Data Min Knowl Disc (2014) 28:1004–1045
DOI 10.1007/s10618-013-0326-x

Behavior-based clustering and analysis of interestingness measures for association rule mining

C. Tew · C. Giraud-Carrier · K. Tanner · S. Burton

Received: 10 July 2012 / Accepted: 6 June 2013 / Published online: 26 June 2013
© The Author(s) 2013

Abstract A number of studies, theoretical, empirical, or both, have been conducted to provide insight into the properties and behavior of interestingness measures for association rule mining. While each has value in its own right, most are either limited in scope or, more importantly, ignore the purpose for which interestingness measures are intended, namely the ultimate ranking of discovered association rules. This paper, therefore, focuses on an analysis of the rule-ranking behavior of 61 well-known interestingness measures tested on the rules generated from 110 different datasets. By clustering based on ranking behavior, we highlight, and formally prove, previously unreported equivalences among interestingness measures. We also show that there appear to be distinct clusters of interestingness measures, but that there remain differences among clusters, confirming that domain knowledge is essential to the selection of an appropriate interestingness measure for a particular task and business objective.

Keywords Interestingness measures · Clustering · Behavior analysis · Association rule mining

Responsible editor: Bart Goethals.

C. Tew · C. Giraud-Carrier (✉) · K. Tanner · S. Burton
Department of Computer Science, Brigham Young University, Provo, UT 84602, USA
e-mail: cgc@cs.byu.edu

C. Tew
e-mail: csuresaw@hotmail.com

K. Tanner
e-mail: keslert@gmail.com

S. Burton
e-mail: scotthburton@gmail.com

1 Introduction

Association rule mining has joined the ranks of other analysis techniques in the toolbox of researchers and data analysts. Mostly derived from the original Apriori algorithm (Agrawal et al. 1993; Agrawal and Srikant 1994), association rule mining algorithms view data records as sets of attribute-value pairs, or items, and extract rules of association among itemsets, of the form $A \rightarrow B$. The meaning of such rules is that whenever items in A are present in a data record, items in B are likely to be present in that same record. Two simple statistical quantities are generally computed to guide the extraction of association rules: support, which is the probability that a data record contains the items of both A and B, and confidence, which is the probability that a data record contains the items of B given that it contains the items of A.

In general, the user specifies a minimum acceptable level of support, and the association rule mining algorithm proceeds in two phases. In the first phase, frequent itemsets (i.e., sets of items whose support exceeds the user-specified minimum support threshold) are generated, capitalizing on the downward closure property of support, namely, that the support of an itemset is always smaller than that of any of its subsets. In the second phase, all rules consisting of bipartitions of items in frequent itemsets are generated, and ranked according to some general measure of quality or *interestingness*. It is also possible to set a threshold on the values of the interestingness measure so that only rules that exceed that threshold are kept. Confidence is often used as a default interestingness measure. However, there are clearly other valid conceptions of interestingness, and over the years, researchers have designed (objective) interestingness measures that attempt to match the varied notion of "interestingness" expected by different users and/or application domains (e.g., Lan et al. 2006).

Recent surveys of the many available interestingness measures have been conducted by McGarry (2005), and Geng and Hamilton (2006). A number of studies, theoretical, empirical, or both, have also been conducted to provide insight into the properties and behavior of these measures, as well as how one might go about selecting among them for a given application. Unfortunately, most of these studies are either limited in scope or ignore the purpose for which interestingness measures are intended. In particular, what matters most to the practitioner is how an interestingness measure ranks rules so she can get the most interesting rules out of her data. As a result, if two interestingness measures produce the same ranking, they are fundamentally the same as far as the end goal is concerned, and discussing other properties by which they may be distinguished becomes mostly a theoretical and/or philosophical exercise. In other words, if the rankings are the same, there is no way to distinguish among the rules produced, so preferring one interestingness measure over another based on other criteria seems of little practical value. We follow through with this idea here, and investigate similarities among interestingness measures by considering rankings. In particular, we report on an extensive analysis, both theoretical and empirical, that focuses on the behavior of a large number of interestingness measures over a large number of tasks. We characterize the behavior of an interestingness measure by the rankings it produces over sets of rules. By then comparing interestingness measures based on the level of agreement among their rankings, we can produce a behavior-based clustering, which adds to our understanding of these measures.

We note at the onset that it has been shown that, for a number of interestingness measures, there exists an interestingness paradox, in that "given some measure of what makes a rule interesting, discovered rules that satisfy that measure are supposed to be interesting. However, some of the discovered rules are in reality uninteresting since they are expected to hold given some of the other rules that were discovered and simple distributional assumptions" (Padmanabhan 2004). For example, if Lift is the selected interestingness measure and if $A \rightarrow B$ is deemed interesting, then all rules of the form $A, X \rightarrow B$, where X is independent of A and B, will also be listed as interesting since $\text{Lift}(A, X \rightarrow B) = \frac{P(AXB)}{P(AX)P(B)} = \frac{P(AB)P(X)}{P(A)P(X)P(B)} = \frac{P(AB)}{P(A)P(B)} = \text{Lift}(A \rightarrow B)$. Yet, nothing of real interest is being added by these extra rules. Approaches to help reduce the set of rules to those more likely to be interesting include optimal rule discovery (Li 2006), screening during the rule generation process (Webb 2010), false discovery avoidance (Webb 2006), and clustering and summarizing following the generation of rules (Sahar 2002). Recently, some researchers have turned their attention to the idea of mining pattern *sets* rather than individual association rules as another means to deal with redundancy among rules, and ranking patterns using background knowledge (Jaroszewicz and Simovici 2004; Gallo et al. 2007; De Bie et al. 2010; Tatti and Mampaey 2010; Mampaey et al. 2011; Webb 2010, 2011; Spyropoulou and De Bie 2011). Interestingly, by contrast to pattern set mining methods, measuring individual rules has the benefit that one knows when one particular rule is "good enough," while in a pattern set, there often exist complex statistical dependencies that determine whether a pattern is included or not. However, we ignore these issues here as we do not aim to solve the "what is interesting" nor the "what is redundant" questions, but rather the more straightforward and very practical question of "how do all of these interestingness measures differ in action?"

The rest of the paper is organized as follows. In Sect. 2, we discuss related work, including previous studies of interestingness measures as well as the use of clustering for behavior analysis in other areas of data mining. In Sect. 3, we describe our methodology for clustering interestingness measures based on their ranking behavior. In Sect. 4, we discuss our results, prove the indistinguishability of several groups of interestingness measures, and point out a number of insightful findings about relationships among measures. Finally, Sect. 5 concludes the paper.

2 Related work

Much work has been done in the field of data mining regarding the design, use and evaluation of interestingness measures for association rule mining. At a high level, McGarry (2005) and Sahar (2010) both make a clear distinction between objective and subjective measures of interestingness. Objective measures quantify the relationship between the antecedent and the consequent of a rule. Subjective measures require input from a user, basing the interestingness of rules on both the data and the user's beliefs and expectations. In this paper, we focus exclusively on objective measures. We give a brief overview of the most relevant work here. More detailed comparisons and discussions are found throughout the paper.

Vaillant et al. (2004) and Lenca et al. (2007, 2008), while designing a multi-criteria decision system to select among interestingness measures, cluster 20 measures based on both theoretical properties and empirical results over 10 datasets using preorder comparison matrices. They find five clusters, which are congruent with ours on the same interestingness measures. The 20 measures considered are all decreasing with respect to $N_{A\overline{B}}$ (i.e., the number of records containing the items in A but not the items in B), reflecting the authors' bias that the fewer the number of counterexamples, the higher the interestingness. The authors also point out that they observe differences depending on the datasets they consider. We address both of these issues by considering a large number of interestingness measures with no pre-selection, and a large number of varied datasets. Interestingly, the authors state that some measures are "monotonically increasing transformations of the confidence, while [others are] monotonically increasing transformation[s] of the lift" and "measures that are monotonically increasing transformations of [others] inherit [these other measures'] properties," so that "such measures will rank the rules according to the same order." This is also what we bring out and prove formally on a much larger scale.

Huynh et al. (2005, 2006, 2007) were the first to propose the clustering of interestingness measures based on ranking behavior. However, while they do consider up to 35 interestingness measures, they cluster based on at most 2 datasets. They are careful to select datasets with very different properties, but the sample is rather small, calling into question the robustness and validity of their results. In some way, their work is really geared to exploring, analyzing and visualizing one dataset at a time (in their tool called ARQAT), not so much to truly compare interestingness measures. In particular, they do not look at formulas and algebraic relations, except for Yule's Y, Yule's Q, and Odds Ratio, and only there in passing. Our analysis goes further by establishing the "algebraic nature" of the observed "functional dependencies" and making them explicit. On the other hand, we borrow from them the very interesting idea of building high correlation (CG+) and high anti-correlation (CG0) graphs, which we adapt by averaging across all datasets rather than across only those datasets where a minimum correlation threshold is met. There are two other significant differences with our work. On the one hand, they use Pearson correlation while we use Spearman rank correlation; on the other hand, they use $1 - |\rho|$ for distance, while we use $\sqrt{\frac{1-\rho}{2}}$. We believe that rank correlation is a better representative of relative behavior since it is ranks rather than raw values that matter most. We likewise believe that there is value in not confounding perfect correlation and perfect anti-correlation. Many of our findings offer additional supporting evidence—a kind of second, converging opinion—to theirs thus confirming their validity, while others extend them (e.g., inclusion of new interestingness measures in existing clusters and discovery of new clusters of interestingness measures).

Ohsaki et al. (2003, 2004) review 38 interestingness measures with a focus on how they perform on (1) a real application (here, in the medical domain), and (2) against expert judgement (here, a medical expert). Their premise is that interestingness measures "are mainly used to remove meaningless rules rather than to discover really interesting ones for a human user, since they do not include domain knowledge." Hence, the measures they consider are not compared against each other, but rather they are gauged against a human gold standard over a specific set of rules obtained by

classification mining (rather than association rule mining) using C5.0 (Ohsaki et al. 2002). They can then show how some interestingness measures match expert judgement (e.g., Recall), while others do not (e.g., Prevalence). Interestingly, they define Yule's Q and Yule's Y as a function of Odds Ratio, but they do not go the step further to conclude that all three give the same rankings, resulting in redundancy among them.

Yao and Zhong (1999) describe a very nice theoretical framework that uses both set theory and a probabilistic point of view to analyze interestingness measures. Their work covers only 16 measures, which they classify based on the distinction they make between 1-way support and 2-way support. They also establish various algebraic relationships among measures, e.g., 2-way Support = Support $\times$ $\log_2$ Lift (in our naming convention), but, while they realize that "different names have been used for essentially the same measure, or a positive monotonic transformation of the same measure (called order preserving transformation)," they remain theoretical and do not address their measures' behavior in terms of ranking.

Wu et al. (2010) analyze five interestingness measures only, as they argue against the sensitivity of some interestingness measures to null transactions (i.e., transactions that contain neither A nor B). They contend that such sensitivity gives poor results when it comes to interestingness and further assert that the five measures they select are the only ones they know that are null-invariant. The very elegant result is that they can rewrite them all in terms of the mathematical generalized mean, with the appropriate exponent, and are then able to establish a total order among them. While there is no analysis of behavior per se, the paper contains a short discussion of rankings obtained by Coherence, Cosine and Kulczynski on one dataset and eight rules, using Kendall's τ rank correlation. Interestingly, the paper also demonstrates understanding of equivalence based on ranking, as the authors prove statements of the form $M_1(r_i) \leq M_1(r_j) \iff M_2(r_i) \leq M_2(r_j)$.

Kannan and Bhaskaran (2009) consider 38 interestingness measures in their design of a methodology for pruning association rules. In particular, they show how each interestingness measure performs in terms of coverage after poor rules have been removed versus when all rules are included. Unfortunately, it is difficult to infer anything from these results of the relative similarity or correlation among interestingness measures.

Jalali-Heravi and Zaïane (2010) discuss 49 distinct interestingness measures, the largest number surveyed until the present study. Their analysis takes place in the context of association classifiers, which are a subset of association rule miners (i.e., they compute only those rules with the target value as consequent). Their focus is on the impact of the choice of interestingness measures in each phase: rule generation/pruning and rule selection (identifying the best measure for each phase). They do, however, cluster their selected interestingness measures, but the clustering seems incidental to their results, and is not really explored further in the paper. Furthermore, the clustering is based on 16 human-coded, binary properties. The present study is directly complementary as we cluster a large number of measures based upon their behavior, rather than predefined properties.

The work of Abe and Tsumoto (2008) on the behavior of 39 interestingness measures over 32 datasets is similar to ours in principle, but in the context of classification. Indeed, they apply classification rules with PART rather than association rules with Apriori. In addition, they are biased by Gain Ratio (used in PART) while we try to

remove as much bias as possible, and they use raw correlation while we use rank correlation. Finally, they have a very different approach where they average interestingness values over sets of rules first (i.e., one for each dataset) and compute correlations thereafter, while we compute rank correlations first and average across datasets next. Not surprisingly, their results are very different from ours. It is likely that the averaging across rules loses some of the information about the behavior. Behavior should be measured at the rule level rather than at the global level. A similar argument has also been made about Classifier Output Difference versus accuracy in classification learning (Peterson and Martinez 2005; Lee and Giraud-Carrier 2011).

In early work, Tan and Kumar (2000) started from the premise that statistical correlation was desirable, analyzed a few interestingness measures with regard to how closely they matched the notion of correlation, and introduced their own interestingness measure. Later, Tan et al. (2002, 2004) added to their work by performing a comparative study based upon 20 objective measures' properties. Their purpose is to show that is it possible to find the most suitable measure using a small set of well-separated contingency tables that would be equivalent to finding the best measure using the entire data set. They present some key properties that would help analysts select the right measure for a given application. Their results suggest that under certain conditions many measures begin to behave consistently with one another. Some of our results confirm this finding and extend it to a wider range of measures.

Lallich et al. (2007) analyze 15 interestingness measures from a theoretical standpoint. They point out a number of results that are identical to some of our lemmas and theorems about the relationships among the measures studied. Yet, since they look at things from a purely theoretical perspective, they do not make the jump to the equivalence in rankings of said measures, which would allow them to reduce their set of 15 measures to a set of 11 groups of equivalent measures. While Lavrač et al. (1999) focus their attention on classification and inductive logic programming, they also describe a theoretical framework for reasoning about a number of interestingness measures. They even provide a small empirical analysis of rank correlation between a human expert and 4 measures over a set of 10 rules. Again, we combine these theoretical views with a thorough empirical study of ranking behavior over a much larger number of interestingness measures, thus confirming and extending these results.

