying $n_{a\bar{b}}$ with fixed n, n_a, and n_b, one can distinguish two different cases:

- Case 1: $n_b \geq \frac{n}{2}$. Then $\frac{n_a n_{\bar{b}}}{n} \leq \frac{n_a}{2}$, and the rule goes through the independence before going through the equilibrium when $n_{a\bar{b}}$ increases.
- Case 2: $n_b \leq \frac{n}{2}$. Then $\frac{n_a n_{\bar{b}}}{n} \geq \frac{n_a}{2}$, and the rule goes through the equilibrium before going through the independence when $n_{a\bar{b}}$ increases.

Let us now compare an IM of deviation from equilibrium m_{eql} and an IM of deviation from independence m_{idp} for these two cases. In order to have a fair comparison, we suppose that the two IMs have similar behaviors: same value for a logical rule, same value for equilibrium/independence, same decrease speed with regard to the negative examples. For example, m_{eql} and m_{idp} can be the Descriptive Confirmed-Confidence [18] and the Loevinger index respectively [22] (Appendix B). As shown in figure 1, m_{idp} is more filtering than m_{eql} in case 1, whereas m_{eql} is more filtering than m_{idp} in case 2. More precisely, in case 1, m_{idp} contributes to rejecting the bad rules, while in case 2 it is m_{eql}. This confirms that the IMs of deviation from equilibrium and the IMs of deviation from independence are complementary, the second ones not being systematically "better" than the first ones[2]. In particular, the IMs of deviation from equilibrium must not be neglected when itemsets are rare (low frequency). In this situation, case 2 is more frequent than case 1.



(a) case 1 ($n_b \geq \frac{n}{2}$) (b) case 2 ($n_b \leq \frac{n}{2}$)

Fig. 1. Comparison of Descriptive Confirmed-Confidence and Loevinger index (E: equilibrium, I: independence)

In our IM taxonomy, the subject of an IM could be the deviation from independence or the deviation from equilibrium. However, as some IMs do not assess any of the two deviation, a third cluster must be added ("other measures" in Tab. 1). The IMs of this cluster generally have a fixed value

[2] Numerous authors consider that a good IM must vanish at independence (principle originally proposed in [25]). This amounts to saying that IMs of deviation from independence are better than IMs of deviation from equilibrium.

only for the rules with no negative examples ($n_{a\bar{b}} = 0$) or for the rules with no examples ($n_{ab} = 0$). Most of them are similarity measures.

3.2 Nature of an IM

The objective IMs can also be classified according to their descriptive or statistical nature.

Descriptive IMs

The descriptive (or frequential) IMs do not vary with the cardinality expansion (when all the data cardinalities are increased or decreased in equal proportion). A descriptive IM m satisfies $m(n, n_a, n_b, n_{a\bar{b}}) = m(\alpha.n, \alpha.n_a, \alpha.n_b, \alpha.n_{a\bar{b}})$ for any strictly positive constant α. These IMs take the data cardinalities into account only in a relative way (by means of the frequencies $p(a)$, $p(b)$, $p(a \; and \; \bar{b})$) and not in an absolute way (by means of the cardinalities n_a, n_b, $n_{a\bar{b}}$).

Statistical IMs

The statistical IMs are those which vary with the cardinality expansion. They take into account the size of the phenomena studied. Indeed, a rule is statistically more valid when it is accessed on a large amount of data. Among the statistical IMs, one can find the probabilistic IMs, which compare the observed distribution to an expected distribution, such as the II measure presented in Appendix A.

3.3 IM taxonomy

A taxonomy according to the nature and subject of the objective IMs is given below (Tab. 1). On the column, we can see that most of the IMs are descriptive. Another observation shows that IPEE is the only one statistical IM computing the deviation from equilibrium.

4 ARQAT tool

ARQAT (Fig. 2) is an exploratory analysis tool that embeds thirty-six objective IMs studied in surveys (See Appendix B for a complete list of selected IMs).

It provides graphical views structured in five task-oriented groups: ruleset analysis, correlation and clustering analysis, interesting rules analysis, sensitivity analysis, and comparative analysis.

Nature / Subject	Descriptive IMs	Statistical IMs
Measures of deviation from equilibrium	– Confidence (5), – Laplace (21), – Sebag & Schoenauer (31), – Example & Contra-Example (13), – Descriptive Confirm (9), – Descriptive Confirmed-Confidence (10), – Least Contradiction (22)	– IPEE (16)
Measures of deviation from independence	– Phi-Coefficient (28), – Lift (23), – Loevinger (25), – Conviction (6), – Dependency (8), – Pavillon (27), – J-measure (18), – Gini-index (14), – TIC (33), – Collective Strength (4), – Odds Ratio (26), – Yules's Q (34), – Yule's Y (35), – Klosgen (20), – Kappa (19)	– II (15), – EII$\alpha = 1$ (11), – EII$\alpha = 2$ (12), – Lerman (24), – Rule Interest (30)
Other measures	– Support (32), – Causal Support (3), – Jaccard (17), – Cosine (7), – Causal Confidence (0), – Causal Confirm (1), – Causal Confirmed-Confidence (2), – Putative Causal Dependency (29)	

Table 1. Taxonomy of the objective IMs

The ARQAT input is a set of association rules R where each association rule $a \rightarrow b$ must be associated with the four cardinalities n, n_a, n_b, and $n_{a\overline{b}}$.

In the first stage, the input ruleset is preprocessed in order to compute the IM values of each rule, and the correlations between all IM pairs. The results are stored in two tables: an IM table (R×I) where rows are rules and columns are IM values, and a correlation matrix (I×I) crossing IMs. At this stage, the ruleset may also be sampled (filtering box in Fig. 2) in order to focus the study on a more restricted subset of rules.

In the second stage, the data-analyst can drive the graphical exploration of results through a classical web-browser. ARQAT is structured in five groups of task-oriented views. The first group (1 in Fig. 2) is dedicated to ruleset and simple IM statistics to better understand the structure of the IM table (R×I). The second group (2) is oriented to the study of IM correlation in table (I×I) and IM clustering in order to select the most suitable IMs. The third one (3) focuses on rule ordering to select the most interesting rules. The fourth group (4) proposes to study the sensitivity of IMs. The last group (5) offers the possibility to compare the results obtained from different rulesets.

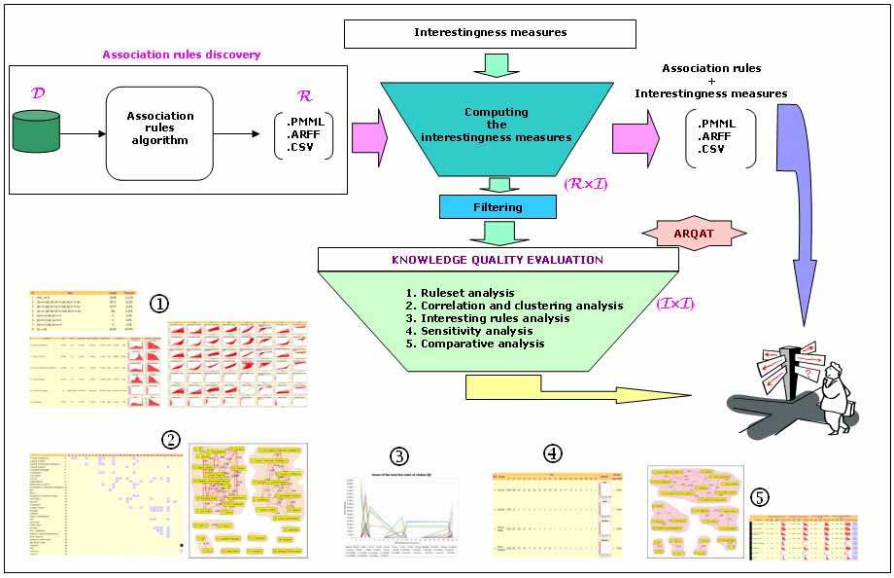


Fig. 2. ARQAT structure.

