ARQAT: An Exploratory Analysis Tool For Interestingness Measures

Xuan-Hiep Huynh, Fabrice Guillet, and Henri Briand

LINA CNRS FRE 2729 - Polytechnic school of Nantes University
La Chantrerie BP 50609
44306 Nantes Cedex 3, France
(e-mail: `xuan-hiep.huynh@polytech.univ-nantes.fr`,
`fabrice.guillet@polytech.univ-nantes.fr`,
`henri.briand@polytech.univ-nantes.fr`)

Abstract. 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 help choose the best 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 to select the best interestingness measures. In this paper, we present ARQAT a new tool to study the specific behavior of a set of 34 interestingness measures in the context of a specific dataset and in an exploratory data analysis perspective. The tool implements 14 graphical and complementary views structured on 5 levels of analysis: ruleset analysis, correlation and clustering analysis, best rules analysis, sensitivity analysis, and comparative analysis. The tool is described and illustrated on the mushroom dataset in order to show the interest of both the exploratory approach and the use of complementary views.
Keywords: interestingness measure, ARQAT, exploratory analysis.

1 Introduction

In the last decade, the designing of Interestingness Measure (IM) to evaluate association rules has become an important knowledge quality challenge in the context of KDD. This is because association rule [Agrawal *et al.*, 1993] is one of the few models dedicated to unsupervised discovery of rule tendencies in data. It is unfortunately confronted to a major difficulty: the user (a decision-maker or a data-analyst) must cope with a large amount of extracted rules in order to validate and select the best ones [Piatetsky-Shapiro, 1991]. One way to reduce the cost of the user's task is to help him/her with the measurement of rule interestingness adapted to both his/her goals and the dataset studied.

In initial research works [Agrawal *et al.*, 1993][Agrawal and Srikant, 1994] on association rules, these precursors have introduced the first two statistical measures: support and confidence. These measures are well adapted to

Apriori algorithm constraints, but are not sufficient to capture rule interestingness. To improve this limit, many complementary IMs have been then introduced in the research literature. As interestingness depends both on the user's goals and data characteristics, two kinds of IMs may be distinguished [Freitas, 1999]: subjective and objective. First, subjective measures depend on the user's goals and his/her knowledge or beliefs, and are combined to specific supervised algorithms in order to compare the extracted rules with what the user knows or wants [Padmanabhan and Tuzhilin, 1998][Liu *et al.*, 1999]. Hence, subjective measures allow capturing rule novelty and unexpectedness in relation to the user's knowledge or beliefs. Second, objective measures are statistical indexes that only rely on data structure and more precisely on itemset frequency. Many interesting surveys summarize their definitions and properties (see [Bayardo and Agrawal, 1999], [Hilderman and Hamilton, 2001],
[Tan *et al.*, 2002], [Tan *et al.*, 2004], [Piatetsky-Shapiro, 1991],
[Lenca *et al.*, 2004], [Guillet, 2004]). These surveys address two joint research issues, the definition of the set of principles or properties that lead to the design of a good IM, and their comparison from a data-analysis point of view to study IM behavior in order to help the user to select the best ones. In [Vaillant *et al.*, 2003] a tool HERBS is also presented.

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 IMs in the context of a specific dataset and in an exploratory analysis perspective. More precisely, ARQAT is a toolbox designed to help a data-analyst to capture the best measures and as a final purpose, the best rules within a specific ruleset.

The paper is structured as follows. In section 2, we introduce the principles and the structure of ARQAT tool. In the three next sections, we describe 3 groups of ARQAT views: ruleset statistics, correlation analysis, and best rules analysis. We illustrate each view on the mushroom dataset, in order to show the interest of the exploratory approach for IM analysis.