Finally, we note that we borrowed the idea of behavior-based clustering from its application elsewhere in data mining, where it was used to gain insight and assist in the selection of classification learning algorithms (Lee and Giraud-Carrier 2011). While the approaches are different, our motivation, analysis and results about equivalences among interestingness measures in the context of association rule mining may also be regarded as analogous to the work of Meilă (2012) on logical equivalences of distances in the context of clustering, as well as the work of Fürnkranz and Flach (2005) on search heuristics in the context of covering algorithms.

3 Behavior-based clustering

In this section, we describe our methodology for clustering interestingness measures in the context of association rule mining and their ranking behavior. We begin with a

short section of preliminaries including notation and a brief review of association rule mining.

3.1 Preliminaries

As stated above, association rule mining algorithms view data records as sets of attribute-value pairs, or items, and extract rules of association that relate the presence of certain feature values with that of others. The datasets we consider here match this expectation. They consist of a number of features each with a finite number of possible values that are used to represent objects of a given domain. For example, in a mycology application the features would represent various characteristics of mushrooms, such as the shape of the stalk, the number of rings, the gill color, and the edibility, so that each record would then encode a particular mushroom with its specific values for each of the features. Similarly, in a health application, the features would represent various measurements and behaviors, such as smoking habit, exercise regimen, (discretized) BMI, and general assessment, so that each record would encode a particular individual. In this latter case, the output of association rule mining could include a rule such as $smoking = daily \wedge exercise = none \rightarrow health = poor$, which states that a habit of daily smoking together with a lack of exercise in an individual's life are likely to be associated with poor health for that individual.

Sets of attribute-value pairs, also known as itemsets, are usually abstracted by single letters so that association rules are of the general form: $A \rightarrow B$. Again, the meaning of such rules is that whenever the items in A are present in a data record, the items in B are likely to be present in that same record. The following notation and associated explanation will be useful to the rest of our discussion. We denote by $P(X)$ the probability of itemset X occurring in a record. Since we do not have access to exact probabilities, we estimate $P(X)$ using frequencies. That is, $P(X)$ is given by the ratio of the number of records in a dataset containing the items in X to the total number of records in that dataset. In fact, all probabilities are similarly estimated by frequencies. We denote by $P(XY)$ the probability that both itemsets X and Y occur in a record and by $P(X|Y)$ the conditional probability of the items in X occurring in a record given that the items in Y are in that record. By definition $P(X|Y) = \frac{P(XY)}{P(Y)}$, so that the frequency based estimate of $P(X|Y)$ is the ratio of the number of records containing both X and Y to the number of records containing Y. Finally, we denote by $P(\overline{X})$ the probability that itemset X does not occur in a record. By definition, $P(\overline{X}) = 1 - P(X)$.

The following identities, which are easy to prove from the foregoing definitions, are used extensively in derivations and proofs throughout the remainder of the paper.

$$
\begin{aligned}
P(X\overline{Y}) &= P(X) - P(XY) \\
P(\overline{X}Y) &= P(Y) - P(XY) \\
P(\overline{XY}) &= 1 - P(X) - P(Y) + P(XY)
\end{aligned}
$$

With respect to association rules, two quantities are of particular importance. The *support* of a rule $A \rightarrow B$ is the quantity $P(AB)$, while its *confidence* is the quantity

$P(B|A)$. As stated above, the typical approach to association rule mining is to first extract what are called frequent itemsets, i.e., sets of items whose support exceeds a user-specified minimum threshold. Once these itemsets have been generated, all possible rules are constructed by considering all partitions of each itemset into two itemsets, and those satisfying a user-specified rule quality condition are kept. The default rule quality condition is a minimum value of confidence. Of course, and this is the subject of this paper, a number of other quality, also known as interestingness, measures can be used.

3.2 Interestingness measures for association rule mining

One of the contributions of our study is its attempt at being as broad as possible, both in terms of the interestingness measures considered and the datasets used to evaluate their behavior. As far as interestingness measures are concerned, our search of the literature, including recent surveys, revealed 100 distinctly named measures. Of these, 34 were found to be redundant names, i.e., different names for the same mathematical quantity, leaving 66 unique interestingness measures.

It is worthwhile to note that duplicate names for measures are not always associated with the exact same mathematical formula. For example, Class Correlation Ratio is given as $\frac{P(AB)P(\overline{A})}{P(\overline{A}B)P(A)}$ by Verhein and Chawla (2007), which modulo a few simple algebraic transformations is equal to: $\frac{P(B|A)}{P(B|\overline{A})}$, the formula given for Relative Risk by Ali et al. (1997). Similarly, Confirmed Confidence Descriptive is given as $P(B|A)-P(\overline{B}|A)$ by Kodratoff (2001), which is again easily seen to be equal to $2P(B|A)-1$, the formula given for Ganascia by Ganascia (1991) and Lallich et al. (2007). Such equivalences, if not anticipated, can easily be overlooked however, which may in turn lead to redundancy in analyses. For example, the study of Jalali-Heravi and Zaïane (2010) has four redundant measures (of 53 listed), and the study of Kannan and Bhaskaran (2009) has three redundant measures (of 39 listed). These redundancies have a tendency to propagate as researchers reuse each other's work, which is particularly problematic for oft-cited surveys (e.g., the recent survey of Geng and Hamilton (2006) has one redundant measure). In fact, we became aware of some of these redundancies only after running our clustering, where, as expected, redundant measures were found to cluster together at a distance of 0.

A further five interestingness measures were left out of our study. Hyper-lift and hyper-confidence (Hahsler and Hornik 2007), while reasonable extensions of lift and confidence respectively, rely on the computation of quantiles for which there does not seem to be a closed form. Credibility, Peculiarity, and Gago and Bento's Interestingness, which are used in Ohsaki et al. (2004) and Abe and Tsumoto (2008), are targeted more specifically at the interestingness of classification rules rather than that of association rules.

The following list shows the final set of 61 interestingness measures selected for our study. For each, we give the precise mathematical formula in terms of probabilities, as well as the list of names we have found it referred as in the literature. Note that, wherever possible, we have been careful to go back to, read and reference the original

papers where the measures were first introduced, together with any necessary details about parameter settings and assumptions. Interestingly, this process also revealed a few errors (typos or otherwise) in several influential papers with the risk of propagation through citation (e.g., three errors in the popular survey of Geng and Hamilton (2006), three errors in Ohsaki et al. (2003), one error in Huynh et al. (2007)). Where there are multiple names for a measure, we list them but the bolded name is the one we will use in the rest of the paper. In some instances, the measure first appeared under a different name than the one we have selected.

1. **1-way Support (1WS)** (Yao and Liu 1997)

$$P(B|A)\log_2 \frac{P(B|A)}{P(B)}$$

2. **2-way Support (2WS)** (Yao and Liu 1997; Yao and Zhong 1999)

$$P(AB)\log_2 \frac{P(B|A)}{P(B)}$$

3. **Accuracy (ACC)** (Geng and Hamilton 2006), Causal Support (Kodratoff 2001), Rule Set Accuracy (Lavrač et al. 1999)

$$P(AB) + P(\overline{AB})$$

4. **Added Value (AV)** (Sahar 2003), Pavillon (Huynh et al. 2007), Change of Support (Yao and Zhong 1999; Kannan and Bhaskaran 2009), Centered Confidence (Vaillant et al. 2004), Dependency (Kodratoff 2001), Relative Accuracy (Lavrač et al. 1999)

$$P(B|A) - P(B)$$

   Interestingly, the definition in Kodratoff (2001) calls for the absolute value of the above. However, the signed values make more sense when ranking association rules, since negative values suggest that the presence of A actually hinders the presence of B, making for poor rules.
5. **Chi-square** (χ^2) (Brin et al. 1997a; Tan and Kumar 2000)

$$\frac{(P(AB) - P(A)P(B))^2 N}{P(A)P(\overline{A})P(B)P(\overline{B})}$$

   where N is the number of records in the dataset under consideration. Brin et al. (1997a) use χ^2 to discover correlation rather than association rules, and Tan and Kumar (2000) argue that χ^2 measures only whether there is independence between A and B, but not the strength of that correlation, so that it cannot be used in ranking. However, as we are restricted to relationships between only two itemsets, there is only one degree of freedom, and the larger the value of χ^2 the more likely the two itemsets are correlated. In that sense, and because measures

derived from χ^2 (e.g., Dilated χ^2) have been used for ranking, we choose to retain χ^2 in our analysis.

6. **Collective Strength (CS)** (Aggarwal and Yu 1998)

$$\frac{P(AB) + P(\overline{AB})}{P(A)P(B) + P(\overline{A})P(\overline{B})} \frac{1 - P(A)P(B) - P(\overline{A})P(\overline{B})}{1 - P(AB) - P(\overline{AB})}$$

Aggarwal and Yu define CS as $\frac{\text{Good Events}}{E[\text{Good Events}]} \frac{E[\text{Bad Events}]}{\text{Bad Events}}$. If we denote the events by A and B, then Good Events $= P(AB) + P(\overline{AB})$ (i.e., A and B "agree"), $E[\text{Good Events}] = P(A)P(B) + P(\overline{A})P(\overline{B})$ (i.e., the probability of chance agreement between A and B), and Bad Events are simply the complement of Good Events, which gives the above formula.

7. **Complement Class Support (CCS)** (Arunasalam and Chawla 2006)

$$\frac{P(A\overline{B})}{P(\overline{B})}$$

Unlike most of our other selected interestingness measures, CCS produces small values for strong rules. Hence, in order to be consistent in the rankings, we use $-$ CCS in our experiments.

8. **Conditional Entropy (CE)** (Blanchard et al. 2005b)

$$-P(B|A)\log_2 P(B|A) - P(\overline{B}|A)\log_2 P(\overline{B}|A)$$

9. **Confidence (CON)** (Agrawal et al. 1993), Absolute Support (Yao and Zhong 1999), Precision (Abe and Tsumoto 2008; Ohsaki et al. 2004), Rule Accuracy (Lavrač et al. 1999)

$$P(B|A)$$

10. **Confidence Causal (CDC)** (Kodratoff 2001)

$$\frac{1}{2}(P(B|A) + P(\overline{A}|\overline{B}))$$

11. **Confirm Causal (CRC)** (Kodratoff 2001)

$$P(AB) + P(\overline{AB}) - 2P(A\overline{B})$$

12. **Confirm Descriptive (CRD)** (Kodratoff 2001)

$$P(AB) - P(A\overline{B})$$

13. **Confirmed Confidence Causal (CCC)** (Kodratoff 2001)

$$\frac{1}{2}(P(B|A) + P(\overline{A}|\overline{B})) - P(\overline{B}|A)$$

14. **Conviction (CVC)** (Brin et al. 1997b)

$$\frac{P(A)P(\overline{B})}{P(A\overline{B})}$$

15. **Correlation Coefficient (CCO)** (Geng and Hamilton 2006)

$$\frac{P(AB) - P(A)P(B)}{\sqrt{P(A)P(B)P(\overline{A})P(\overline{B})}}$$

16. **Cosine (COS)** (Geng and Hamilton 2006), IS (Tan and Kumar 2000)

$$\frac{P(AB)}{\sqrt{P(A)P(B)}}$$

17. **Coverage (COV)** (Geng and Hamilton 2006), Generality (Yao and Zhong 1999)

$$P(A)$$

18. **Dilated Chi-square** (**D**χ^2) (Lan et al. 2004)

$$\left(\frac{P(A)P(\overline{A})P(B)P(\overline{B})}{(\min(\min(P(A),\, P(\overline{A})),\, \min(P(B),\, P(\overline{B})))\, \min(\max(P(A),\, P(\overline{A})),\, \max(P(B),\, P(\overline{B})))^2}\right)^{\alpha} \chi^2$$

Lan and colleagues state that "the parameter α is used to control the impact of global and local maximum χ^2 values and tuned for different classification problems." They then go on to choose a value of 0.5 for their specific example since it gives a result "which is much more reasonable to our intuition." This makes it sound like they picked α to meet their needs. Later in the experiments, they state that "the average error rate…is only 12.81 % if the best parameter α is selected for each dataset." There is no guidance as to how to select α and it seems to be dependent on the dataset. Since we are only interested in relative performance and do not wish to optimize for each dataset, we select a neutral value of $\alpha = 1$.

19. **Directed Information Ratio (DIR)** (Blanchard et al. 2005b)

$$\begin{cases} -\infty & \text{if } P(B)=1 \\ 0 & \text{if } P(B) \le \frac{1}{2} \text{ and } P(B|A) \le \frac{1}{2} \\ 1+P(B|A)\log_2 P(B|A)+P(\overline{B}|A)\log_2 P(\overline{B}|A) & \text{if } P(B) \le \frac{1}{2} \text{ and } P(B|A) > \frac{1}{2} \\ 1+\dfrac{1}{P(B)\log_2 P(B) + P(\overline{B})\log_2 P(\overline{B})} & \text{if } P(B) > \frac{1}{2} \text{ and } P(B|A) \le \frac{1}{2} \\ 1-\dfrac{P(B|A)\log_2 P(B|A) + P(\overline{B}|A)\log_2 P(\overline{B}|A)}{P(B)\log_2 P(B) + P(\overline{B})\log_2 P(\overline{B})} & \text{if } P(B) > \frac{1}{2} \text{ and } P(B|A) > \frac{1}{2} \end{cases}$$

20. **Entropic Implication Intensity 1 (EII1)** (Blanchard et al. 2003; Huynh et al. 2007)
21. **Entropic Implication Intensity 2 (EII2)** (Blanchard et al. 2003; Huynh et al. 2007)

$$\sqrt{\mathrm{IIM}\left(\left(1-H^{\alpha}_{B|A}\right)\left(1-H^{\alpha}_{\overline{A}|\overline{B}}\right)\right)^{\frac{1}{2\alpha}}}$$

where $H_{X|Y} = -P(X|Y)\log_2 P(X|Y) - P(\overline{X}|Y)\log_2 P(\overline{X}|Y)$. We consider the same two versions here as the authors of this measure did, i.e., $\alpha = 1$ and $\alpha = 2$.