In this section, we focus on the description of the first three groups and we illustrate them with the same ruleset: 123228 association rules extracted by Apriori algorithm (support 12%) from the mushroom dataset [23].

4.1 Ruleset statistics

The basic statistics are summarized on three views of ARQAT. The first one, ruleset characteristics, shows the distributions underlying rule cardinalities, in order to detect "borderline cases". For instance, in Tab. 2, the first line gives the number of "logical" rules i.e. rules without negative examples. We can notice that the number of logical rules is here very high (≈13%).

N	Type	Count	Percent
1	$n_{a\bar{b}} = 0$	16158	13.11%
2	$n_{a\bar{b}} = 0$ & $na < nb$	15772	12.80%
3	$n_{a\bar{b}} = 0$ & $na < nb$ & $nb = n$	0	0.00%
4	$n_a > n_b$	61355	49.79%
5	$nb = n$	0	0.00%

Table 2. Some ruleset characteristics of the mushroom ruleset.

The second view, IM distribution (Fig. 3), draws the histograms for each IM. The distributions are also completed with classically statistical indexes :

minimum, maximum, average, standard deviation, skewness and kurtosis values. In Fig. 3, one can see that Confidence (line 5) has an irregular distribution and a great number of rules with 100% confidence; it is very different from Causal Confirm (line 1).

The third view, joint-distribution analysis (Fig. 4), shows the scatterplot matrix of all IM pairs. This graphical matrix is very useful to see the details of the relationships between IMs. For instance, Fig. 4 shows four disagreement shapes: Rule Interest vs Yule's Q (4), Sebag & Schoenauer vs Yule's Y (5), Support vs TIC (6), and Yule's Y vs Support (7) (highly uncorrelated). On the other hand, we can notice four agreement shapes on Phi-Coefficient vs Putative Causal Dependency (1), Phi-Coefficient vs Rule Interest (2), Putative Causal Dependency vs Rule Interest (3), and Yule's Q vs Yule's Y (8) (highly correlated).

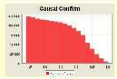
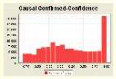
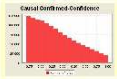
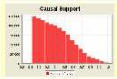
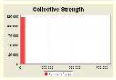
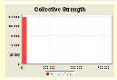
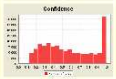
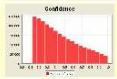
N°	Measure	Min	Max	Average	Std. Deviation	Skewness		Kurtosis		Histogram	Inverse-cumulative
0	Causal Confidence	0.0614	1.0	0.6246	0.2712	-0.2033	Left	-1.0001	Flat		
1	Causal Confirm	-1.5951	1.0	0.1084	0.5744	-1.1203	Left	0.8805	Peaked		
2	Causal Confirmed-Confidence	-0.8169	1.0	0.2018	0.5448	0.0286	Right	-1.2036	Flat		
3	Causal Support	0.1198	1.0	0.5542	0.2051	-0.0995	Left	-0.5643	Flat		
4	Collective Strength	0.0	785856.8441	2328.2323	7944.9114	30.5993	Right	1932.4436	Peaked		
5	Confidence	0.1217	1.0	0.5772	0.2799	0.194	Right	-1.3159	Flat		
6	Conviction	0.0	219.4653	2.1238	6.3156	14.7085	Right	287.2374	Peaked		

Fig. 3. Distribution of some IMs on the mushroom dataset.

4.2 Correlation analysis

This task aims at delivering IM clustering and facilitating the choice of a subset of IMs that is best-adapted to describe the ruleset. The correlation values between IM pairs are computed in the preprocessing stage by using the Pearson's correlation coefficient and stored in the correlation matrix (I×I). Two visual representations are proposed. The first one is a simple summary matrix in which each significant correlation value is visually associated with a

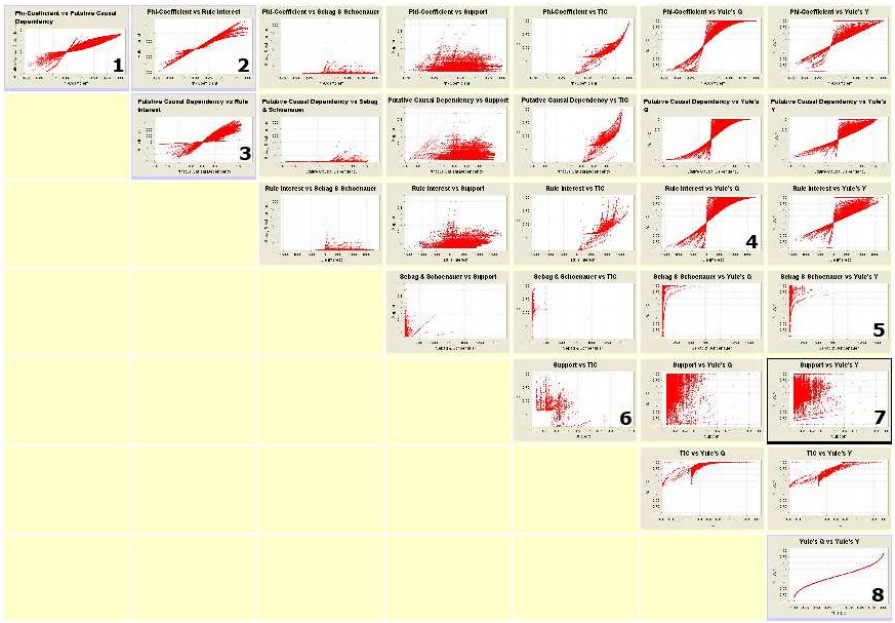


Fig. 4. Scatterplot matrix of joint-distributions on the mushroom dataset.

different color (a level of gray). For instance, the furthest right dark cell from Fig. 5 shows a low correlation value between Yule's Y and Support. The other seventy-nine gray cells correspond to high correlation values.

The second one (Fig. 6) is a graph-based view of the correlation matrix. As graphs are a good means to offer relevant visual insights on data structure, the correlation matrix is used as the relation of an undirected and valued graph, called "correlation graph". In a correlation graph, a vertex represents an IM and an edge value is the correlation value between two vertices/IMs. We also add The possibility to set a minimal threshold τ (maximal threshold θ respectively) to retain only the edges associated with a high correlation (respectively low correlation); the associated subgraphs are denoted by CG+ and CG0.

These two subgraphs can then be processed in order to extract clusters of IMs: each cluster is defined as a connected subgraph. In CG+, each cluster gathers correlated or anti-correlated IMs that may be interpreted similarly: they deliver a close point of view on data. Moreover, in CG0, each cluster contains uncorrelated IMs: i.e. IMs that deliver a different point of view.

Hence, as each graph depends on a specific ruleset, the user can use the graphs as data insight, which graphically help him/her select the minimal set of the IMs best adapted to his/her data. For instance in Fig. 6, CG+ graph contains twelve clusters on thirty-six IMs. The user can select the most representative IM in each cluster, and then retain it to validate the rules.