2 Principles of ARQAT tool

ARQAT is an exploratory analysis tool that embeds 34 objective IMs studied in surveys. We complete this list of IMs with three complementary measures: Implication Intensity (II) introduced by Gras [Gras, 1996] [Guillaume *et al.*, 1998], Entropic Implication Intensity (EII) [Gras *et al.*, 2001] [Blanchard *et al.*, 2003], and the informational ratio modulated by the contra-positive (TIC)
[Blanchard *et al.*, 2004] (See Appendix 1 for a complete list of selected measures).

ARQAT (Fig. 1) implements a set of 14 complementary and graphical views structured in 5 task-oriented groups: ruleset analysis, correlation and clustering analysis, best rules analysis, sensitivity analysis, and comparative analysis.

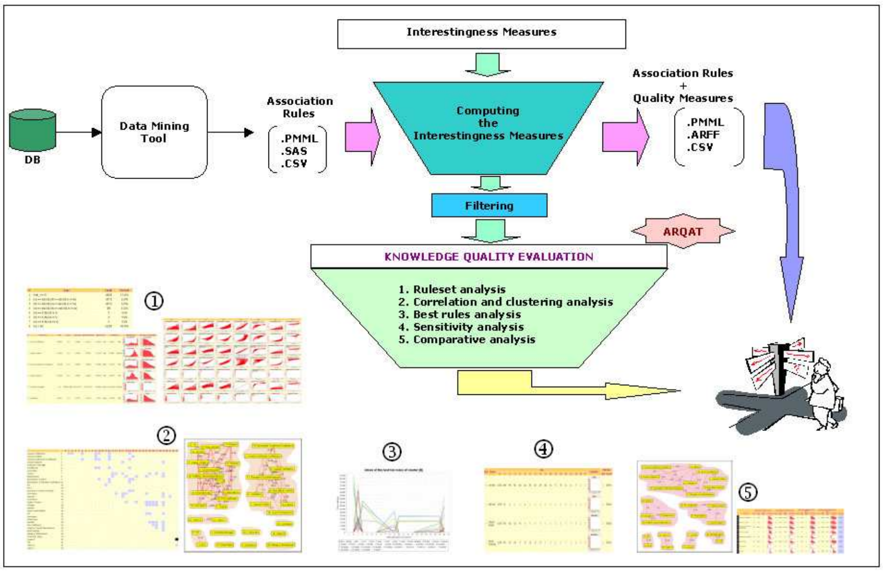


Fig. 1. ARQAT structure.

For the input, ARQAT requires an association ruleset where each association rule $a \Rightarrow b$ must be associated to 4 cardinalities $(n, n_a, n_b, n_{a\overline{b}})$. More precisely, n is the number of transactions, n_a (resp. n_b) the number of transactions satisfying the itemset a (resp. b), and $n_{a\overline{b}}$ is the number of transactions satisfying $a \wedge \overline{b}$ (negative examples).

In a 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 in order to focus the study on a more restricted subset of rules.

In a second stage, the data-analyst can then drive the graphical exploration of results through a classical web-browser. ARQAT is structured in 5 groups of task-oriented views. The first group (1 in Fig. 1) 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 best IMs. The third one (3) focuses on rule ordering to select the best 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.

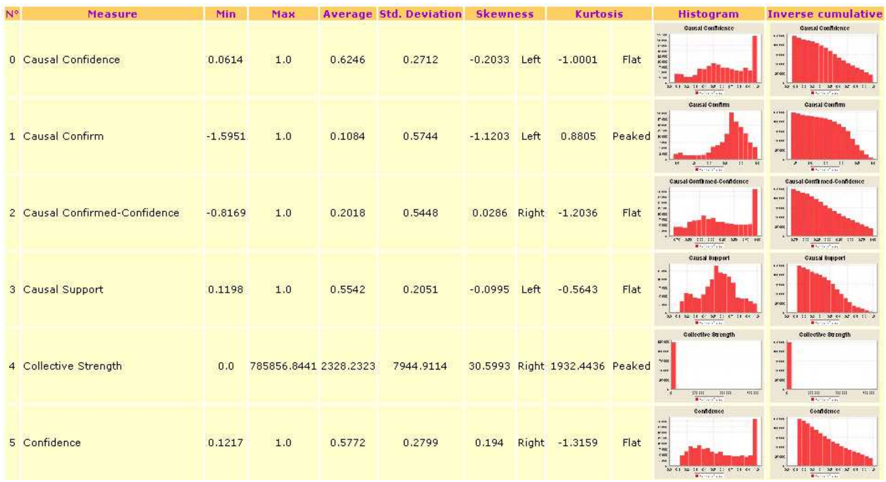
The next sections will focus on the description of the first three groups and will illustrate it with the same ruleset: 120000 association rules extracted by Apriori algorithm (support 10%) from mushroom dataset [Blake and Merz, 1998].