22. **Example and Counterexample Rate (ECR)** (Vaillant et al. 2004)

$$1 - \frac{P(A\overline{B})}{P(AB)}$$

23. **F-measure (FM)** (Ohsaki et al. 2003)

$$\frac{2P(A|B)P(B|A)}{P(A|B)+P(B|A)}$$

24. **Ganascia (GAN)** (Ganascia 1991), Confirmed Confidence Descriptive (Kodratoff 2001)

$$2P(B|A) - 1$$

25. **Gini Index (GI)** (Breiman et al. 1984)

$$P(A)(P(B|A)^2 + P(\overline{B}|A)^2) + P(\overline{A})(P(B|\overline{A})^2 + P(\overline{B}|\overline{A})^2) - P(B)^2 - P(\overline{B})^2$$

26. **Goodman–Kruskal (GK)** (Goodman and Kruskal 1954)

$$\frac{\max(P_1, P_2)+\max(P_3, P_4)+\max(P_1, P_3)+\max(P_2, P_4)-\max(P(A), P(\overline{A}))-\max(P(B), P(\overline{B}))}{2-\max(P(A), P(\overline{A}))-\max(P(B), P(\overline{B}))}$$

where $P_1 = P(AB)$, $P_2 = P(A\overline{B})$, $P_3 = P(\overline{A}B)$, and $P_4 = P(\overline{AB})$

27. **Implication Index (IIN)** (Lerman et al. 1981a,b; Ritschard and Zighed 2006)

$$\sqrt{N}\frac{P(A\overline{B}) - P(A)P(\overline{B})}{\sqrt{P(A)P(\overline{B})}}$$

28. **Indice Probabiliste d'Ecart d'Equilibre (IPE)** (Blanchard et al. 2005a)

$$1 - \frac{1}{2^{N_A}}\sum_{k=0}^{N_{A\overline{B}}}\binom{N_A}{k}$$

where N_A is the number of records containing the items in A and $N_{A\overline{B}}$ is the number of records containing the items in A but not the items in B. The actual formula is $P(|X \cap \overline{B}| > N_{A\overline{B}} \mid H_0)$, which is given by the above in the case of drawing random sets with replacement (i.e., $|X \cap \overline{B}|$ is binomial with parameters N_A and $\frac{1}{2}$ (Blanchard et al. 2005a)).

29. **Information Gain (IG)** (Geng and Hamilton 2006)

$$\log_2 \frac{P(AB)}{P(A)P(B)}$$

30. **Intensity of Implication (IIM)** (Gras and Larher 1992)

$$\frac{1}{2} - \frac{1}{2}\,\mathrm{sgn}\left(\frac{\mathrm{IIN}}{\sqrt{2}}\right)\sqrt{1 - e^{-\left(\frac{\mathrm{IIN}}{\sqrt{2}}\right)^2 \frac{\frac{4}{\pi}+0.147\left(\frac{\mathrm{IIN}}{\sqrt{2}}\right)^2}{1+0.147\left(\frac{\mathrm{IIN}}{\sqrt{2}}\right)^2}}}$$

The actual formula is $P(\mathcal{N}(0,1) \geq \mathrm{IIN}) = \frac{1}{\sqrt{2\pi}} \int_{\mathrm{IIN}}^{+\infty} e^{\frac{-t^2}{2}} dt = 1 - \int_{-\infty}^{\mathrm{IIN}} e^{\frac{-t^2}{2}} dt = 1 - (\frac{1}{2} + \frac{1}{2}\,\mathrm{erf}(\frac{\mathrm{IIN}}{\sqrt{2}})) = \frac{1}{2} - \frac{1}{2}\,\mathrm{erf}(\frac{\mathrm{IIN}}{\sqrt{2}})$. The error function, erf, does not have a closed form solution. We use the approximation to a maximum error of $1.2\ 10^{-4}$ proposed by Winitzki (2003, 2008).

31. **Interestingness Weighting Dependency (IWD)** (Gray and Orlowska 1998)

$$\left(\left(\frac{P(AB)}{P(A)P(B)}\right)^l - 1\right) P(AB)^m$$

Gray and Orlowska state that "l and m are parameters to weight the relative importance of the two measures" (p. 136) and in their application "the parameters of the algorithm were determined through experimentation and set to the following values: $l = 2$, $m = 1$" (p. 140). These values apply only to one dataset they studied. It is not desirable to run experiments for each of our datasets to optimize these values. Rather we pick the values $l = m = 1$, which appear to be neutral, giving the same weight to both. Since we use the same values for all datasets, there should not be any unfair advantage to any.

32. **Jaccard (JAC)** (Jaccard 1901), Mutual Support (Yao and Zhong 1999), Coherence (Kannan and Bhaskaran 2009; Wu et al. 2010)

$$\frac{P(AB)}{P(A) + P(B) - P(AB)}$$

33. **J-measure (JM)** (Blachman 1968; Smyth and Goodman 1992), Information Content (Yao and Zhong 1999)

$$P(AB)\log_2 \frac{P(B|A)}{P(B)} + P(A\overline{B})\log_2 \frac{P(\overline{B}|A)}{P(\overline{B})}$$

34. **Kappa** (κ) (Cohen 1960)

$$\frac{P(B|A)P(A) + P(\overline{B}|\overline{A})P(\overline{A}) - P(A)P(B) - P(\overline{A})P(\overline{B})}{1 - P(A)P(B) - P(\overline{A})P(\overline{B})}$$

Cohen actually introduces κ as a measure of agreement between judges, defined by $\frac{p_0 - p_c}{1 - p_c}$, where p_0 is the proportion of items on which the judges agree and p_c the proportion of items on which agreement is expected by chance. If A and B represent the judges and there are only two items, then $p_0 = P(AB) + P(\overline{A}\overline{B})$ and $p_c = P(A)P(B) + P(\overline{A})P(\overline{B})$, which gives the above formula.

35. **Klösgen (KLO)** (Klösgen 1996)

$$\sqrt{P(A)}(P(B|A) - P(B))$$

Klösgen actually introduced the more general formula: $P(A)^{\alpha}(P(B|A) - P(B))$, which allowed him to compare with others (e.g., when $\alpha = 1$, this is identical to Piatetsky-Shapiro). His specific formula (from 1992) is with $\alpha = \frac{1}{2}$.

36. **K-measure (KM)** (Ohsaki et al. 2003)

$$P(B|A)\log_2 \frac{P(B|A)}{P(B)} + P(\overline{B}|\overline{A})\log_2 \frac{P(\overline{B}|\overline{A})}{P(\overline{B})} - P(B|A)\log_2 \frac{P(B|A)}{P(\overline{B})} - P(\overline{B}|\overline{A})\log_2 \frac{P(\overline{B}|\overline{A})}{P(B)}$$

There is an error on the J-measure in the paper where the K-measure is introduced. We checked the source for the J-measure (see above), so we know the formula in their paper is indeed incorrect. We were tempted to make the same correction to their K-measure formula, since it is parallel to the J-measure (namely, the "given not A" in the conditional probabilities may actually be "given A" in the numerator). However, we have no grounds for such a change and choose to stick with the author's formula, since they introduced it.

37. **Kulczynski 1 (KU1)** (Kulczynski 1927), Agreement–Disagreement Index (Plasse et al. 2007)

$$\frac{P(AB)}{P(A\overline{B}) + P(\overline{A}B)}$$

38. **Kulczynski 2 (KU2)** (Kulczynski 1927; Wu et al. 2010)

$$\frac{1}{2}\left(\frac{P(AB)}{P(A)} + \frac{P(AB)}{P(B)}\right)$$

39. **Laplace Correction (LAC)** (Clark and Boswell 1991)

$$\frac{NP(AB) + 1}{NP(A) + k}$$

All uses of Laplace Correction in the context of interestingness measures that we have surveyed set $k = 2$. We follow this convention here as well.

40. **Least Contradiction (LEC)** (Azé and Kodratoff 2002)

$$\frac{P(AB) - P(A\overline{B})}{P(B)}$$

41. **Leverage (LEV)** (Geng and Hamilton 2006)

$$P(B|A) - P(A)P(B)$$

There is a significant chance that Leverage is actually an erroneous version of Piatetsky-Shapiro, as several authors seem to use the name Leverage together with the above formula but reference Piatetsky-Shapiro (1991). Furthermore, the formula is not particularly intuitive. However, since many studies include both Leverage and Piatetsky-Shapiro, we choose to retain Leverage in our analysis.

42. **Lift (LIF)** (Geng and Hamilton 2006), Brin's Interest (Brin et al. 1997b), Interest (Tan and Kumar 2000; Kodratoff 2001), Independence (Yao and Zhong 1999)

$$\frac{P(B|A)}{P(B)}$$

43. **Loevinger (LOE)** (Loevinger 1947; Bertrand and Bel Mufti 2006), Certainty Factor (Berzal et al. 2002; Kannan and Bhaskaran 2009), Satisfaction (Lavrač et al. 1999)

$$1 - \frac{P(A\overline{B})}{P(A)P(\overline{B})}$$

44. **Logical Necessity (LON)** (Duda et al. 1981; Kamber and Shinghal 1996)

$$\frac{P(\overline{A}|B)}{P(\overline{A}|\overline{B})}$$

Small values of LON suggest that A is logically necessary for B, while large values of LON indicate that the absence of A is encouraging for B. Hence, in order to be consistent in the rankings, we treat LON as CCS and use $-$LON in our experiments.

45. **Mutual Information (MI)** (Blanchard et al. 2005b), 2-way Support Variation (Geng and Hamilton 2006)

$$P(AB)\log_2 \frac{P(AB)}{P(A)P(B)} + P(A\overline{B})\log_2 \frac{P(A\overline{B})}{P(A)P(\overline{B})} + P(\overline{A}B)\log_2 \frac{P(\overline{A}B)}{P(\overline{A})P(B)} + P(\overline{AB})\log_2 \frac{P(\overline{AB})}{P(\overline{A})P(\overline{B})}$$

46. **Normalized Mutual Information (NMI)** (Geng and Hamilton 2006)

$$\frac{\text{MI}}{-P(A)\log_2 P(A) - P(\overline{A})\log_2 P(\overline{A})}$$

47. **Odd Multiplier (OM)** (Geng and Hamilton 2006), Bayes Factor (Lenca et al. 2007), Logical Sufficiency (Duda et al. 1981; Kamber and Shinghal 1996)

$$\frac{P(AB)P(\overline{B})}{P(B)P(A\overline{B})}$$

48. **Odds Ratio (OR)** (Mosteller 1968)

$$\frac{P(AB)P(\overline{AB})}{P(A\overline{B})P(\overline{A}B)}$$

49. **Piatetsky-Shapiro (PS)** (Piatetsky-Shapiro 1991), Leverage-2 (Kannan and Bhaskaran 2009), Rule Interest (Huynh et al. 2007), Novelty (Lavrač et al. 1999), Weighted Relative Accuracy (Lavrač et al. 1999)

$$N(P(AB) - P(A)P(B))$$

50. **Prevalence (PRE)** (Geng and Hamilton 2006), Expected Confidence (McGarry 2005)

$$P(B)$$

51. **Putative Causal Dependency (PCD)** (Kodratoff 2001; Huynh et al. 2007)

$$\frac{1}{2}(P(B|A)-P(B)) + (P(\overline{A}|\overline{B})-P(\overline{A}))-(P(\overline{B}|A)-P(\overline{B}))-(P(A|\overline{B})-P(A))$$

52. **Recall (REC)** (Geng and Hamilton 2006), Local Support (Geng and Hamilton 2006), Sensitivity (Lavrač et al, 1999; Kannan and Bhaskaran 2009)

$$P(A|B)$$

53. **Relative Risk (REL)** (Ali et al. 1997), Class Correlation Ratio (Verhein and Chawla 2007; Jalali-Heravi and Zaïane 2010)

$$\frac{P(B|A)}{P(B|\overline{A})}$$

54. **Sebag–Schoenauer (SS)** (Sebag and Schoenauer 1988)

$$\frac{P(AB)}{P(A\overline{B})}$$

55. **Specificity (SPE)** (Geng and Hamilton 2006), Negative Reliability (Lavrač et al. 1999)

$$P(\overline{B}|\overline{A})$$

56. **Support (SUP)** (Agrawal et al. 1993)

$$P(AB)$$

57. **Theil Uncertainty Coefficient (TUC)** (Blanchard et al. 2005b)

$$\frac{\text{MI}}{-P(B)\log_2 P(B) - P(\overline{B})\log_2 P(\overline{B})}$$

58. **TIC** (Blanchard et al. 2004)

$$\sqrt{\text{DIR}(A \Rightarrow B)\,\text{DIR}(\overline{B} \Rightarrow A)}$$

59. **Yule's Q (YQ)** (Yule 1900)

$$\frac{P(AB)P(\overline{A}\overline{B}) - P(A\overline{B})P(\overline{A}B)}{P(AB)P(\overline{A}\overline{B}) + P(A\overline{B})P(\overline{A}B)}$$

60. **Yule's Y (YY)** (Yule 1912)

$$\frac{\sqrt{P(AB)P(\overline{A}\overline{B})} - \sqrt{P(A\overline{B})P(\overline{A}B)}}{\sqrt{P(AB)P(\overline{A}\overline{B})} + \sqrt{P(A\overline{B})P(\overline{A}B)}}$$

Yule actually introduced this measure as the function ω defined by the equivalent formula $\frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}$, where $\kappa = \frac{P(A\overline{B})P(\overline{A}B)}{P(AB)P(\overline{A}\overline{B})}$.

61. **Zhang (ZHA)** (Zhang 2000)

$$\frac{P(AB) - P(A)P(B)}{\max(P(AB)(1 - P(B)), P(B)(P(A) - P(AB))}$$

As a comparison of the current study with past studies of interestingness measures, Table 1 shows which of our measures appeared in which significant prior study, together with a total for each measure and a total for each study. It is clear that (1) our study is significantly more extensive than any of its predecessors and (2) some interestingness measures are better represented than others in the literature: 11 measures appear only once in the survey/studies we examined, but 33 appear 5 or more times.

3.3 Experimental setup

Since the results of clustering may be sensitive to the data used, we also consider a large and diverse number of datasets. This reduces the risk of selection bias as well as increases the reliability of the results so they are a more accurate reflection of how each metric performs. We consider a total of 110 datasets, as shown in Table 2.