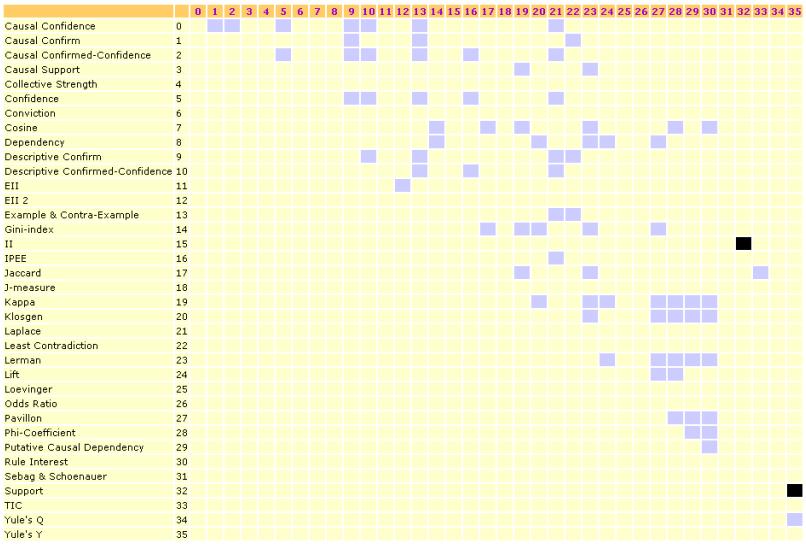


Fig. 5. Summary matrix of correlations on the mushroom dataset.

A close observation on the CG0 graph (Fig. 6) shows an uncorrelated cluster formed by II, Support and Yule's Y measures (also the two dark cells in Fig. 5). This observation is confirmed on Fig. 4 (7). CG+ graph shows a trivial cluster where Yule's Q and Yule's Y are strongly correlated. This is also confirmed in Fig. 4 (8), showing a functional dependency between the two IMs. These two examples show the interest of using the scatterplot matrix complementarily (Fig. 4) with the correlation graphs CG0, CG+ (Fig. 6) in order to evaluate the nature of the correlation links, and overcome the limits of the correlation coefficient.

4.3 Interesting rule analysis

In order to help a user select the most interesting rules, two specific views are implemented. The first view (Fig. 7) collects a set of a given number of interesting rules for each IM in one cluster, in order to answer the question: how interesting are the rules of this cluster?. The selected rules can alternatively be visualized with parallel coordinate drawing (Fig. 8). The main interest of such a drawing is to rapidly see the IM rankings of the rules.

These two views can be used with the IM values of a rule or alternatively with the rank of the value. For instance, Fig. 7 and Fig. 8 use the rank to evaluate the union of the ten interesting rules for each of the ten IMs in the C0 cluster (see Fig. 6). The Y-axis in Fig. 8 holds the rule rank for the corresponding IM. By observing the concentration lines on low rank values, one can obtain four IMs: Confidence(5), Descriptive Confirmed-Confidence(10),

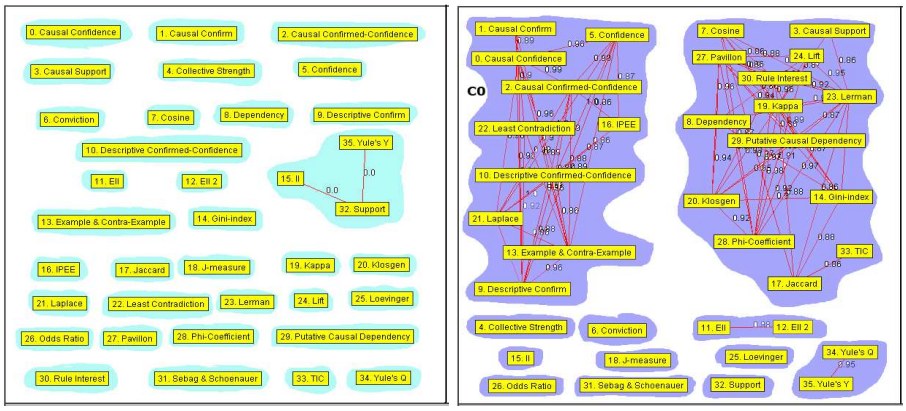


Fig. 6. CG0 and CG+ graphs on the mushroom dataset (clusters are highlighted with a gray background).

Example & Contra-Example(13), and IPEE (16) (on points 1, 2, 3, 4 respectively) that are good for a majority of interesting rules. This can also be retrieved from columns 5, 10, 13, 16 of Fig. 7. Among these four IMs, IPEE is the most suitable choice because of the lowest rule ranks obtained.

Measure Order		0	1	2	5	9	10	13	21	22	16	Rule's presentation
30	R107560	1	19121	1	1	41	1	1	8	5388	1	BROAD FREE ONE ==>veil_color=WHITE
31	R107562	1	18997	1	1	41	1	1	8	5361	1	BROAD ONE veil_color=WHITE ==>FREE
32	R107594	1	8972	1	1	18	1	1	3	2574	1	CLOSE FREE ONE ==>veil_color=WHITE
33	R107596	1	8914	1	1	18	1	1	3	2564	1	CLOSE ONE veil_color=WHITE ==>FREE
34	R122275	1	13800	1	1	32	1	1	5	3977	1	BROAD FREE ==>veil_color=WHITE
35	R122283	1	18299	1	1	38	1	1	6	5145	1	FREE stalk_surf_above=SMOOTH ==>veil_color=WHITE
36	R122285	1	18179	1	1	38	1	1	6	5134	1	stalk_surf_above=SMOOTH veil_color=WHITE ==>FREE
37	R122296	1	20903	1	1	55	1	1	10	6193	1	FREE stalk_surf_below=SMOOTH ==>veil_color=WHITE
38	R122308	65969	8772	40612	23743	10	23743	23743	23714	1013	1	FREE ==>ONE veil_color=WHITE

Fig. 7. Union of the ten interesting rules of the cluster $C0$ on the mushroom dataset (extract).

5 Focus on graph-based clustering approach

When considering a large set of IMs, the graph-based view of the correlation matrix may be quite complex. In order to highlight the more "natural" clusters, we propose to construct two types of subgraphs : the correlated ($CG+$) and the uncorrelated ($CG0$) partial subgraph. In this section we present the different filtering thresholds for their construction. We also extend the correlation graphs to graphs of stable clusters ($\overline{CG0}$ and $\overline{CG+}$) in order to compare several rulesets.

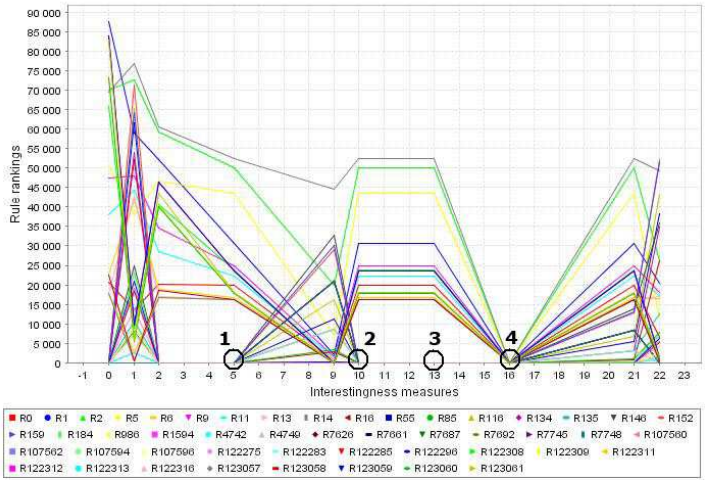


Fig. 8. Plot of the union of the ten interesting rules of the cluster $C0$ on the mushroom dataset.