3 Ruleset statistics

This first group of ARQAT tools delivers 3 views summarizing some simple statistics in the ruleset structure. The first one, ruleset characteristics , shows the distributions underlying rule cardinalities, in order to detect borderline cases.

The second view, IM distribution (Fig. 2), draws the histograms for each IM. The distributions are also completed with minimum, maximum, average, standard deviation, skewness and kurtosis values. In Fig. 2, 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. 3), shows the scatter-plot matrix of all IM pairs. This graphical matrix is very useful to see the details of the relationships between IMs. For instance, Fig. 3 shows four disagreement shapes: Rule Interest vs Yule's Q (4), Sebag & Schoenauer vs Yule's Y (5), Similarity Index vs Support (6), and Yule's Y vs Support (7) (strongly uncorrelated). On the other hand, we can notice four agreement shapes on Putative Causal Dependency vs Rule Interest (1), Putative Causal Dependency vs Similarity Index (2), Rule Interest vs Similarity Index (3), and Yule's Q vs Yule's Y (8) (strongly 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		

Fig. 2. Distribution of some measures on mushroom dataset.

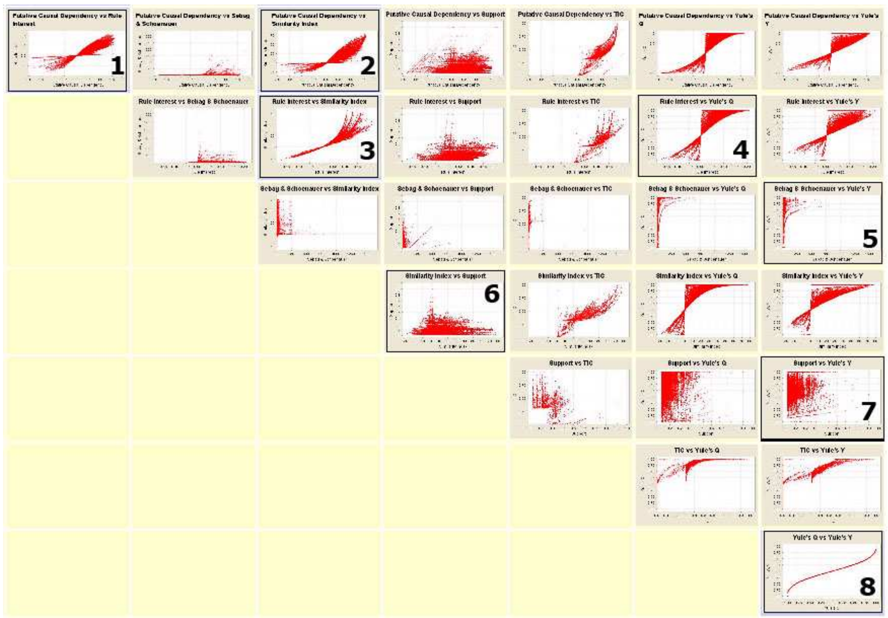


Fig. 3. Scatterplot matrix of joint-distributions on mushroom dataset.