Table 1 Interestingness measures: previous studies

	G	T	A	J	K	Y	O	H	L	B	P	W	
1WS	✓		✓	✓		✓	✓						5
2WS	✓		✓	✓		✓	✓						5
ACC	✓		✓	✓	✓		✓						5
AV	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		10
χ^2			✓	✓			✓					✓	4
CS	✓	✓	✓	✓	✓		✓	✓					7
CCS			✓	✓									2
CE										✓			1
CON	✓	✓	✓	✓	✓	✓	✓	✓	✓				9
CDC				✓				✓					2
CRC				✓				✓					2
CRD				✓				✓					2
CCC				✓				✓					2
CVC	✓	✓	✓	✓	✓		✓	✓	✓				8
CCO	✓	✓	✓	✓	✓		✓	✓	✓				8
COS	✓	✓	✓	✓	✓		✓	✓				✓	8
COV	✓		✓		✓	✓	✓						5
Dχ^2				✓									1
DIR										✓			1
EII1								✓					1
EII2								✓	✓				2
ECR	✓			✓	✓			✓	✓				5
FM			✓	✓									2
GAN				✓				✓					2
GI	✓	✓	✓	✓	✓		✓	✓		✓			8
GK	✓	✓		✓			✓						4
IIN				✓				✓	✓				3
IPE								✓					1
IG	✓			✓	✓				✓				4
IIM				✓				✓	✓				3
IWD	✓		✓	✓		✓	✓						5
JAC	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	10
JM	✓	✓	✓	✓	✓	✓	✓	✓		✓			9
κ		✓	✓	✓			✓	✓	✓				6
KLO	✓	✓	✓	✓	✓	✓	✓	✓					8
KM				✓			✓						2
KU1											✓		1
KU2												✓	1
LAC	✓	✓	✓	✓	✓		✓	✓	✓				8
LEC	✓			✓	✓			✓	✓				5

Table 1 continued

	G	T	A	J	K	Y	O	H	L	B	P	W	
LEV	✓		✓	✓	✓		✓						5
LIF	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	11
LOE	✓	✓	✓	✓	✓		✓	✓	✓		✓		9
LON						✓							1
MI	✓	✓	✓	✓		✓	✓			✓			7
NMI	✓						✓						2
OM	✓			✓	✓	✓			✓		✓		6
OR	✓	✓	✓	✓	✓		✓	✓					7
PS	✓	✓	✓	✓	✓	✓	✓		✓				8
PRE	✓		✓		✓		✓						4
PCD								✓					1
REC	✓		✓	✓	✓		✓						5
REL	✓		✓	✓	✓	✓	✓						6
SS	✓			✓	✓			✓	✓				5
SPE	✓		✓	✓	✓		✓						5
SUP	✓	✓	✓	✓	✓	✓	✓	✓	✓				9
TUC										✓			1
TIC								✓					1
YQ	✓	✓	✓	✓	✓		✓	✓					7
YY	✓	✓	✓	✓	✓		✓	✓					7
ZHA	✓			✓	✓				✓				4
	37	21	33	47	31	16	34	33	19	6	6	5	

Significant prior studies are shown with the measures they contain, together with counts for each measure and each study. References are abbreviated as follows: *G* Geng and Hamilton (2006), *T* Tan et al. (2002), *A* Abe and Tsumoto (2008), *J* Jalali-Heravi and Zaïane (2010), *K* Kannan and Bhaskaran (2009), *Y* Yao and Zhong (1999), *O* Ohsaki et al. (2003), *H* Huynh et al. (2007), *L* Lenca et al. (2007), *B* Blanchard et al. (2005b), *P* Plasse et al. (2007), *W* Wu et al. (2010)

45 datasets are from the Gene Expression Machine Learning Repository (Stiglic and Kokol 2009), 53 datasets are from the UCI Machine Learning Repository (Asuncion and Newman 2007), and 12 datasets are from the Multi-class Protein Fold Recognition Data (see http://cs.odu.edu/~sji/resources.html), where we treated training and test datasets as individual datasets for clustering purposes here. Note that these datasets are generally used in the context of classification learning, but they can naturally be used for association learning by treating the class attribute as any other attribute.

Numerical attributes were discretized into three bins of equal frequency. No other data transformations were performed. The following summarizes some of the characteristics of our selected datasets.

– *Number of examples* 35 datasets have between 16 and 200 examples; 36 datasets have between 201 and 500 examples; 18 datasets have between 501 and 1,000 examples; and 21 datasets have between 1,001 and 48,842 examples.

Table 2 Datasets

UCI ML Repository	
Abalone, Adult, Balance Scale, Balloons adult-stretch, Balloons adult + stretch, Balloons yellow-small, Balloons small-yellow + adult-stretch, Blood Transfusion, Breast Cancer, Breast Cancer W-Diagnostic, Breast Cancer W-Original, Breast Cancer W-Prognostic, Car, Chess (kr-vs-kp), Congressional Voting Records, Connect-4, Diabetes, Ecoli, Glass Identification, Haberman's Survival, Image Segmentation, Internet Ads, Ionosphere, Iris, Lenses, Letter Recognition, Liver Disorders (BUPA), Low Resolution Spectrometer, Lung Cancer, Lymphography, MAGIC Gamma Telescope, Molecular Biology (Splice-junction), Mushroom, Nursery, Ozone Level Detection, Page Blocks, Parkinsons String Removed, Pen-Based Recognition, Pima Indians Diabetes, Poker Hand, Primary Tumor, Sonar, Soybean (Large), Spambase, SPECT Heart, Statlog (Heart), Statlog (Landsat), Statlog (Shuttle), Tic-Tac-Toe, Waveform-5000, Wine, Yeast, Zoo	53
Gene Expression ML Repository	
AP Breast Colon, AP Breast Kidney, AP Breast Lung, AP Breast Omentum, AP Breast Ovary, AP Breast Prostate, AP Breast Uterus, AP Colon Kidney, AP Colon Lung, AP Colon Omentum, AP Colon Ovary, AP Colon Prostate, AP Colon Uterus, AP Endometrium Breast, AP Endometrium Colon, AP Endometrium Kidney, AP Endometrium Lung, AP Endometrium Omentum, AP Endometrium Ovary, AP Endometrium Prostate, AP Endometrium Uterus, AP Lung Kidney, AP Lung Uterus, AP Omentum Kidney, AP Omentum Lung, AP Omentum Ovary, AP Omentum Prostate, AP Omentum Uterus, AP Ovary Kidney, AP Ovary Lung, AP Ovary Uterus, AP Prostate Kidney, AP Prostate Lung, AP Prostate Ovary, AP Prostate Uterus, AP Uterus Kidney, OVA Breast, OVA Colon, OVA Endometrium, OVA Kidney, OVA Lung, OVA Omentum, OVA Ovary, OVA Prostate, OVA Uterus	45
Multi-class Protein Folding	
c-train, c-test, h-train, h-test, p-train, p-test, s-train, s-test, v-train, v-test, z-train, z-test	12

The list of 110 datasets selected for our analysis of the ranking behavior of interestingness measures

- *Number of attributes* 36 datasets have between 4 and 20 attributes; 30 datasets have between 21 and 100 attributes; 20 datasets have between 101 and 200 attributes; 19 datasets have between 201 and 300 attributes; and 5 datasets have between 301 and 1,559 attributes.
- *Number of attribute-value pairs* 23 datasets have between 8 and 30 attribute-value pairs; 28 datasets have between 31 and 100 attribute-value pairs; 25 datasets have between 101 and 500 attribute-value pairs; 31 datasets have between 501 and 1,000 attribute-value pairs; and 3 datasets have between 1,001 and 3,121 attribute-value pairs. (This statistic gives an idea of the size of the search space for the association mining algorithm since each item in our context consists of an attribute-value pair.)

It is clearly unfeasible, and most likely unnecessary, to compare our interestingness measures over the entire set of rules generated by association rule mining. A simple, practical approach would be to perform association rule mining and consider only the top N rules it generates. There is a serious risk of bias in this approach, however. Indeed, most algorithms generate only rules that match the minimum confidence threshold and order them in decreasing order of confidence. Therefore, considering the top N rules may give unfair advantage to some of the interestingness measures, namely confidence and its derivatives. The problem is made worse when the value of N is small, as most rules will then likely have both very high support and very high confidence, and hence may artificially score similarly across most interestingness measures. To alleviate this

difficulty and reduce the amount of bias, we (1) run the standard Apriori algorithm as implemented in Weka (Witten and Eibe 2000) with a minimum confidence threshold of 0.0; (2) set the minimum support threshold low enough for each dataset to generate at least 1,000 rules; and (3) randomly select 100 rules from the result set. In doing so, we keep computational costs reasonable while increasing diversity and reducing bias in the set of rules used in our analysis.

Each interestingness measure can be viewed as a function that takes a rule as input and produces a numerical score based upon probabilities about the premise and consequent of the rule. Because we use frequencies as estimates of probabilities, some formulas may contain indeterminate forms (i.e., $\frac{0}{0}$). In such cases, we consider the function to be undefined and the corresponding rule to be uninteresting by default. Hence, we assign the function the smallest value of $-\infty$, which pushes the rule to the end of the ranked list. Note that while indeterminate forms are not widespread, the way we handle them means that they mix with rules whose true value is $-\infty$, which may have minor effects on rank correlations. Consistent with common practice, we set $0 \log_2 0 = 0$.

3.4 Distance computation

A matrix of pairwise distances between interestingness measures is computed using the procedure described in Algorithm 1. For each dataset (Line 1), 100 rules are selected at random as described above, and labeled arbitrarily from 1 to 100 (Line 2). For each interestingness measure (Line 3), Lines 4–6 compute the raw scores assigned by that measure to each of the 100 randomly selected rules. In Line 7, these scores are ordered in decreasing values. Since raw scores are not always directly comparable (i.e., they can range over significantly different values from one interestingness measure to another), we opt to have interestingness measures place value on rules in terms of a ranking system rather than a numerical score. Hence, in Line 8, the raw scores are replaced by corresponding ranks, accounting for ties as follows. All rules with the same value are assigned a rank that is the average of what their ranks would be if ties were ignored. For example, if the values were 0.9, 0.7, 0.7, 0.6, 0.4, 0.4, 0.4, and 0.2, then the resulting ranks would be 1, 2.5, 2.5, 4, 6, 6, 6, and 8. In Line 8, the rules are re-ordered by increasing label values to guarantee consistency across interestingness measures (i.e., the first rule is the same for all measures, with its corresponding rank by that measure, etc.). By Line 10, each of the 100 rules has a corresponding rank for each of the interestingness measures, and it is now possible to compare them in terms of rank correlation.

In Lines 11–16, a Spearman rank correlation coefficient is computed for each pair of distinct interestingness measures (Line 13). Spearman rank correlation gives a score of +1 to identical rankings, a score of -1 to perfectly reversed rankings, and a score between -1 and +1 in all other cases. There is one important technical point about correlations that requires our attention here as it affects our algorithm. It is an established fact that most correlation measures are not additive (Hill and Lewicki 2007), and that averaging Pearson correlation coefficients thus creates bias and should be used with caution (if at all) in test statistics (Silver and Dunlap 1987; Corey et al. 1998). The same result holds, of course for Spearman rank correlation (Fieller et al.

Algorithm 1: Distance matrix computation

Data: A set D of datasets and a set I of interestingness measures
Result: A $|I| \times |I|$ distance matrix M

```
1  for d ∈ D do
2      A ← 100 labeled association rules generated by running Apriori on d;
3      for i ∈ I do
4          for a ∈ A do
5              v_i(a) ← value of i on a;
6          end
7          Order the pairs (a, v_i(a)) by decreasing values of v_i(a);
8          Create new pairs (a, r_i(a)) by replacing v_i(a) by its rank r_i(a);
9          Order the pairs (a, r_i(a)) by increasing label values;
10     end
11     for i ∈ I do
12         for j ∈ I, j ≠ i do
13             s^d_ij ← Spearman rank correlation of {(a, r_i(a))} and {(a, r_j(a))};
14             z^d_ij ← Fisher's z transform of s^d_ij;
15         end
16     end
17 end
18 for i ∈ I do
19     for j ∈ I, j ≠ i do
20         Z_ij ← Average over D of the z^d_ij 's;
21         S_ij ← Fisher's back-transform of Z_ij;
22         M_ij ← sqrt((1 − S_ij)/2);
23     end
24     M_ii ← 0;
25 end
```

1957). However, for both Pearson and Spearman rank correlations, it has also been shown that this bias can be reduced by first using Fisher's z transformation, averaging the transformed values, and back-transforming the result into an aggregate correlation value. Recall that we perform the above computations over a number of datasets (Line 1). The purpose of doing so is to limit the impact of the choice of datasets on our results, and so we will wish to average over these datasets. Hence, in Line 14, rank correlation values are subjected to Fisher's z transformation, i.e., $z = 0.5 \log \frac{1+s}{1-s}$.

By Line 17, each z^d_{ij} is a $|I| \times |I|$ matrix that, for each dataset d, captures information about the amount of similarity/dissimilarity of behavior among interestingness measures, where behavior is measured in terms of rule rankings. It is now possible to compute a summary $|I| \times |I|$ distance matrix as shown in Lines 18–24. In Line 20, the transformed correlation values are averaged, and in Line 21, they are back-transformed, i.e., $s = \frac{e^{2z}-1}{e^{2z}+1}$ into correlation values. Subtracting these values from 1 and dividing the result by 2 transforms the correlations, which range between -1 (opposite) and +1 (identical) into distances, which range between 0 (identical) and 1 (complete opposite). However, in order to obtain a true distance metric for purposes of clustering, we must take the square root of these distance values, as shown in Line 22 (Greenacre and Primicerio 2013). Finally, in Line 24, we trivially set the distance

between all interestingness measures and themselves to 0. The final matrix M is a relatively unbiased, experiment-based summary of the relative similarity among the rankings produced by the interestingness measures in I.

Equipped with the distance matrix M, we can proceed to cluster interestingness measures. Since we are interested in gaining insight into the relative behavior of various interestingness measures, we choose to use hierarchical agglomerative clustering (HAC) (Johnson 1967; Jain and Dubes 1988). Indeed, one of the advantages of HAC is that it produces a complete sequence of nested clusterings, by starting with each interestingness measure in its own cluster and successively merging the two closest clusters into a new cluster until a single cluster containing all of the measures is obtained. In addition to a distance between individual items, HAC requires a distance between clusters. Several approaches are possible, the more common of which are complete linkage (maximum distance between all pairs of objects across clusters), single linkage (minimum distance between all pairs of objects across clusters), and average linkage (average of all inter-cluster distances). We choose complete linkage here as it has a tendency to create more compact, clique-like clusters (Jain et al. 1999). We note that complete linkage is also known to be more sensitive to outliers, but given our normalized distance measure, we do not face this issue here. For our HAC implementation, we use the *hclust* function from the *stats* package of R (R Development Core Team 2007).