5.1 Principles

Let $R(D) = \{r_1, r_2, ..., r_p\}$ denote a set of p association rules derived from a dataset D. Each rule $a \rightarrow b$ is described by its itemsets (a, b) and its cardinalities $(n, n_a, n_b, n_{a\bar{b}})$. Let M be the set of q available IMs for our analysis $M = \{m_1, m_2, ..., m_q\}$. Each IM is a numerical function on rule cardinalities: $m(a \rightarrow b) = f(n, n_a, n_b, n_{a\bar{b}})$. For each IM $m_i \in M$, we can construct a vector $m_i(R) = \{m_{i1}, m_{i2}, ..., m_{ip}\}, i = 1..q$, where m_{ij} corresponds to the calculated value of the IM m_i for a given rule r_j.

The correlation value between any two IMs $m_i, m_j\{i, j = 1..q\}$ on the set of rules R is calculated by using a Pearson's correlation coefficient $\rho(m_i, m_j)$ [27], where $\overline{m_i}, \overline{m_j}$ are the average values calculated of vector $m_i(R)$ and $m_j(R)$ respectively:

$$\rho(m_i, m_j) = \frac{\sum_{k=1}^{p}[(m_{ik} - \overline{m_i})(m_{jk} - \overline{m_j})]}{\sqrt{[\sum_{k=1}^{p}(m_{ik} - \overline{m_i})^2][\sum_{k=1}^{p}(m_{jk} - \overline{m_j})^2]}}$$

In order to make the interpretation of the large set of correlation values easier, we introduce the following definitions:

Definition 3. Two IMs m_i and m_j are *τ-correlated* with respect to the dataset D if their absolute correlation value is greater than or equal to a given threshold τ: $|\rho(m_i, m_j)| \geq \tau$. And, conversely, two IMs m_i and m_j are *θ-uncorrelated* with respect to the dataset D if the absolute value of their correlation value is lower than or equal to a threshold value θ: $|\rho(m_i, m_j)| \leq \theta$.

For θ-uncorrelated IMs, we use a statistical test of significance by choosing a level of significance of the test $\alpha = 0.05$ for hypothesis testing (common values for α are: $\alpha = 0.1, 0.05, 0.005$). The threshold θ is then calculated by the following formula: $\theta = 1.960/\sqrt{p}$ in a population of size p [27]. The assignment $\tau = 0.85$ of τ-correlated is used because this value is widely acceptable in the literature.

As the correlation coefficient is symmetrical, the $q(q-1)/2$ correlation values can be stored in one half of the table $q \times q$. This table $(I \times I)$ can also be viewed as the relation of an undirected and valued graph called correlation graph, in which a vertex value is an IM and an edge value is the correlation value between two vertices/IMs.

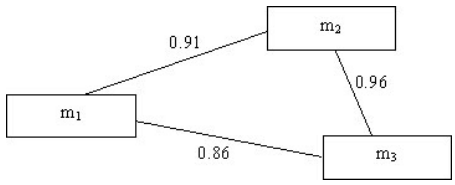


Fig. 9. An illustration of the correlation graph.

For instance, Fig. 9 can be the correlation graph obtained on five association rules $R(D) = \{r_1, r_2, r_3, r_4, r_5\}$ extracted from a dataset D and three IMs $M = \{m_1, m_2, m_3\}$ whose values and correlations are given in Tab. 3.

$R \times I$	m_1	m_2	m_3
r_1	0.84	0.89	0.91
r_2	0.86	0.90	0.93
r_3	0.88	0.94	0.97
r_4	0.94	0.95	0.99
r_5	0.83	0.87	0.84

$I \times I$	m_1	m_2	m_3
m_1		0.91	0.86
m_2			0.96
m_3			

Table 3. Correlation values for three IMs and five association rules.

5.2 Correlated versus uncorrelated graphs

Unfortunately, when the correlation graph is complete, it is not directly human-readable. We need to define two transformations in order to extract more limited and readable subgraphs. By using definition 3, we can extract the *correlated partial subgraph (CG+)*: the subgraph composed of edges associated with a τ-correlated. On the same way, the *uncorrelated partial subgraph (CG0)* where we only retain edges associated with correlation values close to 0 (θ-uncorrelated).

These two partial subgraphs can then be used as a visualization support in order to observe the correlative liaisons between IMs.

We can also observe the clusters of IMs corresponding with the connected parts of the graphs.

5.3 Extension to graph of stable clusters

In order to facilitate the comparison between several correlation matrices, we have introduced some extensions to define the stable clusters between IMs.

Definition 4. The $\overline{CG+}$ graph (respectively $\overline{CG0}$ graph) of a set of k rulesets $R = \{R(D_1), ..., R(D_k)\}$ is defined as the average graph of intersection of the k partially correlated (respectively uncorrelated) subgraphs $CG+_k$ (respectively $CG0_k$) calculated on R. Hence, each edge of $\overline{CG+}$ (respectively $\overline{CG0}$) is associated with the average value of the corresponding edge in the k $CG+_k$ graphs. Therefore, the $\overline{CG+}$ (respectively $\overline{CG0}$) graph allows visualizing the strongly (respectively weakly) stable correlations, as being common to k studied rulesets.

Definition 5. We call τ-stable (respectively θ-stable) clusters the connected part of the $\overline{CG+}$ (respectively $\overline{CG0}$) graph.

6 Study of IM behavior on two prototypical and opposite datasets

We have applied our method to two "opposite" datasets: D_1 and D_2, in order to compare correlation behavior and more precisely, to discover some stable clusters.

6.1 Data description

Our experiments are based on the categorical mushroom dataset (D_1) from Irvine machine-learning database repository and a synthetic dataset (D_2). The latter is obtained by simulating the transactions of customers in retail businesses, the dataset was generated using the IBM synthetic data generator [3]. D_2 has the typical characteristic of the Agrawal dataset T5.I2.D10k. We also generate the set of association rules (ruleset) R_1 (respectively R_2) from the dataset D_1 (respectively D_2) using the Apriori algorithm [2] [3]. For a closer evaluation of the IM behavior of the most interesting rules from these two rulesets, we have extracted $R_1^{'}$ (respectively $R_2^{'}$) from R_1 (respectively R_2) as the union of the first 1000 rules ($\approx 1\%$, ordered by decreasing IM values) issued from each IM (see Tab. 4).

In our experiment, we compared and analyzed the thirty-six IMs defined in Appendix B. We must notice that EII($\alpha = 1$) and EII($\alpha = 2$) are two entropic versions of the II measure.

Dataset	Items (Average length)	Transactions	Number of rules (support threshold)	$R(D)$	θ	τ	$R(D)$
D_1	118 (22)	8416	123228 (12%)	R_1	0.005	0.85	R_1
			10431 (12%)	$R_1^{'}$	0.020	0.85	$R_1^{'}$
D_2	81 (5)	9650	102808 (0.093%)	R_2	0.003	0.85	R_2
			7452 (0.093%)	$R_2^{'}$	0.012	0.85	$R_2^{'}$

Table 4. Description of the datasets.

6.2 Discussion

The analysis aims at finding stable relations between the IMs studied over the four rulesets. We investigate in: (1) the $\overline{CG0}$ graphs in order to identify the IMs that do not agree for ranking the rules, (2) the $\overline{CG+}$ graph in order to find the IMs that do agree for ranking the rules.

Ruleset	Number of correlations		Number of clusters	
	τ-correlated	θ-uncorrelated	CG+	CG0
R_1	79	2	12	34
$R_1^{'}$	91	15	12	21
R_2	65	0	14	36
$R_2^{'}$	67	17	12	20

Table 5. Comparison of correlation.