4 Correlation analysis

This second group is dedicated to IM correlation study in order to deliver IM clustering and facilitate the choice of the subset of IMs that is the best-adapted to describe the ruleset. The correlations between IM pairs were computed in the preprocessing stage by using the Pearson's correlation coefficient and stored in the correlation matrix (I $\times$ I). The user has two visual possibilities to explore the matrix. The first one is a simple summary matrix in which each significant correlation value is visually associated to a different color (a level of gray). For instance, the only one dark cell from Fig. 4 shows a low correlation value between Yule's Y and Support. The other seventy-four gray cells correspond to high correlation values.

The second one (Fig. 5) is a graph-based view of the correlation matrix. As graphs are a good way to offer relevant graphical insights on data structure, we use the correlation matrix 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 2 vertices/measures. We also add the possibility to set a minimal threshold τ (resp. maximal threshold θ) to retain only the edges associated to a high correlation (resp. low correlation), that deliver a partial subgraph CG+ (resp. CG0).

These two partial subgraphs can then be processed in order to extract clusters of measures. Each cluster is defined as a maximal connected subgraph. In CG+, each cluster will gather correlated or anti-correlated mea-

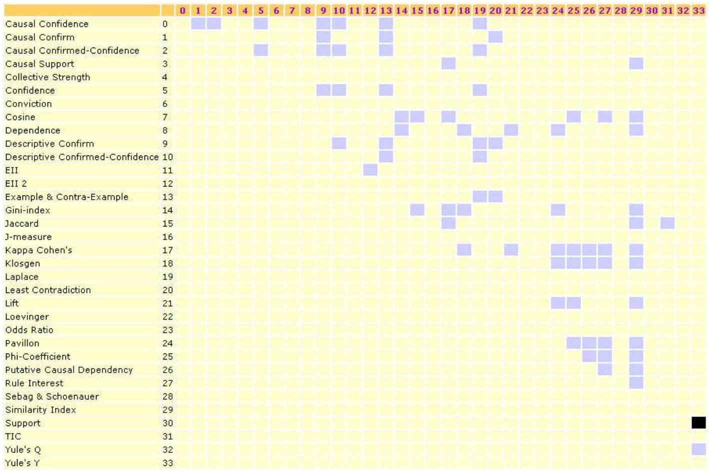


Fig. 4. Summary matrix of correlation on mushroom dataset.

sures that may be interpreted similarly: they deliver a close point of view on data. Moreover, in CG0 each cluster will contain uncorrelated measures: measures that deliver a different point of view.

Hence, as each graph depends on a specific ruleset, the user will use the graphs as data insight, which will graphically help him/her to select the minimal set of the measures best adapted to his/her data. For instance in Fig. 5, CG+ graph contains 11 clusters on 34 measures, the user can select in each cluster the most representative measure, and then retain it to validate the rules.

A close watch on the CG0 graph (Fig. 5) shows an uncorrelated cluster formed by Support and Yule's Y measures (also the dark cell in Fig. 4). This observation is confirmed on Fig. 3 (7). CG+ graph shows a trivial cluster where Yule's Q and Yule's Y are strongly correlated. This is also confirmed on Fig. 3 (8) showing a functional dependency between the two measures. These two examples show the interest to use the scatterplot matrix complementarily (Fig. 3) with the correlation graphs CG0, CG+ (Fig. 5) in order to evaluate the nature of the correlation links, and overcome the limits of the correlation coefficient.

5 Best rule analysis

In order to help a user to select the best rules, we have implemented two specific views. The first view (Fig. 6) collects a set of given number of best rules for each measure in one cluster, in order to answer the question "How interesting are the rules of this cluster?"