4 Analysis and discussion

Upon running Algorithm 1 with D containing all 110 available datasets and I containing all 61 interestingness measures, it became apparent that the computation of the Indice Probabiliste d'Ecart d'Equilibre (IPE) is unfeasible for several of our large datasets. Its formula calls for the sum of large factorials, which even after simplification cannot be computed (the numbers involved exceed the representation capacity of standard data types). We did try to use the BigInteger class available in C# but the computational time remained impractical. In order to get a sense for how IPE behaves relative to the other interestingness measures, we ran Algorithm 1 with all of the interestingness measures, but only the 71 of our selected datasets that have less than 500 examples. We then performed the clustering and found that the rankings produced by IPE are most like those produced by Confidence and Confirm Descriptive (rank correlation greater than 0.975). In order to be able to use all of the available datasets in the rest of our analyses, we will omit IPE from the experiments, and simply recall that it should be grouped with Confidence and Confirm Descriptive.

4.1 Indistinguishability

Even after redundant interestingness measures are removed, there still remain several groups of measures that cluster at a distance of 0, namely:

- {Yule's Y, Odds Ratio, Yule's Q}
- {Kulczynski 1, Jaccard, F-measure}
- {Lift, Information Gain}

- {Conviction, Loevinger}
- {Odd Multiplier, Zhang}
- {Ganascia, Example and Counterexample Rate, Confidence, Sebag–Schoenauer}

Recall that our analysis focuses on the behavior of interestingness measures as it pertains to how they rank association rules. Hence, two interestingness measures, M_1 and M_2, exhibit the same behavior if, given a set of rules $R = \{r_1, r_2, \ldots, r_n\}$, they produce the same ranking over R. In other words, for any two rules, $r_i, r_j \in R$, if M_1 ranks r_i before r_j, then so does M_2, and vice-versa. While this clearly holds when the formulas for M_1 and M_2 are the same, as is the case in the redundancies highlighted in our list of interestingness measures, it may also hold in the more general cases highlighted in the dendrogram.

We show that these observed clusters at 0 are not the result of an artifact of the data used for clustering, but the reflection of true equivalences among the interestingness measures involved. To show that M_1 and M_2 are equivalent, or indistinguishable in terms of their ranking behavior, we must prove that

$$\forall r_i, r_j \in R \quad M_1(r_i) \leq M_1(r_j) \iff M_2(r_i) \leq M_2(r_j)$$

where $M(r)$ denotes the rank that interesting measure M gives to rule r. In what follows, we prove corresponding instantiations of this statement for our discovered clusters at 0. Note that Fürnkranz and Flach (2005) use a similar approach in their analysis of heuristics for covering algorithms.

Theorem 1 *Lift and Information Gain are indistinguishable.*

Proof It is easy to see that Information Gain $= \log_2$ Lift, and the result follows immediately since, as also pointed out by Lenca et al. (2007), Information Gain is a monotonically increasing transformation of Lift. □

Lemma 1 *Yule's Q and Odds Ratio are indistinguishable.*

Proof We first note, as also pointed out by Tan et al. (2002) and Ohsaki et al. (2004), that Yule's Q $= \frac{\text{Odds Ratio} - 1}{\text{Odds Ratio} + 1}$. For simplicity, let $x_1 = \text{Odds Ratio}(r_i)$ and $x_2 = \text{Odds Ratio}(r_j)$ for any two rules r_i and r_j. To show that Yule's Q and Odds Ratio are indistinguishable, we need only show that:

$$x_1 \leq x_2 \iff \frac{x_1 - 1}{x_1 + 1} \leq \frac{x_2 - 1}{x_2 + 1}$$

We have

$$\begin{aligned}
\frac{x_1 - 1}{x_1 + 1} \leq \frac{x_2 - 1}{x_2 + 1} &\iff (x_1 - 1)(x_2 + 1) \leq (x_1 + 1)(x_2 - 1) \\
&\iff x_1 x_2 + x_1 - x_2 - 1 \leq x_1 x_2 - x_1 + x_2 - 1 \\
&\iff 2x_1 \leq 2x_2 \\
&\iff x_1 \leq x_2
\end{aligned}$$

which establishes the needed result. □

Lemma 2 *Yule's Y and Odds Ratio are indistinguishable.*

Proof It is easy to see that Yule's Y $= \frac{\sqrt{\text{Odds Ratio}}-1}{\sqrt{\text{Odds Ratio}}+1}$, so that a simple adaptation of Lemma 1 establishes the result. □

It follows immediately from Lemmas 1 and 2 that:

Theorem 2 *Odds Ratio, Yule's Q and Yule's Y are indistinguishable.*

Lemma 3 *Confidence and Ganascia are indistinguishable.*

Proof We first note, as also pointed out by Lallich et al. (2007), that Ganascia $=$ 2 Confidence -1. As before, let $x_1 = \text{Confidence}(r_i)$ and $x_2 = \text{Confidence}(r_j)$ for any two rules r_i and r_j. To show that Ganascia and Confidence are indistinguishable, we need only show that:

$$x_1 \leq x_2 \iff 2x_1 - 1 \leq 2x_2 - 1$$

which is trivial. □

Lemma 4 *Sebag–Schoenauer and Example and Counterexample Rate are indistinguishable.*

Proof We first note that Example and Counterexample Rate (ECR) is not defined when $P(AB) = 0$. Let us begin by assuming that $P(AB) \neq 0$. Then, as also pointed out by Lallich et al. (2007), ECR $= 1 - \frac{1}{\text{Sebag–Schoenauer}}$. Again, let $x_1 =$ Sebag–Schoenauer(r_i) and $x_2 =$ Sebag–Schoenauer(r_j) for any two rules r_i and r_j. To show that Sebag–Schoenauer and ECR are indistinguishable, we need only show that:

$$x_1 \leq x_2 \iff 1 - \frac{1}{x_1} \leq 1 - \frac{1}{x_2}$$

which follows from

$$\begin{aligned} 1 - \frac{1}{x_1} \leq 1 - \frac{1}{x_2} &\iff -\frac{1}{x_1} \leq -\frac{1}{x_2} \\ &\iff \frac{1}{x_1} \geq \frac{1}{x_2} \\ &\iff x_1 \leq x_2 \end{aligned}$$

Now, assume that $P(AB) = 0$. Then, Sebag–Schoenauer $= 0$. Since 0 is the smallest value that Sebag–Schoenauer can take, rules it assigns that value to are ranked last. While ECR is not defined when $P(AB) = 0$, it makes sense, from a computational standpoint, to assign it its limit as $P(AB)$ approaches 0, namely $-\infty$. Since this is also the smallest value that ECR can take, rules it assigns that value to are ranked last. Hence, the two measures remain indistinguishable. □

Lemma 5 *Confidence and Example and Counterexample Rate are indistinguishable.*

Proof Note, as in Lemma 4 that ECR is not defined when $P(AB) = 0$. Let us again begin by assuming that $P(AB) \neq 0$. Then, simple algebraic transformations show that ECR $= 2 - \frac{1}{\text{Confidence}}$, and a simple adaptation of Lemma 4 establishes the needed result. Now, assume that $P(AB) = 0$. Then, Confidence $= 0$ (provided $P(A) \neq 0$, which we can safely assume). Since 0 is the smallest value that Confidence can take, rules it assigns that value to are ranked last. As in Lemma 4, we assign ECR the value $-\infty$ in this situation. And it follows that the two measures remain indistinguishable. □

It follows immediately from Lemmas 3, 4 and 5 that:

Theorem 3 *Confidence, Ganascia, Sebag–Schoenauer, and Example and Counterexample Rate are indistinguishable.*

Note that although in theory Laplace Correction and Confidence are not strictly indistinguishable, in practice they may be considered that way. Indeed, we have LAC $= \frac{NP(AB)+1}{NP(A)+2} = \frac{P(AB)+\frac{1}{N}}{P(A)+\frac{2}{N}}$, which, for large N, approaches $\frac{P(AB)}{P(A)} = P(B|A)$, and in the limit, Laplace Correction is equal to Confidence. Given the sizes of the datasets we have selected, and of those typical in applications of association rule mining, there are many instances where the values produced by Laplace Correction and Confidence are close enough that the ensuing rankings are almost the same, leading to very small distances between them. In our experiment, the average value of the distance between Laplace Correction and Confidence is about 0.01, or a rank correlation of 0.999. Hence, for all practical purposes, the rankings produced by Laplace Correction and Confidence are indistinguishable. We shall therefore add Laplace Correction to the above group of measures.

Theorem 4 *Loevinger and Conviction are indistinguishable.*

Proof We first note that Loevinger is not defined when $P(\overline{B}) = 0$. Again, we safely assume that $P(A) \neq 0$, since otherwise no rule involving A could have been generated. Let us begin by assuming that $P(\overline{B}) \neq 0$. Then, as also pointed out by Berzal et al. (2002) and Lallich et al. (2007), we have Loevinger $= 1 - \frac{1}{\text{Conviction}}$, and the needed result follows from the proof of Lemma 4. Now, assume that $P(\overline{B}) = 0$. Then, Conviction $= 0$ (provided $P(A) \neq 0$, which we are assuming). Since 0 is the smallest value that Conviction can take, rules it assigns that value to are ranked last. Similarly to what we do with Example and Counterexample Rate, we assign to Loevinger the value $-\infty$ in this situation. And it follows that the two measures remain indistinguishable. □

Lemma 6 *F-measure and Jaccard are indistinguishable.*

Proof We first note that

$$\text{F-measure} = \frac{2P(A|B)P(B|A)}{P(A|B)+P(B|A)} = 2\frac{\frac{P(AB)}{P(B)}\frac{P(AB)}{P(A)}}{\frac{P(AB)}{P(B)}+\frac{P(AB)}{P(A)}} = 2\frac{\frac{P(AB)}{P(A)P(B)}}{\frac{1}{P(A)}+\frac{1}{P(B)}}$$
$$= 2\frac{P(AB)}{P(A)P(B)}\frac{P(A)P(B)}{P(A)+P(B)} = 2\frac{P(AB)}{P(A)+P(B)}$$

and

$$1 + \text{Jaccard} = 1 + \frac{P(AB)}{P(A) + P(B) - P(AB)} = \frac{P(A) + P(B)}{P(A) + P(B) - P(AB)}$$

Since Jaccard $\neq -1$, we can take the inverse to get

$$\frac{1}{1 + \text{Jaccard}} = 1 - \frac{P(AB)}{P(A) + P(B)} = 1 - \frac{1}{2}\,\text{F-measure}$$

and it follows that:

$$\text{F-measure} = \frac{2\,\text{Jaccard}}{1 + \text{Jaccard}}$$

As before, let $x_1 = \text{Jaccard}(r_i)$ and $x_2 = \text{Jaccard}(r_j)$ for any two rules r_i and r_j. To show that F-measure and Jaccard are indistinguishable, we need only show that:

$$x_1 \leq x_2 \iff \frac{x_1}{1 + x_1} \leq \frac{x_2}{1 + x_2}$$

We have (note that x_1 and x_2 are positive)

$$\begin{aligned} \frac{x_1}{1 + x_1} \leq \frac{x_2}{1 + x_2} &\iff (1 + x_2)x_1 \leq (1 + x_1)x_2 \\ &\iff x_1 + x_1x_2 \leq x_2 + x_1x_2 \\ &\iff x_1 \leq x_2 \end{aligned}$$

which establishes the needed result. □

Lemma 7 *Jaccard and Kulczynski 1 are indistinguishable.*

Proof We first note, as also pointed out by Plasse et al. (2007), that $\frac{1}{\text{Jaccard}} = \frac{1}{\text{Kulczynski 1}} + 1$, so that

$$\text{Jaccard} = \frac{\text{Kulczynski 1}}{1 + \text{Kulczynski 1}}$$

and the proof of indistinguishability is the same as in Lemma 6. □

It follows immediately from Lemmas 6 and 7 that:

Theorem 5 *F-measure, Jaccard and Kulczynski 1 are indistinguishable.*

Theorem 6 *Zhang and Odd Multiplier are indistinguishable.*

Proof We first note that

$$\begin{aligned}
\text{Zhang} &= \frac{P(AB) - P(A)P(B)}{\max(P(AB)P(\overline{B}),\, P(A\overline{B})P(B))} \\
&= \frac{-P(B)[P(A) - P(AB)] + P(AB) - P(B)P(AB)}{\max(P(AB)P(\overline{B}),\, P(A\overline{B})P(B))} \\
&= \frac{-P(B)P(A\overline{B}) + P(AB)(1 - P(B))}{\max(P(AB)P(\overline{B}),\, P(A\overline{B})P(B))} \\
&= \frac{P(AB)P(\overline{B}) - P(A\overline{B})P(B)}{\max(P(AB)P(\overline{B}),\, P(A\overline{B})P(B))} \\
&= \frac{\frac{P(AB)P(\overline{B})}{P(A\overline{B})P(B)} - 1}{\max\left(\frac{P(AB)P(\overline{B})}{P(A\overline{B})P(B)}, 1\right)} = \frac{\text{Odd Multiplier} - 1}{\max(1, \text{Odd Multiplier})}
\end{aligned}$$

As before, let $x_1 = \text{Odd Multiplier}(r_i)$ and $x_2 = \text{Odd Multiplier}(r_j)$ for any two rules r_i and r_j. To show that Zhang and Odd Multiplier are indistinguishable, we need only show that:

$$x_1 \leq x_2 \iff \frac{x_1 - 1}{\max(1, x_1)} \leq \frac{x_2 - 1}{\max(1, x_2)}$$

We consider four possible cases (note that x_1 and x_2 are positive).