CG+ and *CG0*

Fig. 10 shows four $CG+$ graphs obtained from the four rulesets. As seen before, the sample rulesets and the original rulesets have close results so we can use the sample rulesets for representing the original rulesets. This observation is useful when we evaluate the CG+ graphs but not for CG0 graphs. For example, with the CG+ graph of R_1 (Fig. 10), one can choose the largest cluster containing the fourteen IMs (Causal Support, Pavillon, Lift, Lerman, Putative Causal Dependency, Rule Interest, Phi-Coefficient, Klosgen, Dependency, Kappa, Gini-index, Cosine, Jaccard, TIC) for his/her first choice. In this cluster one can also see the weak relation between TIC and the other IMs of the cluster. Tab. 5 also shows the two opposite tendencies obtained from the number of τ-correlated computed: $79(R_1) \rightarrow 91(R_1^{'}), 65(R_2) \rightarrow 67(R_2^{'})$.

With the four CG0 graphs (Fig. 11), one can easily see that the number of θ-uncorrelated increases when the most interesting rules are selected: $2(R_1) \rightarrow 15(R_1^{'}), 0(R_2) \rightarrow 17(R_2^{'})$ (Fig. 11, Tab. 5).

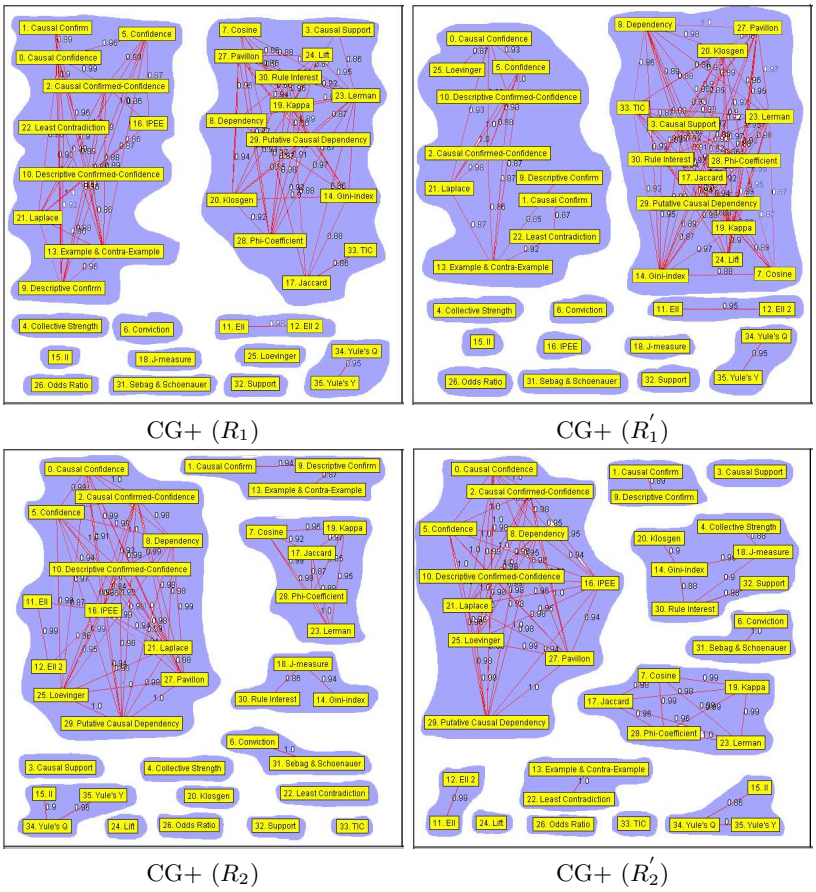


Fig. 10. CG+ graphs (clusters are highlighted in gray).

$\overline{CG0}$ graphs: uncorrelated stability

Uncorrelated graphs first show that there are no θ-stable clusters that appear on the four rulesets studied in Fig. 11. Secondly, there is no $\overline{CG0}$ graph from these datasets. A close observation of four CG0 graphs shows that at least one IM in each cluster will later be clustered around in a τ-stable cluster of $\overline{CG+}$ graph (Fig. 11, Fig. 12) like Yule's Y, Putative Causal Dependency, EII($\alpha = 2$), Cosine, Laplace so that the stronger the θ-uncorrelated, the more interesting the IM that participated in the θ-uncorrelated.

$\overline{CG+}$ graph: correlated stability

From Tab. 5, we can see that, R_1' is approximately twice as correlated as R_2'. As seen in Fig. 12, five τ-stable clusters found come from the datasets studied.

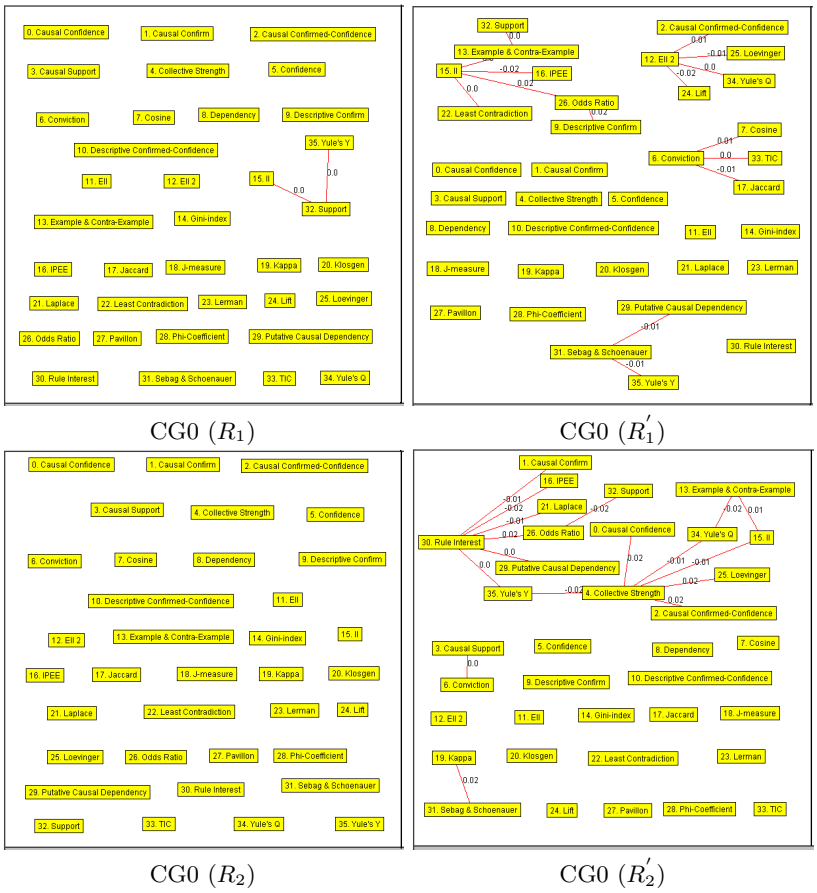


Fig. 11. CG0 graphs.

By briefly analyzing these τ-stable clusters, some interesting observations are drawn.

(C1), the largest cluster, (Confidence, Causal Confidence, Causal Confirmed-Confidence, Descriptive Confirmed-Confidence, Laplace) has most of its IMs extended from Confidence measure. From this cluster, we can easily see a highly connected component – each vertex must have an edge with the other vertices – indicating the strong agreement of the five IMs.

According to the taxonomy (Tab. 1), this cluster is associated with descriptive IMs that are sensible to equilibrium.

(C2), another cluster, has two highly connected components which are formed by Phi-Coefficient, Lerman, Kappa, Cosine and Jaccard. Most of these IMs are similarity measures. According to the taxonomy (Tab. 1) this cluster is to measure the deviation from independence.