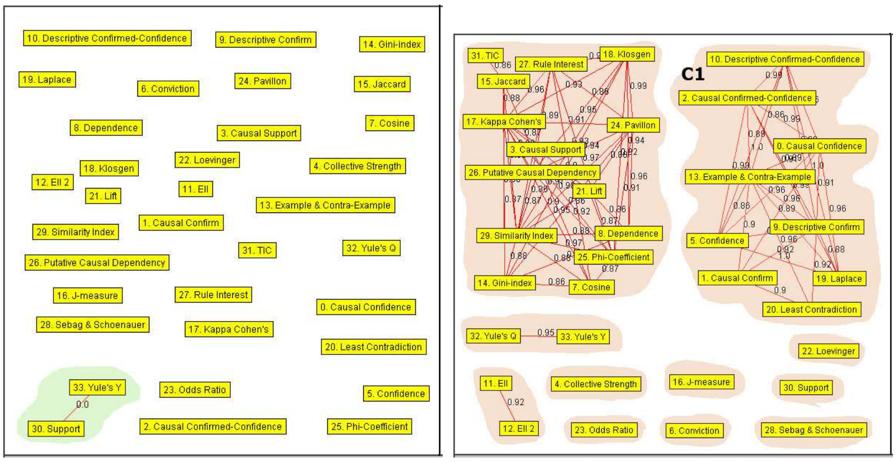


Fig. 5. CG0 and CG+ graphs on mushroom dataset (clusters are highlighted with a gray background).

The selected rules can alternatively be visualized with parallel coordinates drawing (Fig. 7). The main interest of such a drawing is to rapidly see the IM rankings of the rules, and then to facilitate their interpretation.

These two views can be used with IM values of a rule or alternatively with the rank of the value. For instance, Fig. 6 and Fig. 7 use the rank to evaluate the union of the ten best rules for each of the nine IMs in the C1 cluster (see Fig. 5). The Y-axis in Fig. 7 holds the rule rank for the corresponding measure. By observing the concentration lines on low rank values, we can obtain 3 measures: Confidence(5), Decsriptive Confirmed-Confidence(10), and Example & Contra-Example(13) (on points 1, 2, 3 respectively) that are good for a majority of best rules. This can also be retrieved from columns 5, 10, 13 of Fig. 6.

Measure Order		0	1	2	5	9	10	13	19	20	Rule's presentation
21	R107560	1	19121	1	1	41	1	1	8	5388	BROAD FREE ONE ==>veil_color=WHITE
22	R107562	1	18997	1	1	41	1	1	8	5361	BROAD ONE veil_color=WHITE ==>FREE
23	R107594	1	8972	1	1	18	1	1	3	2574	CLOSE FREE ONE ==>veil_color=WHITE
24	R107596	1	8914	1	1	18	1	1	3	2564	CLOSE ONE veil_color=WHITE ==>FREE
25	R122275	1	13800	1	1	32	1	1	5	3977	BROAD FREE ==>veil_color=WHITE
26	R122283	1	18299	1	1	38	1	1	6	5145	FREE stalk_surf_above=SMOOTH ==>veil_color=WHITE
27	R122285	1	18179	1	1	38	1	1	6	5134	stalk_surf_above=SMOOTH veil_color=WHITE ==>FREE
28	R122296	1	20903	1	1	55	1	1	10	6193	FREE stalk_surf_below=SMOOTH ==>veil_color=WHITE
29	R122308	65969	8772	40612	23743	10	23743	23743	23714	1013	FREE ==>ONE veil_color=WHITE

Fig. 6. Union of the ten best rules of the first cluster on mushroom dataset (extract).