Case 1 $\max(1, x_1) = \max(1, x_2) = 1$. Then,

$$\begin{aligned}
\frac{x_1 - 1}{\max(1, x_1)} \leq \frac{x_2 - 1}{\max(1, x_2)} &\iff (x_1 - 1) \leq (x_2 - 1) \\
&\iff x_1 \leq x_2
\end{aligned}$$

Case 2 $\max(1, x_1) = x_1$ and $\max(1, x_2) = x_2$. Then,

$$\begin{aligned}
\frac{x_1 - 1}{\max(1, x_1)} \leq \frac{x_2 - 1}{\max(1, x_2)} &\iff \frac{x_1 - 1}{x_1} \leq \frac{x_2 - 1}{x_2} \\
&\iff 1 - \frac{1}{x_1} \leq 1 - \frac{1}{x_2} \\
&\iff x_1 \leq x_2 \ \text{ from Lemma 4}
\end{aligned}$$

Case 3 $\max(1, x_1) = 1$ and $\max(1, x_2) = x_2$. Then, we have $x_1 - 1 < 0$, while $\frac{x_2 - 1}{x_2} > 0$. It follows that $x_1 \leq x_2$ (since $x_1 < 1$ and $x_2 > 1$), and the result holds trivially.

Case 4 $\max(1, x_1) = x_1$ and $\max(1, x_2) = 1$. Then, we have $\frac{x_1 - 1}{x_1} > 0$, while $x_2 - 1 < 0$, so that $\frac{x_1 - 1}{\max(1, x_1)} \geq \frac{x_2 - 1}{\max(1, x_2)}$. It also follows that $x_1 \geq x_2$ (since $x_1 > 1$ and $x_2 < 1$), and the result holds vacuously.

As these four cases are the only possible ones, the result follows. □

There are clearly many other relationships among interestingness measures, such as $\chi^2 = N \times \text{Correlation Coefficient}^2$, Support $=$ Prevalence $\times$ Recall, and Piatetsky-Shapiro $= N \times$ Coverage $\times$ Added Value, but these clearly do not lead to any similarity in rankings.

From a practical standpoint, the equivalences captured by Theorems 1–6 are significant, since interestingness measures are used as a means to an end by practitioners, rather than as an end in themselves. In other words, what really matters is how the interestingness measures rank the rules generated by the mining algorithm so the user can get the most interesting rules out of his/her data. As a result, if two measures produce the same ranking, they are fundamentally the same since there is no way to distinguish among the rules produced. As we discuss further here, we feel that there is indeed tremendous value in focusing on ranking behavior.

Because there are so many existing interestingness measures, it is unlikely that we, or others, would ever have analyzed any of these formulas well enough—there are $\frac{61 \times 60}{2} = 1,830$ pairs of measures—to discover these relationships without a clustering of their rankings. Beyond even the sheer number of possibilities is the fact that some would likely be missed a priori due to a lack of similarity in their closed forms. Typical examples here include Indice Probabiliste d'Ecart d'Equilibre and Confidence, and Kappa and 2-way Support. While their formulas show little obvious similarity, they actually cluster together at a very small distance suggesting that their rankings are actually rather similar. In general, our behavior-based analysis may highlight similarities that would otherwise easily be overlooked. Conversely, behavior-based clustering can be used to confirm theoretical results and may help catch possible errors. For example, Berzal et al. (2002) offer a proof that Certainty Factor (or Loevinger, here) is equal to 1—Lift, and thus would produce the same rankings. Yet, they do not, as clearly shown in our dendrogram. It turns out, upon closer examination, that the proposed proof does contain an error.

Some researchers have also worked on clustering interestingness measures, as we do, but they have used Pearson correlation (i.e., correlation among raw values of interestingness) rather than Spearman rank correlation (Huynh et al. 2005; Abe and Tsumoto 2008). Unfortunately, Pearson correlation may exhibit non-zero values when Spearman rank correlation does not, thus creating a false impression of differences. For example, while the foregoing makes it clear that Sebag–Schoenauer, Example and Counterexample Rate, and Confidence, as well as Yule's Q, Yule's Y and Odds Ratio, always produce the same rankings, the dendrogram obtained by Huynh et al. (2005) would have them in separate, somewhat distant clusters. Our proposed clustering, which focuses on relative rankings rather than actual scores, allows us to capture these similarities.

On the other hand, several researchers have studied interestingness measures with a view of selecting among them, from a rather theoretical standpoint, relying on properties such as linearity and intelligibility (Vaillant et al. 2004; Lenca et al. 2007, 2008), or various symmetries and invariances (Tan et al. 2002, 2004), with little concern for behavior. The work of Jalali-Heravi and Zaïane (2010) is particularly noteworthy in that respect as it considers a mixture of 16 of these properties, and is directly comparable to our study. The authors represent interestingness measures as binary vectors of properties and, using Hamming distance, cluster them with average linkage HAC.

While some of their groupings agree with ours (e.g., Yule's Q and Yule's Y, Conviction and Loevinger, Confidence and Ganascia), there are significant differences. For example, while property-based clustering captures the similarity of Yule's Q and Yule's Y, it fails to recognize their behavioral similarity to Odds Ratio; conversely, while behavior-based clustering highlights differences between Confirm Causal and Leverage, property-based clustering has them cluster at 0 suggesting that the measures are indistinguishable. In the context of selection, these differences show that handcrafted properties do not correlate well with ranking behavior, which can have a significant impact on outcomes for the end user.

Furthermore, one could make the argument that expert-crafted properties are of value mostly once one knows more about the behavior of interestingness measures. Indeed, if two measures produce radically different rankings, it may be harder to justify a choice between them based solely on extrinsic properties. Just because an interestingness measure is intelligible, for example, may have little bearing on the quality of the rankings it produces in practice. On the other hand, if two metrics behave rather similarly, one may now select between them based on such other criteria. For example, when considering Jaccard and Agreement–Disagreement Index (i.e., Kulczyniski 1), Plasse et al. (2007) state, after pointing out that the two would produce the same ranking of rules (see Lemma 7 for a proof), that "while they lead to the same classification of rules, the Jaccard coefficient has the advantage of varying between 0 and 1," a property they obviously find desirable.

From our behavioral perspective, indistinguishable interestingness measures are redundant in the sense that one in each group clustering at 0 is sufficient for purposes of ranking. Consequently, we can collapse each group into a single representative measure, as follows. We provide a short justification for our choice of representative in each group.

- {Yule's Y, **Odds Ratio**, Yule's Q}. Odds Ratio is computationally simpler.
- {Kulczynski 1, Jaccard, **F-measure**}. No significant computational difference, but F-measure is probably more popular.
- {**Lift**, Information Gain}. Lift is computationally simpler and very popular in association rule mining.
- {Conviction, **Loevinger**}. No significant computational difference, but Loevinger predates Conviction.
- {**Odd Multiplier**, Zhang}. Odd Multiplier is computationally simpler.
- {Laplace Correction, Ganascia, Example and Counterexample Rate, **Confidence**, Sebag–Schoenauer}. Confidence is computationally simpler, and the interestingness measure *par excellence* in association rule mining.

Hence, we can safely remove 11 interestingness measures from our original set, leading to a total of 50 behaviorally distinct interestingness measures only.

4.2 Empirical ranking behavior analysis

Running Algorithm 1 on the reduced set of interestingness measures, but considering all available datasets, produces a distance matrix which, after clustering, leads to a

dendrogram as illustrated in Fig. 1. Recall that, although not shown, IPE would cluster with Confidence and Confirm Descriptive.

A first look at Fig. 1 reveals a significant amount of structure to the dendrogram, with a number of distinct small groups of positively correlated measures progressively merging into larger groups, and a few relatively independent measures (e.g., Support, TIC, Implication Index). At a high-level, there seems to be two sizable groups of generally positively correlated measures (i.e., distance less than 0.5, or correlation greater than 0.5), one from EII1 to Odd Multiplier (9 measures), and the other from Odds Ratio to Normalized Mutual Information (26 measures), as well as a distinct group of 7 measures (from Recall to Implication Index) that tend to be negatively correlated with these. Focusing on the small distances, we notice that there remain several groups of interestingness measures that behave very similarly. It is possible to argue that, for practical purposes, very small distances among measures are of little consequence on the respective final rankings of rules (recall the extreme case of Laplace Correction and Confidence above). Consequently, we could also group such measures together, and collapse them into a single representative measure. While the details of the dendrogram may vary over repeated experiments, the aforementioned structure remains consistent across them.

Given the fully nested sequence of clusterings obtained by HAC, the choice of a specific final grouping is typically made by selecting a level at which to cut through the dendrogram, and defining the clusters as the groups of elements hanging from the subtrees whose top branches intersect with the horizontal line corresponding to the chosen level. The higher the cut point, the less similar the elements of the corresponding clusters. The lower the cut point, the less the level of generalization. Finding a best cut point is difficult and requires some measure of clustering goodness or quality. Here, however, such a formal quality measure may not be strictly necessary as the metric we use for clustering embeds the very notion of quality we seek, namely behavior similarity as measured by rank correlation. Therefore, an appropriate threshold on that value is sufficient. Here, we select a rather conservative value of 0.9, or a cut point of $\sqrt{0.05}$ (recall that we use $\sqrt{\frac{1-r}{2}}$ as our distance metric). Examining the results over 10 random experiments gives rise to the following 21 clusters.

1. {Support}
2. {Prevalence}
3. {K-measure}
4. {Least Contradiction}
5. {Confidence, Confirm Descriptive, Indice Probabiliste d'Ecart d'Equilibre}
6. {TIC}
7. {EII1, EII2}
8. {Complement Class Support, Leverage, Confidence Causal, Confirmed Confidence Causal}
9. {Directed Information Ratio}
10. {Loevinger, Odd Multiplier}
11. {Odds Ratio}
12. {Dilated Chi-square}

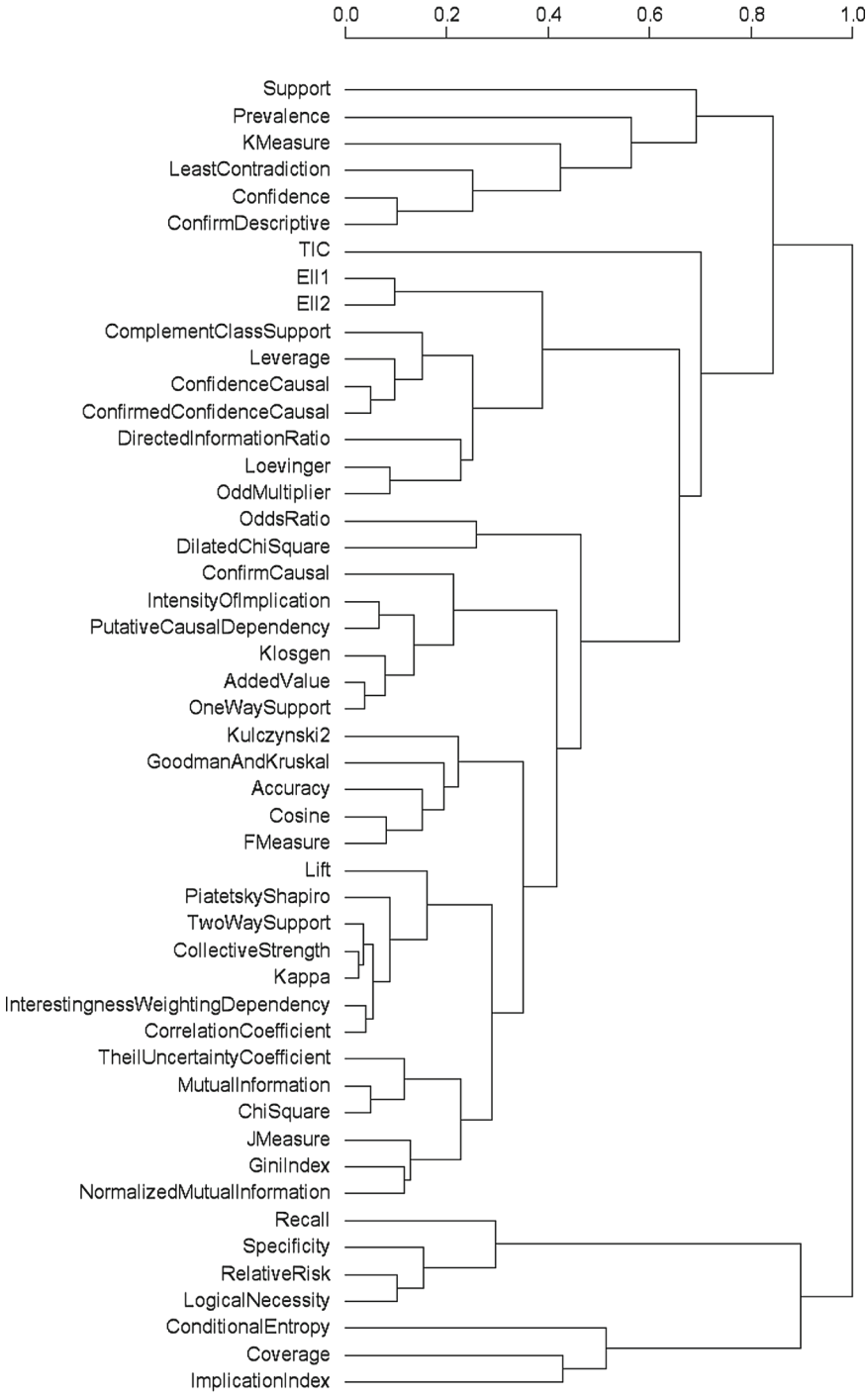

Fig. 1 Behavior-based clustering of interestingness measures

13. {Confirm Causal, Intensity of Implication, Putative Causal Dependency, Klosgen, Added Value, 1-way Support}
14. {Kulczynski 2, Goodman–Kruskal, Accuracy, Cosine, F-measure}
15. {Lift, Piatetsky-Shapiro, 2-way Support, Collective Strength, Kappa, Interestingness Weighting Dependency, Correlation Coefficient}
16. {Theil Uncertainty Coefficient, Mutual Information, Chi-square, J-measure, Gini Index, Normalized Mutual Information}
17. {Recall}
18. {Specificity, Relative Risk, Logical Necessity}
19. {Conditional Entropy}
20. {Coverage}
21. {Implication Index}

Where there is an overlap in the interestingness measures under consideration, these clusters are generally consistent with those obtained by Huynh et al. (2005) with a threshold of 0.85. Where they differ, we argue that our results are more reliable and more robust since we have verified the interestingness measures' formulas and experimented with a large and varied set of applications. Note that while it is not always straightforward, it is possible to confirm some of the groupings analytically. For example, 2-way Support $= P(AB)\log_2 \frac{P(AB)}{P(A)P(B)}$. Since in many cases $\frac{P(AB)}{P(A)P(B)}$ (= Lift) $\simeq 1$, we have 2-way Support $\simeq P(AB)(\frac{P(AB)}{P(A)P(B)} - 1)$ by Taylor expansion, and thus 2-way Support $\simeq$ Interestingness Weighting Dependency, leading to similar rankings for these two measures.