(C3), this cluster (Dependency, Pavillon, Putative Causal Dependency) is interesting because almost all the IMs of this cluster are reasonably well correlated. The nature of these IMs are descriptive.

(C4), is a cluster formed by EII and EII 2, which are two IMs obtained with different parameters of the same original formula. This cluster has many extended directions to evaluate the entropy of II.

(C5), Yule's Q and Yule's Y, brings out a trivial observation because these IMs are derived from Odds Ratio measure. Both IMs are descriptive and measuring of deviation from independence.

In looking for τ-stable clusters, we have found the τ-correlated that exist between various IMs and we have identified five τ-stable clusters. Each τ-stable cluster forms a subgraph in a $\overline{CG+}$ graph, also contains a highly connected component. Therefore, we can choose a representative IM for each cluster. For example, in our experiment, we have five representative IMs for all the thirty-six IMs. How we can choose a representative IM is also an interesting study for the future. In the first approach, we can select the IM that has the highest number of relations with the others: Causal Confidence, Cosine, Klosgen, EII($\alpha = 2$), and Yule's Y. The stronger the τ-stable cluster, the more interesting the representative IM. An important observation is that, the existence of highly connected graphs represents a strong agreement with a τ-stable cluster. We have reached significant information: *τ-stable clusters can be obtained from different IMs and rulesets*. The different IMs imply taking into account both their mathematical definitions and their respective significance. The datasets are both highly correlated and lowly correlated.

7 Conclusion

We have studied and compared the various IMs described in the literature in order to help the decision-maker to better understand the behavior of the IMs in the stage of post-processing of association rules. A new approach called correlation graph implemented by a new tool, ARQAT, with two types: CG+ and CG0 is proposed to evaluate IMs by using graphs as a visual insight on the data.

With this approach, the decision-maker has a few IMs to decide and as a graphical representation to select the most interesting rules to examine. Another interesting result obtained from this work is that we have found some stable clusters between IMs, five such τ-stable clusters have been found with the $\overline{CG+}$ graph. Our approach is highly related to the real value of the dataset and the number of proposed IMs.

Our future research will investigate the two following directions: first, we will improve the correlation analysis by introducing a better measure than linear correlation whose limits are stressed in the literature; second, we will also improve the IM clustering analysis with IM aggregation techniques to facilitate the user's decision making from the most suitable IMs.

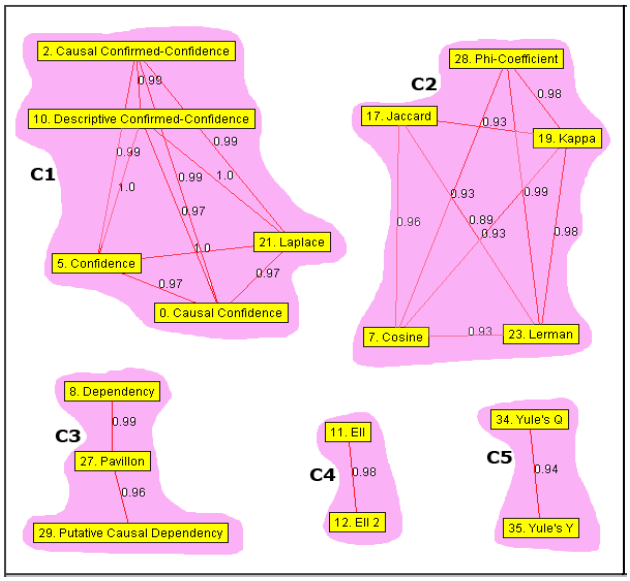


Fig. 12. $\overline{CG+}$ graph.

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. Proceedings of the ACM-SIGMOD International Conference on Management of Data. Washington DC, USA (1993) 207–216
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. VLDB'94, Proceedings of the 20th International Conference on Very Large Data Bases. Santiago, Chile (1994) 487–499
3. Agrawal, R., Mannila, H., Srikant, R., Toivonen H., Verkano, A.I.: Fast discovery of association rules. Advances in Knowledge Discovery in Databases. (1996) 307–328
4. Azé, J., Kodratoff, Y.: A study of the Effect of Noisy Data in Rule Extraction Systems. EMCSR'02, Proceedings of the Sixteenth European Meeting on Cybernetics and Systems Research. (2002) 781–786
5. Bayardo, Jr.R.J., Agrawal, R.: Mining the most interesting rules. KDD'99, Proceedings of the Fifth ACM SIGKDD international conference on Knowledge discovery and data mining. San Diego, CA, USA (1999) 145–154
6. Blanchard, J., Guillet, F., Gras, R., and Briand, H.: Using information-theoretic measures to assess association rule interestingness. ICDM'05, Proceedings of the 5th IEEE International Conference on Data Mining, IEEE Computer Society Press, (2005) 66–73.
7. Blanchard, J., Guillet, F., Gras, R., Briand, H.: Assessing rule interestingness with a probabilistic measure of deviation from equilibrium. ASMDA'05, Proceedings of the 11th International Symposium on Applied Stochastic Models and Data Analysis. (2005) 191–200

8. Blanchard, J., Guillet, F., Gras, R., Briand, H.: Mesurer la qualité des règles et de leurs contraposées avec le taux informationnel TIC. EGC'04, Actes de 4èmes journées d'Extraction et de Gestion des Connaissances, RNTI-E-2, Vol. 1. Cépaduès Editions, Clermont Ferrand, France (2004) 287–298 (in French)
9. Blanchard, J., Kuntz, P., Guillet, F., Gras, R.: Implication Intensity: from the basic statistical definition to the entropic version. Statistical Data Mining and Knowledge Discovery, Chapter 28. Chapman & Hall, CRC Press (2003) 475–493
10. Freitas, A.A.: On rule interestingness measures. Knowledge-Based Systems, 12(5-6). (1999) 309–315
11. Gavrilov, M., Anguelov, D., Indyk, P., and Motwani, R.: Mining the stock market: which measure is best?. KDD'00, Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining . Boston, MA, USA (2000) 487–496.
12. Gras, R., Couturier, R., Blanchard, J., Briand, H., Kuntz, P., Peter, P.: Quelques critères pour une mesure de qualité de règles d'association. Mesures de Qualité pour la Fouille de Données, RNTI-E-1. Cépaduès Editions (2004) 3–31 (in French)
13. Gras, R.: L'implication statistique - Nouvelle méthode exploratoire de données. La Pensée Sauvage Édition (1996) (in French)
14. Hilderman, R.J., Hamilton, H.J.: Knowledge Discovery and Measures of Interestingness. Kluwer Academic Publishers (2001)
15. Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., and Verkano, A.I.: Finding interesting rules from larges sets of discovered association rules. ICIKM'94, Proceedings of the Third International Conference on Information and Knowledge Management. Ed. Nabil R. Adam, Bharat K. Bhargava and Yelena Yesha, Gaithersburg, Maryland. ACM Press, (1994) 401–407.
16. Huynh, X.-H., Guillet, F., Briand, H.: Clustering interestingness measures with positive correlation. ICEIS'05, Proceedings of the 7th International Conference on Enterprise Information Systems. (2005) 248–253
17. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York, (1990)
18. Kodratoff, Y.: Comparing Machine Learning and Knowledge Discovery in DataBases: An Application to Knowledge Discovery in Texts. Machine Learning and Its Applications, LNCS 2049. Springer-Verlag, (2001) 1–21
19. Kononenco, I.: On biases in estimating multi-valued attributes. IJCAI'95. (1995) 1034–1040
20. Lenca, P., Lallich, S., Vaillant, B.: On the robustness of association rules. Proceedings of the IEEE International Conference on Cybernetics and Intelligent Systems. (2006) 596–601
21. Liu, B., Hsu, W., Mun, L., Lee, H.: Finding interestingness patterns using user expectations. IEEE Transactions on Knowledge and Data Mining (11). (1999) 817–832
22. Loevinger, J.: A systematic approach to the construction and evaluation of tests of ability. Psychological Monographs. (1947)
23. Newman, D.J., Hettich,S., Blake, C.L., Merz, C.J.: [UCI] Repository of machine learning databases, http://www.ics.uci.edu/∼mlearn/MLRepository.html. University of California, Irvine, Department of Information and Computer Sciences, (1998).
24. Padmanabhan, B., Tuzhilin, A. : A belief-driven method for discovering unexpected patterns. KDD'98, Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining. (1998) 94–100