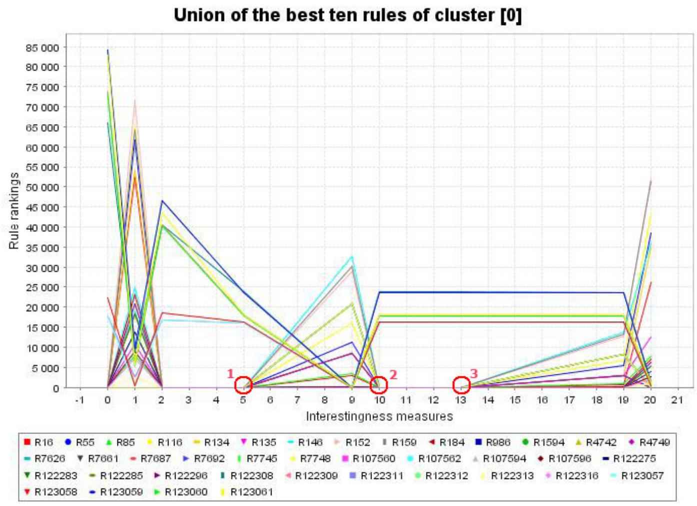


Fig. 7. Plot of the union of the ten best rules of the first cluster on mushroom dataset.

6 Conclusion

We have designed and described some features of a new tool, ARQAT, implementing an exploratory data-analysis approach for IM behavior analysis on a specific dataset.

Technically, ARQAT is written in Java and embeds a set of 14 graphical tools. For exchange facilities, three common file formats are used for importing/exporting the rulesets: PMML (XML data-mining standard), CSV (Excel and SAS) and ARFF (used by WEKA). ARQAT will be freely available at www.polytech.univ-nantes.fr/arqat.

In this paper, we have shown the interest of such an exploratory approach, where the intensive use of graphical and complementary visualizations improves and facilitates data insight for the user.

ARQAT is a first step toward a larger analysis platform in the domain of knowledge quality research. Our future research will investigate the two following directions. First, we will improve the correlation analysis by introducing a better measure than Pearson coefficient 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 best IMs.

References

[Agrawal and Srikant, 1994]R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th VLDB Conference*, pages 487–499, 1994.

[Agrawal *et al.*, 1993]R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of 1993 ACM-SIGMOD International Conference on Management of Data*, pages 207–216, 1993.

[Bayardo and Agrawal, 1999]Jr.R.J. Bayardo and R. Agrawal. Mining the most interestingness rules. In *Proceedings of the Fifth ACM SIGKDD International Confeference On Knowledge Discovery and Data Mining*, pages 145–154, 1999.

[Blake and Merz, 1998]C.L. Blake and C.J. Merz. *UCI Repository of machine learning databases, http://www.ics.uci.edu/~mlearn/MLRepository.html*. University of California, Irvine, Dept. of Information and Computer Sciences, 1998.

[Blanchard *et al.*, 2003]J. Blanchard, P. Kuntz, F. Guillet, and Gras R. Implication intensity: from the basic statistical definition to the entropic version. In *Statistical Data Mining and Knowledge Discovery*, pages 475–493, 2003.

[Blanchard *et al.*, 2004]J. Blanchard, F. Guillet, R. Gras, and H. Briand. Mesurer la qualité des règles et de leurs contraposés avec le taux informationnel tic. In *Revue des Nouvelles Technologies de l'Information (RNTI)*, pages 287–298, 2004.

[Freitas, 1999]A.A. Freitas. On rule interestingness measures. In *Knowledge-Based Systems*, pages 309–315, 1999.

[Gras *et al.*, 2001]R. Gras, P. Kuntz, R. Couturier, and F. Guillet. Une version entropique de l'intensité d'implication pour les corpus volumineux. In *Extraction des Connaissances et Apprentissage (ECA)*, pages 69–80, 2001.

[Gras, 1996]R. Gras. *L'implication statistique - Nouvelle méthode exploratoire de données*. La pensée sauvage édition, 1996.

[Guillaume *et al.*, 1998]S. Guillaume, F. Guillet, and J. Philippé. Improving the discovery of association rules with intensity of implication. In Lecture Notes in Compuper Science, editor, *Proceedings of 2nd European Symp. on Principles of Data Mining and Knowledge Discovery, PKDD'98*, pages 318–327, 1998.

[Guillet, 2004]F. Guillet. Mesures de la qualité des connaissances en ecd. In *Actes des tutoriels, 4ème Conférence francophone Extraction et Gestion des Connaissances (EGC'2004), http://www.isima.fr/ egc2004/*, pages 1–60, 2004.