To gain further insight into the behavior of our selected interestingness measures, we build three complementary graphs over the set of measures, as follows. We use the distance matrix that led to the clustering of Fig. 1.

- In P^+, the graph of strongly correlated interestingness measures, there is an edge between any pair of measures when their distance is less than $\sqrt{0.025}$ (i.e., rank correlation above 0.95).
- In P^-, the graph of strongly anti-correlated interestingness measures, there is an edge between any pair of measures when their distance is greater than $\sqrt{0.975}$ (i.e., rank correlation below -0.95).
- In P^0, the graph of uncorrelated interestingness measures, there is an edge between any pair of measures when their distance is between $\sqrt{0.4875}$ and $\sqrt{0.5125}$ (i.e., rank correlation between -0.025 and $+0.025$).

P^+ and P^0 correspond roughly to Huynh et al. (2007)'s CG+ and CG0 graphs, except that, in addition to using raw correlation rather than rank correlation, the latter are computed by averaging only over applications where the correlation meets the threshold condition whereas we average over all applications. Hence, P^+ and P^0 are better representatives of overall performance. Since Huynh et al. (2007) use the absolute value of the correlation, they have no equivalent for P^-. We feel that there is value in distinguishing correlation and anti-correlation. The graphs appear in Figs. 2, 3, and 4. To avoid unnecessary clutter, isolates have been left out.

Figure 2 comes as no surprise as it mostly confirms the findings of the above clustering. Since we used complete linkage, the graph does add some information

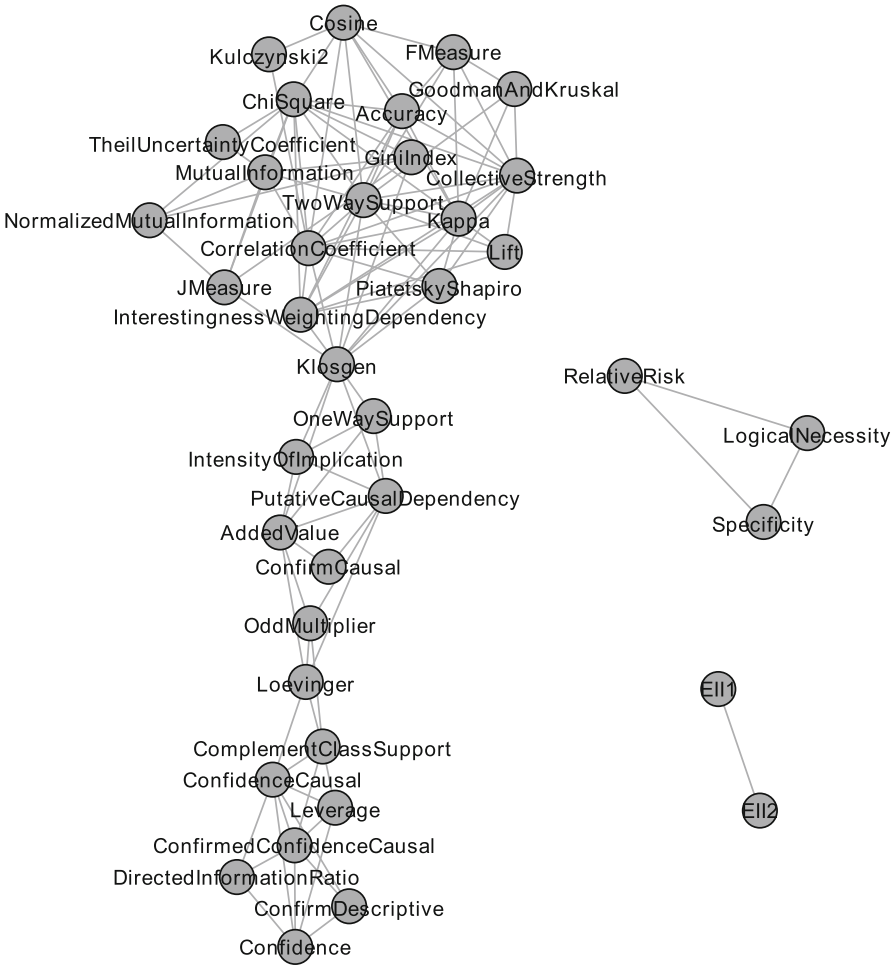

Fig. 2 Strongly positively correlated interestingness measures (P^+)

about pairwise relationships. For example, although Cosine and Chi-square, or Added Value and Odd Multiplier, do not cluster together, their rank correlation is above 0.95. However, one must also be careful not to infer any transitive closure from the graph. Only fully connected subgraphs (e.g., {Specificity, Relative Risk, Logical Necessity}) correspond to clusters in the above sense. Of course, since we use a distance metric, it is true that if $d(x, y) \leq \epsilon$ and $d(x, z) \leq \epsilon$ then $d(y, z) \leq 2\epsilon$, so the transitive closure of the graph built based on distance ϵ corresponds to clusters at distance 2ϵ (i.e., a lower degree of rank correlation).

From Fig. 2, we also observe three subgroups of measures in the large connected component. The top group is rather tightly interconnected with several nodes holding rather central positions with the highest degrees, e.g., Two Way Support (degree 13), Collective Strength, Correlation Coefficient, Kappa and Chi-square (degree 12).

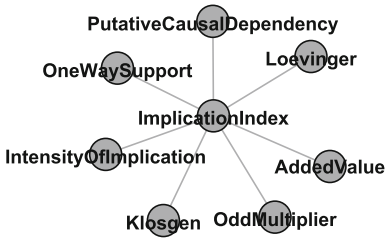

Fig. 3 Strongly anti-correlated interestingness measures (P^-)

In contrast, other nodes act as bridges, or cut-points, between groups of strongly positively correlated interestingness measures. For example, Klosgen acts a bridge between the otherwise disconnected top (Clusters 14, 15 and 16) and middle (Cluster 13) groups, while Odd Multiplier and Loevinger together (Cluster 10) bridge between the otherwise disconnected middle and bottom (Clusters 5, 8 and 9) groups.

The value of P^- is in highlighting interesting measures whose rankings are close to opposite of each other. There are two situations in which this may arise. One is when the two measures are indeed trying to capture diametrically opposed notions of interestingness. For example, one may decide that the most interesting rules are those corresponding to most frequent events (i.e., high support), while another may decide that the most interesting rules are actually those corresponding to rare events (i.e., low support). The other cause of anti-correlation is error in formula or reverse rankings (i.e., where smaller values correspond to more interesting rules). The graph of Fig. 3 is remarkably sparse, suggesting that very few measures in our set are strongly anti-correlated. It shows that Implication Index is most strongly anti-correlated with Clusters 10 and 13 above, as there is an edge between Implication Index and every member of these clusters (except for Confirm Causal). The average rank correlation (after Fisher's z back-transformation) between Implication Index and the interestingness measures of Clusters 10 and 13 is -0.97. In fact, Implication Index is negatively correlated with all other measures, except Coverage and Conditional Entropy. There is a possibility that this may be an example of ranking reversal. Indeed, while we checked the original definition of Implication Index and several authors use the same formula as we do, at least one set of authors uses the opposite (i.e., $-$ IIN) in their work (Vaillant et al. 2004; Lenca et al. 2007, 2008). If they are correct, Implication Index would be grouped with Cluster 13. Either way, one may argue that the difference is one of semantics and little is gained by using Implication Index in the presence of the measures of cluster 13. If Implication Index is as we define it here, one could get a close approximation of its behavior on a set of rules by reversing the ranking obtained by any of the measures in cluster 13. Hence, we could either omit Implication Index or add it to Cluster 13, reducing the total number of clusters to 20.

As far as other measures capturing somewhat opposite notions of interestingness, Fig. 3 suggests that we have none at the level of anti-correlation selected, and the foregoing argument would cause us to group them with their counterparts and simply tag them as reversed. On the other hand, if we lower the threshold used to build P^- to -0.8, the pairs {Specificity, Prevalence} and {Recall, Prevalence} appear, which are examples of measures focusing on rather different views of interestingness. In general,

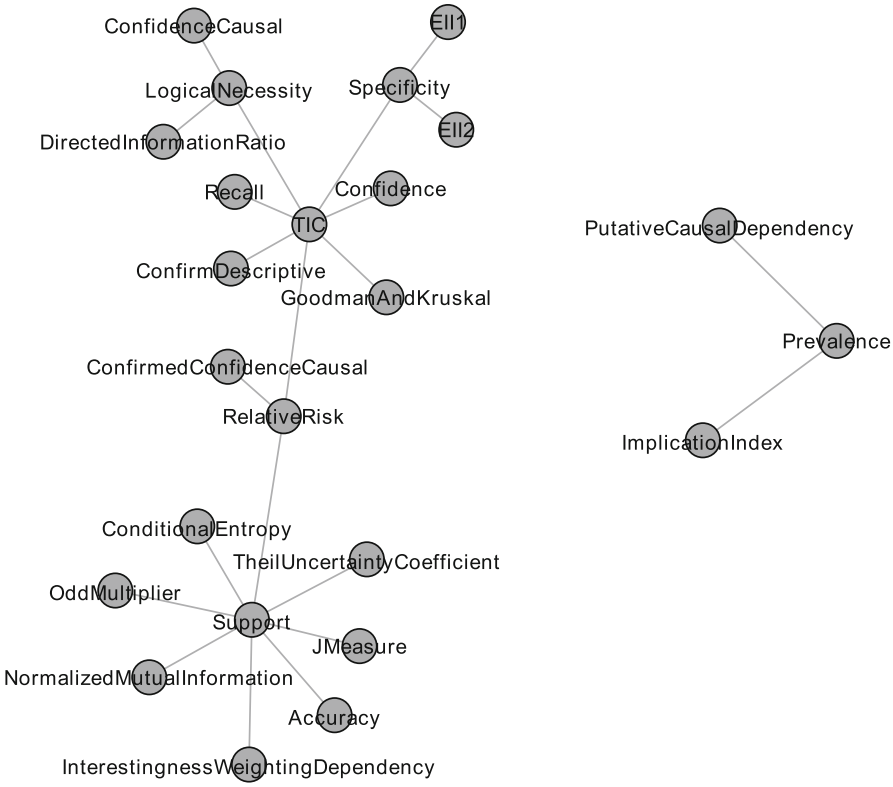

Fig. 4 Most uncorrelated interestingness measures (P^0)

the sparsity of P^-, together with the large cluster of positively correlated measures mentioned in connection with the dendrogram of Fig. 1, show that there is a significant level of consistency in what is regarded as interesting by the designers of our selected measures.

Figure 4 shows two components, and the larger component further highlights a clear separation between two subgroups of measures, bridged across by a single link each to Relative Risk. The graph makes clear that Support and TIC, which are central to each group, are the two interestingness measures least correlated with the others, with degrees 8 and 7, respectively. Closer examination of our experimental results reveals that this observation about TIC may be somewhat of an aberration. Recall that TIC is defined as $\sqrt{\mathrm{DIR}(A \Rightarrow B)\,\mathrm{DIR}(\overline{B} \Rightarrow A)}$. It turns out that for most of our datasets and most of the rules induced by Apriori, at least one of $\mathrm{DIR}(A \Rightarrow B)$ or $\mathrm{DIR}(\overline{B} \Rightarrow A)$ is equal to 0 (see the definition of DIR), so that TIC is also often 0. It follows that in most cases the rankings produced by TIC are essentially random, since ties among rules with the same interestingness value are broken arbitrarily based on rule numbers. Hence, averaged across all datasets, the rank correlation between TIC and most other measures is close to 0. This observation calls into question the value and appropriateness of TIC as an interestingness measure for association rule mining.

Support, on the other hand, is a well-accepted measure almost universally used in implementations of association rule mining, at least as a mechanism to control the number of frequent itemsets, and consequently rules, generated by the algorithm. Interestingly, Support is not strongly correlated with any of the other measures, but it is uncorrelated with eight of them. One is left to wonder whether using Support in the first phase of the algorithm (i.e, frequent itemset generation), followed by an uncorrelated interestingness measure in the second phase (i.e., rule generation and ranking) is not liable to producing unwanted results. For example, if Support is of little concern, the two-stage filtering may prevent the later creation of highly rated rules according to the target interestingness measure (e.g., rare, yet interesting, rules). Incidentally, approaches specifically aimed at capturing rules corresponding to rare events have been designed, including mining with different minimum support thresholds for different items (Liu et al. 1999; Kiran et al. 2009), and clustering transactions prior to finding associations among items belonging to the same clusters only (Koh and Pears 2008). On the other hand, if one assumes implicitly that interesting rules are those that satisfy both filters (i.e., Support and the desired measure), then it makes sense to use one of the uncorrelated measures since they add value.

Overall, our analysis suggests that, while there is significant behavior similarity among some measures (see the 21 clusters obtained from an original set of 50 interestingness measures), there are also significant differences with each measure focusing on unique characteristics of rules, making it difficult to narrow down which specific interestingness measures would be best for a certain scenario. We argue, as others have, that domain knowledge is essential to the selection of an appropriate interestingness measure for a particular task and business objective. Now, while the choice of an interestingness measure is ultimately a subjective decision, our results show that (1) it is possible to narrow the set of possibilities to a smaller subset of clusters of similar measures, and (2) selections should be made from clusters whose elements are generally uncorrelated.

5 Conclusion and future work

This paper presents an empirical, behavior-based clustering of 61 interestingness measures for association rule mining, evaluated against 110 datasets. This extensive analysis makes the following contributions.

- We have carefully defined each interestingness measure in terms of its formula and name. By going back to the source, we have resolved some of the discrepancies found in the existing literature.
- We show the valuable and exploitable interplay between theory and empirical evidence. In particular, we show how experimental results make it easy to spot relatively obvious redundant measures, and offer a way to discover novel equivalences (which may easily be overlooked given the sheer number of possibilities), that may then be confirmed by analytical means. Similarly, existing theoretical results may be verified empirically, or analytical errors uncovered. In other words, it is possible on the one hand to check theoretical finding against empirical results, and on the other, to focus theoretical work based on empirical results (e.g., 0-clusters).

– A focus on ranking behavior makes it possible to reduce the number of interestingness measures from 61 to 21. This has implication for the design of selection systems, since not only is the number of choices more limited, it is also possible to use less computationally expensive measures in place of those which require more computation.
– We highlight strong correlation, anti-correlation and independence among interesting measures, which may be leveraged when considering what measure to use in what context.