25. Piatetsky-Shapiro, G.: Discovery, analysis and presentation of strong rules. Knowledge Discovery in Databases, G. Piatetsky-Shapiro and W. Frawley editors. MIT Press, Cambridge, MA (1991) 229–248
26. Piatetsky-Shapiro, G., Steingold, S.: Measuring Lift Quality in Database Marketing. SIGKDD Explorations 2(2). (2000) 76–80
27. Ross, S.M.: Introduction to probability and statistics for engineers and scientists. Wiley, (1987)
28. Sebag, M., Schoenauer, M.: Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases. EKAW'88, Proceedings of the European Knowledge Acquisition Workshop. Gesellschaft fr Mathematik und Datenverarbeitung mbH (1988) 28.1–28.20
29. Silberschatz, A., Tuzhilin, A.: What makes patterns interesting in knowledge discovery systems. IEEE Transactions on Knowledge Data Engineering 8(6). (1996) 970–974
30. Tan, P.N., Kumar, V., Srivastava, J.: Selecting the right objective measure for association analysis. Information Systems 29(4). (2004) 293–313
31. Vaillant, B., Lenca, P., Lallich, S.: A clustering of interestingness measures. DS'04, the 7th International Conference on Discovery Science LNAI 3245. (2004) 290–297
32. Vaillant, B., Lallich, S., Lenca, P.: Modeling of the counter-examples and association rules interestingness measures behavior. The 2006 International Conference on Data Mining. (2006)
33. Zhao, Y., Karypis, G.: Criterion functions for document clustering: experiments and analysis. Technical Report TR01-40, Deparment of Computer Science, University of Minnesota. (2001) 1–30

A Complementary IMs: II, EII, TIC, IPEE

A.1 Implication Intensity

Initially introduced by Gras [13], the implicative intensity II aims at quantifying the "surprisingness" of a rule.

Intuitively, it is more surprising to discover that a rule has a small number of negative examples when the dataset is large. Hence, the objective of the implicative intensity is to express the unlikelihood of $n_{a\bar{b}}$ in T.

More precisely, we compare the observed number of negative examples $n_{a\bar{b}}$ with the number $N_{a\bar{b}}$ of expected negative examples for an independence hypothesis. Let us assume that we randomly draw two subsets U and V in T with respectively n_a and n_b transactions. Then, $N_{a\bar{b}} = \left|U \bigcap \overline{V}\right|$ is the random variable associated with the number of negative examples in this random model.

Definition 6. The implicative intensity II of the rule $a \rightarrow b$ is defined by

$$II\left(a \rightarrow b\right) = 1 - p\left(N_{a\bar{b}} \leq n_{a\bar{b}}\right)$$

if $n_a \neq n$; otherwise

$$II\left(a \rightarrow b\right) = 0$$

.

In practice, the distribution of $N_{a\overline{b}}$ depends on the random drawing pattern. We here consider a hyper-geometric law: $p\left(N_{a\overline{b}} = k\right) = \frac{C_{n_{\overline{b}}}^{n_{\overline{b}}-k} C_{n_a}^{k}}{C_n^{n_b}}$. The effective value of II can be easily computed with this recursive formula. Other models based on the binomial law and the Poisson distribution have been proposed.

A.2 Entropic Implication Intensity

Definition 6 essentially measures the surprisingness of the rule $a \rightarrow b$. However, taking the contrapositive $\overline{b} \rightarrow \overline{a}$ into account could reinforce the assertion of the implication between a and b. Moreover, it could improve the quality of discrimination of II when the transaction set T increases: if A and B are small compared to T, their complementary sets are large and vice-versa.

For these reasons, we have introduced a weighted version of the implication intensity $\left(E\left(a,b\right).II\left(a \rightarrow b\right)\right)^{1/2}$ where $E\left(a,b\right)$ measures the disequilibrium between n_{ab} and $n_{a\overline{b}}$ – associated with $a \rightarrow b$ –, and the disequilibrium between $n_{a\overline{b}}$ and $n_{\overline{a}\overline{b}}$ – associated with its contrapositive – [9]. Intuitively, the surprise must be softened (respectively confirmed) when the number of negative examples $n_{a\overline{b}}$ is high (respectively small) for the rule and its contrapositive considering the observed cardinalities n_a and $n_{\overline{b}}$.

A well-known index for taking the cardinalities into account non-linearly is the Shannon conditional entropy. The conditional entropy $H_{b/a}$ of cases (a and b) and (a and $\overline{b}$) given a is defined by

$$H_{b/a} = -\frac{n_{ab}}{n_a} log_2 \frac{n_{ab}}{n_a} - \frac{n_{ab}}{n_a} log_2 \frac{n_{a\overline{b}}}{n_a}$$

and, similarly, we obtain the conditional entropy $H_{\overline{a}/\overline{b}}$ of cases ($\overline{a}$ and $\overline{b}$) and (a and $\overline{b}$) given $\overline{b}$. The complements of 1 for these uncertainties $1 - H$ can be interpreted as the average information collected by the realization of these experiments; the higher this information, the stronger the quality of the implication and its contrapositive.

The expected behavior of the weighted version of II is determined in three stages: *(i)* a slow reaction to the first negative examples (robustness to noise), *(ii)* an acceleration of the rejection in the neighborhood of the equilibrium, *(iii)* an increasing rejection beyond the equilibrium. The adjustment of $1 - H$ proposed in definition 6 satisfies these requirements.

Definition 7. Let $\alpha > 1$ be a fixed number. The disequilibriums are measured by $E\left(a,b\right)$, is defined by

$$E\left(a,b\right)=\left(\left(1-H_{b/a}\right)^{\alpha}.\left(1-H_{\overline{a}/\overline{b}}\right)^{\alpha}\right)^{1/2\alpha}$$

if $\frac{n_{a\overline{b}}}{n}\in\left[0,\frac{n_a}{2n}\right[\bigcap\left[0,\frac{n_b}{2n}\right[$;

$$E(a,b)=0$$

otherwise.

And, the weighted version of the implication intensity – called the entropic implication intensity – is given by

$$EII\left(a\rightarrow b\right)=\left(E\left(a,b\right).II\left(a\rightarrow b\right)\right)^{1/2}$$

Raising the conditional entropies to the power α reinforces the contrast between the different stages presented above.

A.3 TIC

In [6], we introduced DIR (Directed Information Ratio), a new rule IM which is based on information theory. DIR is the entropy decrease rate of the consequent due to the truth of the antecedent, but it is not calculated with a classical entropy function. We use an asymmetric entropy function which considers that the uncertainty is maximal (entropy = 1) when the studied modality is not the more likely. This allows DIR to differentiate two opposite rules $a\rightarrow b$ and $a\rightarrow\overline{b}$, which is not possible with the other information-theoretic measures of rule interestingness. Moreover, to our knowledge, DIR is the only rule IM which rejects both independence and equilibrium, i.e. it discards both the rules whose antecedent and consequent are negatively correlated, and the rules which have more negative examples than examples.