[Hilderman and Hamilton, 2001]R.J. Hilderman and H.J. Hamilton. *Knowledge Discovery and Measures of Interestingness*. Kluwer Academic Publishers, 2001.

[Lenca *et al.*, 2004]P. Lenca, P. Meyer, P. Picouet, B. Vaillant, and S. Lallich. Evaluation et analyse multi-critères des mesures de qualité des règles d'association. In *Revue des Nouvelles Technologies de l'Information - Mesures de Qualité pour la Fouille de Données, RNTI-E-1*, pages 219–246, 2004.

[Liu *et al.*, 1999]B. Liu, W. Hsu, L. Mun, and H. Lee. Finding interestingness patterns using user expectations. In *IEEE Transactions on knowledge and data mining 11(1999)*, pages 817–832, 1999.

[Padmanabhan and Tuzhilin, 1998]B. Padmanabhan and A. Tuzhilin. A belief-driven method for discovering unexpected patterns. In *Proceedings of the 4th international conference on knowledge discovery and data mining*, pages 94–100, 1998.

[Piatetsky-Shapiro, 1991]G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In G. Piatetsky-Shapiro and W. Frawley, editors, *Knowledge Discovery in Databases*, pages 229–248, 1991.

[Tan *et al.*, 2002]P.N. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *Proc. of the Eighth ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002), pages 32–41, 2002.

[Tan *et al.*, 2004]P.N. Tan, V. Kumar, and J. Srivastava. Selecting the right objective measure for association analysis. In *Information Systems 29(4)*, pages 293–313, 2004.

[Vaillant *et al.*, 2003]B. Vaillant, P. Picouet, and P. Lenca. An extensible platform for rule quality measure benchmarking. In R. Bisdorff, editor, *Human Centered Processes (HCP'2003)*, pages 187–191, 2003.

Appendix 1: IM formulas

N°	Interestingness Measure	$f(n, n_a, n_b, n_{a\overline{b}})$
0	Causal Confidence	$1-\frac{1}{2}(\frac{1}{n_a}+\frac{1}{n_{\overline{b}}})n_{a\overline{b}}$
1	Causal Confirm	$\frac{n_a+n_{\overline{b}}-4n_{a\overline{b}}}{n}$
2	Causal Confirmed-Confidence	$1-\frac{1}{2}(\frac{3}{n_a}+\frac{1}{n_{\overline{b}}})n_{a\overline{b}}$
3	Causal Support	$\frac{n_a+n_{\overline{b}}-2n_{a\overline{b}}}{n}$
4	Collective Strength	$\frac{(n_a-n_{a\overline{b}})(n_{\overline{b}}-n_{a\overline{b}})(n_a n_{\overline{b}}+n_b n_{\overline{a}})}{(n_a n_b+n_{\overline{a}}n_{\overline{b}})(n_b-n_a+2n_{a\overline{b}})}$
5	Confidence	$1-\frac{n_{a\overline{b}}}{n_a}$
6	Conviction	$\frac{n_a n_{\overline{b}}}{nn_{a\overline{b}}}$
7	Cosine	$\frac{n_a-n_{a\overline{b}}}{\sqrt{n_a n_b}}$
8	Dependence	$\lvert\frac{n_{\overline{b}}}{n}-\frac{n_{a\overline{b}}}{n_a}\rvert$
9	Descriptive Confirm	$\frac{n_a-2n_{a\overline{b}}}{n}$
10	Descriptive Confirmed-Confidence	$1-2\frac{n_{a\overline{b}}}{n_a}$
11	EII ($\alpha=1$)	$\sqrt{\varphi\times I^{\frac{1}{2\alpha}}}$
12	EII ($\alpha=2$)	$\sqrt{\varphi\times I^{\frac{1}{2\alpha}}}$
13	Example & Contra-Example	$1-\frac{n_{a\overline{b}}}{n_a-n_{a\overline{b}}}$
14	Gini-index	$\frac{(n_a-n_{a\overline{b}})^2+n_{a\overline{b}}^2}{nn_a}+\frac{(n_b-n_a+n_{a\overline{b}})^2+(n_{\overline{b}}-n_{a\overline{b}})^2}{nn_{\overline{a}}}-\frac{n_b^2}{n^2}-\frac{n_{\overline{b}}^2}{n^2}$
15	Jaccard	$\frac{n_a-n_{a\overline{b}}}{n_b+n_{a\overline{b}}}$
16	J-measure	$\frac{n_a-n_{a\overline{b}}}{n}log_2\frac{n(n_a-n_{a\overline{b}})}{n_a n_b}+\frac{n_{a\overline{b}}}{n}log_2\frac{nn_{a\overline{b}}}{n_a n_{\overline{b}}}$
17	Kappa Cohen's	$\frac{2(n_a n_{\overline{b}}-nn_{a\overline{b}})}{n_a n_{\overline{b}}+n_{\overline{a}}n_b}$
18	Klosgen	$\sqrt{\frac{n_a-n_{a\overline{b}}}{n}}(\frac{n_{\overline{b}}}{n}-\frac{n_{a\overline{b}}}{n_a})$
19	Laplace	$\frac{n_a+1-n_{a\overline{b}}}{n_a+2}$
20	Least Contradiction	$\frac{n_a-2n_{a\overline{b}}}{n_b}$
21	Lift	$\frac{n(n_a-n_{a\overline{b}})}{n_a n_b}$
22	Loevinger	$1-\frac{nn_{a\overline{b}}}{n_a n_{\overline{b}}}$
23	Odds Ratio	$\frac{(n_a-n_{a\overline{b}})(n_{\overline{b}}-n_{a\overline{b}})}{n_{a\overline{b}}(n_b-n_a+n_{a\overline{b}})}$
24	Pavillon	$\frac{n_{\overline{b}}}{n}-\frac{n_{a\overline{b}}}{n_a}$
25	Phi-Coefficient	$\frac{n_a n_{\overline{b}}-nn_{a\overline{b}}}{\sqrt{n_a n_b n_{\overline{a}} n_{\overline{b}}}}$
26	Putative Causal Dependency	$\frac{3}{2}+\frac{4n_a-3n_b}{2n}-(\frac{3}{2n_a}+\frac{2}{n_{\overline{b}}})n_{a\overline{b}}$
27	Rule Interest	$\frac{1}{n}(\frac{n_a n_{\overline{b}}}{n}-n_{a\overline{b}})$
28	Sebag & Schoenauer	$1-\frac{n_{a\overline{b}}}{n_a-n_{a\overline{b}}}$
29	Similarity Index	$\frac{n_a-n_{a\overline{b}}-\frac{n_a n_b}{n}}{\sqrt{\frac{n_a n_b}{n}}}$
30	Support	$\frac{n_a-n_{a\overline{b}}}{n}$
31	TIC	$\sqrt{TI(a\rightarrow b)\times TI(\overline{b}\rightarrow a)}$
32	Yule's Q	$\frac{n_a n_{\overline{b}}-nn_{a\overline{b}}}{n_a n_{\overline{b}}+(n_b-n_{\overline{b}}-2n_a)n_{a\overline{b}}+2n_{a\overline{b}}^2}$
33	Yule's Y	$\frac{\sqrt{(n_a-n_{a\overline{b}})(n_{\overline{b}}-n_{a\overline{b}})}-\sqrt{n_{a\overline{b}}(n_b-n_a+n_{a\overline{b}})}}{\sqrt{(n_a-n_{a\overline{b}})(n_{\overline{b}}-n_{a\overline{b}})}+\sqrt{n_{a\overline{b}}(n_b-n_a+n_{a\overline{b}})}}$