In [8], we proposed another IM, derived from DIR, which assesses the rules by taking their contrapositives into account. This new IM called TIC (*Taux Informationnel modulé par la Contraposée, in French*) is the geometric mean of the values of DIR for a rule and its contrapositive (if one of the two values of DIR is negative, then TIC is worth zero). Considering both the rule and its contrapositive allows to discover rules that are closer to logical implication.

A.4 IPEE

As there was no statistical IMs evaluating the deviation from equilibrium, we proposed the new measure IPEE in [7]. Following II, IPEE is based on a probabilistic model. However, while II evaluates the statistical significance of the deviation from independence, IPEE evaluates the statistical significance of the deviation from equilibrium.

B Formulas of IMs

N	Interestingness measure	$f(n, n_a, n_b, n_{a\bar{b}})$	Reference
0	Causal Confidence	$1 - \frac{1}{2}(\frac{1}{n_a} + \frac{1}{n_{\bar{b}}})n_{a\bar{b}}$	[18]
1	Causal Confirm	$\frac{n_a + n_{\bar{b}} - 4n_{a\bar{b}}}{n}$	[18]
2	Causal Confirmed-Confidence	$1 - \frac{1}{2}(\frac{3}{n_a} + \frac{1}{n_{\bar{b}}})n_{a\bar{b}}$	[18]
3	Causal Support	$\frac{n_a + n_{\bar{b}} - 2n_{a\bar{b}}}{n}$	[18]
4	Collective Strength	$\frac{(n_a - n_{a\bar{b}})(n_{\bar{b}} - n_{a\bar{b}})(n_a n_{\bar{b}} + n_b n_{\bar{a}})}{(n_a n_b + n_{\bar{a}} n_{\bar{b}})(n_b - n_a + 2n_{a\bar{b}})}$	[30]
5	Confidence	$1 - \frac{n_{a\bar{b}}}{n_a}$	[2]
6	Conviction	$\frac{n_a n_{\bar{b}}}{n n_{a\bar{b}}}$	[30]
7	Cosine	$\frac{n_a - n_{a\bar{b}}}{\sqrt{n_a n_b}}$	[30]
8	Dependency	$\lvert \frac{n_{\bar{b}}}{n} - \frac{n_{a\bar{b}}}{n_a} \rvert$	[18]
9	Descriptive Confirm	$\frac{n_a - 2n_{a\bar{b}}}{n_a}$	[18]
10	Descriptive Confirmed-Confidence	$1 - 2\frac{n_{a\bar{b}}}{n_a}$	[18]
11	EII ($\alpha = 1$)	$\sqrt{\varphi \times I}^{\frac{1}{2\alpha}}$	[9]
12	EII ($\alpha = 2$)	$\sqrt{\varphi \times I}^{\frac{1}{2\alpha}}$	[9]
13	Example & Contra-Example	$1 - \frac{n_{a\bar{b}}}{n_a - n_{a\bar{b}}}$	[13]
14	Gini-index	$\frac{(n_a - n_{a\bar{b}})^2 + n_{a\bar{b}}^2}{n n_a} + \frac{(n_b - n_a + n_{a\bar{b}})^2 + (n_{\bar{b}} - n_{a\bar{b}})^2}{n n_{\bar{a}}} - \frac{n_{\bar{b}}^2}{n^2} - \frac{n_{\bar{b}}^2}{n^2}$	[30]
15	II	$1 - \sum_{k=max(0, n_a - n_b)}^{n_{a\bar{b}}} \frac{C_{n_b}^{n_a - k} C_{n_{\bar{b}}}^{k}}{C_n^{n_a}}$	[13]
16	IPEE	$1 - \frac{1}{2^{n_a}} \sum_{k=0}^{n_{a\bar{b}}} C_{n_a}^{k}$	[7]
17	Jaccard	$\frac{n_a - n_{a\bar{b}}}{n_b + n_{a\bar{b}}}$	[30]
18	J-measure	$\frac{n_a - n_{a\bar{b}}}{n} log_2 \frac{n(n_a - n_{a\bar{b}})}{n_a n_b} + \frac{n_{a\bar{b}}}{n} log_2 \frac{n n_{a\bar{b}}}{n_a n_{\bar{b}}}$	[30]
19	Kappa	$\frac{2(n_a n_{\bar{b}} - n n_{a\bar{b}})}{n_a n_{\bar{b}} + n_{\bar{a}} n_b}$	[30]
20	Klosgen	$\sqrt{\frac{n_a - n_{a\bar{b}}}{n}}(\frac{n_{\bar{b}}}{n} - \frac{n_{a\bar{b}}}{n_a})$	[30]
21	Laplace	$\frac{n_a + 1 - n_{a\bar{b}}}{n_a + 2}$	[30]
22	Least Contradiction	$\frac{n_a - 2n_{a\bar{b}}}{n_b}$	[4]
23	Lift	$\frac{n(n_a - n_{a\bar{b}})}{n_a n_b}$	[26]
24	Lerman	$\frac{n_a - n_{a\bar{b}} - \frac{n_a n_b}{n}}{\sqrt{\frac{n_a n_b}{n}}}$	[13]
25	Loevinger	$1 - \frac{n n_{a\bar{b}}}{n_a n_{\bar{b}}}$	[22]
26	Odds Ratio	$\frac{(n_a - n_{a\bar{b}})(n_{\bar{b}} - n_{a\bar{b}})}{n_{a\bar{b}}(n_b - n_a + n_{a\bar{b}})}$	[30]
27	Pavillon/Added Value	$\frac{n_{\bar{b}}}{n} - \frac{n_{a\bar{b}}}{n_a}$	[30]
28	Phi-Coefficient	$\frac{n_a n_{\bar{b}} - n n_{a\bar{b}}}{\sqrt{n_a n_b n_{\bar{a}} n_{\bar{b}}}}$	[30]
29	Putative Causal Dependency	$\frac{3}{2} + \frac{4n_a - 3n_b}{2n} - (\frac{3}{2n_a} + \frac{2}{n_{\bar{b}}})n_{a\bar{b}}$	[18]
30	Rule Interest	$\frac{n_a n_{\bar{b}}}{n} - n_{a\bar{b}}$	[25]
31	Sebag & Schoenauer	$\frac{n_a}{n_{a\bar{b}}} - 1$	[28]
32	Support	$\frac{n_a - n_{a\bar{b}}}{n}$	[1]
33	TIC	$\sqrt{DIR(a \rightarrow b) \times DIR(\bar{b} \rightarrow a)}$	[8] [6]
34	Yule's Q	$\frac{n_a n_{\bar{b}} - n n_{a\bar{b}}}{n_a n_{\bar{b}} + (n_b - n_{\bar{b}} - 2n_a) n_{a\bar{b}} + 2n_{a\bar{b}}^2}$	[30]
35	Yule's Y	$\frac{\sqrt{(n_a - n_{a\bar{b}})(n_{\bar{b}} - n_{a\bar{b}})} - \sqrt{n_{a\bar{b}}(n_b - n_a + n_{a\bar{b}})}}{\sqrt{(n_a - n_{a\bar{b}})(n_{\bar{b}} - n_{a\bar{b}})} + \sqrt{n_{a\bar{b}}(n_b - n_a + n_{a\bar{b}})}}$	[30]