Although the amount of choices can be reduced to some degree, this reduction is still limited because most measures generally focus on different and unique characteristics of rules. It has, in fact, been argued that interestingness is ultimately subjective (Sahar 1999) and that interestingness measures "are mainly used to remove meaningless rules rather than to discover really interesting ones for a human user, since they do not include domain knowledge" (Ohsaki et al. 2004). Our results certainly confirm that domain knowledge is essential to the selection of an appropriate interestingness measure for a particular task and business objective.

Finally, we recall that our work has focused on interestingness measures in the context of association rule mining. Given the recent interest in pattern sets, it could be valuable to apply our methodology and/or extend our study to include interestingness measures for pattern sets as this has not been systematically done yet.

References

Abe H, Tsumoto S (2008) Analyzing behavior of objective rule evaluation indices based on a correlation coefficient. In: Proceedings of the 12th international conference on knowledge-based intelligent information and engineering systems (LNAI 5178), pp 758–765

Aggarwal C, Yu P (1998) A new framework for itemset generation. In: Proceedings of the 7th ACM symposium on principles of database systems, pp 18–24

Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of the 20th international conference on very large data bases, pp 487–499

Agrawal R, Imieliński T, Swami A (1993) Mining association rules between sets of items in large databases. ACM SIGMOD Rec 22(2):207–216

Ali K, Manganaris S, Srikant R (1997) Partial classification using association rules. In: Proceedings of the 3rd ACM SIGKDD international conference on knowledge discovery and data mining, pp 115–118

Arunasalam B, Chawla S (2006) CCCS: a top-down associative classifier for imbalanced class distribution. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, pp 517–522

Asuncion A, Newman D (2007) UCI machine learning repository. School of Information and Computer Sciences, University of California, Irvine. http://www.ics.uci.edu/mlearn/mlrepository.html

Azé J, Kodratoff Y (2002) Evaluation de la résistance au bruit de quelques mesures d'extraction de règles d'association. In: Actes des 2èmes Journées Extraction et Gestion des Connaissances, pp 143–154

Bertrand P, Bel Mufti G (2006) Loevinger's measures of rule quality for assessing cluster stability. Comput Stat Data Anal 50(4):992–1015

Berzal F, Blanco I, Sánchez D, Vila MA (2002) Measuring the accuracy and interest of association rules: a new framework. Intell Data Anal 6(3):221–235

Blachman N (1968) The amount of information that y gives about x. IEEE Trans Inf Theory 14(1):27–31

Blanchard J, Kuntz P, Guillet F, Gras R (2003) Implication intensity: from the basic statistical definition to the entropic version. In: Bozdogan H (ed) Statistical data mining and knowledge discovery. Chapman & Hall/CRC Press, Boca Raton, pp 475–493

Blanchard J, Guillet F, Gras R, Briand H (2004) Mesurer la qualité des règles et de leurs contraposées avec le taux informationnel tic. In: Actes des 4èmes Journées Extraction et Gestion des Connaissances, pp 287–298

Blanchard J, Guillet F, Briand H, Gras R (2005a) Assessing rule interestingness with a probabilistic measure of deviation from equilibrium. In: Proceedings of the 11th international symposium on applied stochastic models and data analysis, pp 191–200

Blanchard J, Guillet F, Gras R, Briand H (2005b) Using information-theoretic measures to assess association rule interestingness. In: Proceedings of the 5th IEEE international conference on data mining, pp 66–73

Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Chapman & Hall/CRC Press, Boca Raton

Brin S, Motwani R, Silverstein C (1997a) Beyond market baskets: generalizing association rules to correlations. In: Proceedings of the ACM SIGMOD international conference on management of data, pp 265–276

Brin S, Motwani R, Ullman J, Tsur S (1997b) Dynamic itemset counting and implication rules for market basket data. In: Proceedings of the ACM SIGMOD international conference on management of data, pp 255–264

Clark P, Boswell R (1991) Rule induction with CN2: some recent improvements. In: Proceedings of the 5th European working session on, learning, pp 151–163

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46

Corey D, Dunlap W, Burke M (1998) Averaging correlations: expected values and bias in combined Pearson rs and Fisher's z transformations. J Gen Psychol 125(3):245–261

De Bie T, Kontonasios KN, Spyropoulou E (2010) A framework for mining interesting pattern sets. SIGKDD Explor 12(2):92–100

Duda R, Gaschnig J, Hart P (1981) Model design in the prospector consultant system for mineral exploration. In: Webber B, Nilsson N (eds) Readings in artificial intelligence. Tioga, Palo Alto, pp 334–348

Fieller E, Hartley H, Pearson E (1957) Test for rank correlation coefficients. I. Biometrika 44(3/4):470–481

Fürnkranz J, Flach P (2005) Roc n rule learning—towards a better understanding of covering algorithms. Mach Learn 58(1):39–77

Gallo A, De Bie T, Cristianini N (2007) MINI: mining informative non-redundant itemsets. In: Proceedings of the 11th conference on principles and practice of knowledge discovery in databases, pp 438–445

Ganascia J (1991) CHARADE: Apprentissage de bases de connaissances. In: Kodratoff Y, Diday E (eds) Induction Symbolique-Numérique à Partir de Données. Cépaduès-éditions, Toulouse

Geng L, Hamilton H (2006) Interestingness measures for data mining: a survey. ACM Comput Surv 38(3):1–32

Goodman L, Kruskal W (1954) Measures of association for cross-classifications. J Am Stat Soc 49(268):732–764

Gras R, Larher A (1992) L'implication statistique, une nouvelle méthode d'analyse de données. Mathématiques et Sciences Humaines 120:5–31

Gray B, Orlowska M (1998) CCAIIA: clustering categorical attributes into interesting association rules. In: Proceedings of the 2nd Pacific Asia conference on knowledge discovery and data mining, pp 132–143

Greenacre M, Primicerio R (2013) Multivariate data analysis for ecologists. Foundation BBVA, Madrid

Hahsler M, Hornik K (2007) New probabilistic interest measures for association rules. Intell Data Anal 11(5):437–455

Hill T, Lewicki P (2007) Statistics: methods and applications. StatSoft, Tulsa. http://www.statsoft.com/textbook/

Huynh XH, Guillet F, Briand H (2005) A data analysis approach for evaluating the behavior of interestingness measures. In: Proceedings of the 8th international conference on discovery science (LNAI 3735), pp 330–337

Huynh XH, Guillet F, Briand H (2006) Discovering the stable clusters between interestingness measures. In: Proceedings of the 8th international conference on enterprise information systems: databases and information systems integration, pp 196–201

Huynh XH, Guillet F, Blanchard J, Kuntz P, Briand H, Gras R (2007) A graph-based clustering approach to evaluate interestingness measures: a tool and a comparative study. In: Guillet F, Hamilton H (eds) Quality measures in data mining, vol 43. Studies in computational intelligence, Springer, Heidelberg, pp 25–50

Jaccard P (1901) Etude comparative de la distribution florale dans une portion des alpes et des jura. Bulletin de la Société Vaudoise des Sciences Naturelles 37:547–579

Jain A, Dubes R (1988) Algorithms for clustering data. Prentice-Hall, Inc., Englewood Cliffs
Jain A, Murty M, Flynn P (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323
Jalali-Heravi M, Zaïane O (2010) A study on interestingness measures for associative classifiers. In: Proceedings of the 25th ACM symposium on applied computing, pp 1039–1046
Jaroszewicz S, Simovici D (2004) Interestingness of frequent itemsets using Bayesian networks as background knowledge. In: Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining, pp 178–186
Johnson S (1967) Hierarchical clustering schemes. Psychometrika 2:241–254
Kamber M, Shinghal R (1996) Evaluating the interestingness of characteristic rules. In: Proceedings of the 2nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 263–266
Kannan S, Bhaskaran R (2009) Association rule pruning based on interestingness measures with clustering. Int J Comput Sci Issues 6(1):35–45
Kiran U, Re K et al (2009) An improved multiple minimum support-based approach to mine rare association rules. In: Proceedings of the IEEE symposium on computational intelligence and data mining, pp 340–347
Klösgen W (1996) Explora: a multipattern and multistrategy discovery assistant. In: Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds) Advances in knowledge discovery and data mining. AAAI Press, Cambridge, pp 249–271
Kodratoff Y (2001) Comparing machine learning and knowledge discovery in databases: an application to knowledge discovery in texts. In: Paliouras G, Karkaletsis V, Spyropoulos CD (eds) Machine learning and its applications. Springer, New York, pp 1–21
Koh Y, Pears R (2008) Rare association rule mining via transaction clustering. In: Proceedings of the 7th Australasian conference on knowledge discovery and data mining, pp 87–94
Kulczynski S (1927) Die pflanzenassoziationen der pieninen. Bulletin International de l'Académie Polonaise des Sciences et des Lettres, Classe des Sciences Mathématiques et Naturelles B 2:57–203
Lallich S, Teytaud O, Prudhomme E (2007) Association rule interestingness: measure and statistical validation. In: Guillet F, Hamilton H (eds) Quality measures in data mining, vol 43. Studies in computational intelligence. Springer, Heidelberg, pp 251–275
Lan Y, Chen G, Janssens D, Wets G (2004) Dilated chi-square: a novel interestingness measure to build accurate and compact decision list. In: Proceedings of the international conference on intelligent information processing, pp 233–237
Lan Y, Janssens D, Chen G, Wets G (2006) Improving associative classification by incorporating novel interestingness measures. Expert Syst Appl 31(1):184–192
Lavrač N, Flach P, Zupan B (1999) Rule evaluation measures: a unifying view. In: Proceedings of the 9th international workshop on inductive logic programming (LNAI 1634), pp 174–185
Lee J, Giraud-Carrier C (2011) A metric for unsupervised metalearning. Intell Data Anal 15(6):827–841
Lenca P, Vaillant B, Meyer P, Lallich S (2007) Association rule interestingness measures: experimental and theoretical studies. ReCALL 43:51–76
Lenca P, Meyer P, Vaillant B, Lallich S (2008) On selecting interestingness measures for association rules: user oriented description and multiple criteria decision aid. Eur J Oper Res 184(2):610–626
Lerman I, Gras R, Rostam H (1981a) Elaboration et evaluation d'un indice d' implication pour des données binaires 1. Mathématiques et Sciences Humaines 74:5–35
Lerman I, Gras R, Rostam H (1981b) Elaboration et evaluation d'un indice d' implication pour des données binaires 2. Mathématiques et Sciences Humaines 75:5–47
Li J (2006) On optimal rule discovery. IEEE Trans Knowl Data Eng 18(4):460–471
Liu B, Hsu W, Ma Y (1999) Mining association rules with multiple minimum supports. In: Proceedings of the 5th ACM SIGKDD international conference on knowledge discovery and data mining, pp 337–341
Loevinger J (1947) A systematic approach to the construction and evaluation of tests of ability. Psychol Monogr 61(4):1–49
Mampaey M, Tatti N, Vreeken J (2011) Tell me what I need to know: succinctly summarizing data with itemsets. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, pp 573–581
McGarry K (2005) A survey of interestingness measures for knowledge discovery. Knowl Eng Rev 20(1):39–61
Meilă M (2012) Logical equivalences of distances between clusterings—a geometric perspective. Mach Learn 86(3):369–389
Mosteller F (1968) Association and estimation in contingency tables. J Am Stat Soc 63(321):1–28

Ohsaki M, Sato Y, Yokoi H, Yamaguchi T (2002) A rule discovery support system doe sequential medical data–in the case study of a chronic hepatitis dataset. In: Proceedings of the ICDM workshop on active mining, pp 97–102

Ohsaki M, Kitaguchi S, Yokoi H, Yamaguchi T (2003) Investigation of rule interestingness in medical data mining. In: Proceedings of the 2nd international workshop on active mining (LNAI 3430), pp 174–189

Ohsaki M, Kitaguchi S, Okamoto K, Yokoi H, Yamaguchi T (2004) Evaluation of rule interestingness measures with a clinical dataset on hepatitis. In: Proceedings of the 8th European conference on principles and practice of knowledge discovery in databases (LNAI 3203), pp 362–373

Padmanabhan B (2004) The interestingness paradox in pattern discovery. J Appl Stat 31(8):1019–1035

Peterson A, Martinez T (2005) Estimating the potential for combining learning models. In: Proceedings of the ICML workshop on meta-learning, pp 68–75

Piatetsky-Shapiro G (1991) Discovery, analysis, and presentation of strong rules. In: Piatetsky-Shapiro G, Frawley WJ (eds) Knowledge discovery in databases. AAAI Press, Cambridge, pp 229–248

Plasse M, Niang N, Saportaa G, Villeminot A, Leblond L (2007) Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. Comput Stat Data Anal 52(1):596–613

R Development Core Team (2007) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Ritschard G, Zighed D (2006) Implication strength of classification rules. In: Proceedings of the 16th international symposium on methodologies for intelligent systems (LNCS 4203), pp 463–472

Sahar S (1999) Interestingness via what is not interesting. In: Proceedings of the 5th ACM SIGKDD international conference on knowledge discovery and data mining, pp 332–336

Sahar S (2002) Exploring interestingness through clustering: a framework. In: Proceedings of the 2nd IEEE international conference on data mining, pp 677–680

Sahar S (2003) What is interesting: studies on interestingness in knowledge discovery. PhD thesis, School of Computer Science, Tel-Aviv University

Sahar S (2010) Interestingness measures—on determining what is interesting. In: Maimon O, Rokach L (eds) Data mining and knowledge discovery handbook, 2nd edn. Springer, New York, pp 603–612

Sebag M, Schoenauer M (1988) Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases. In: Proceedings of the European knowledge acquisition, workshop, pp 28.1-28.20

Silver N, Dunlap W (1987) Averaging correlation coefficients: should Fisher's z transformation be used? J Appl Psychol 72(1):146–148

Smyth P, Goodman R (1992) An information theoretic approach to rule induction from databases. IEEE Trans Knowl Data Eng 4(4):301–316

Spyropoulou E, De Bie T (2011) Interesting multi-relational patterns. In: Proceedings of the 11th international conference on data mining, pp 675–684

Stiglic G, Kokol P (2009) GEMLeR: gene expression machine learning repository. Faculty of Health Sciences, University of Maribor. http://gemler.fzv.uni-mb.si/

Tan P, Kumar V (2000) Interestingness measures for association patterns: a perspective. In: Proceedings of the KDD'00 workshop on postprocessing in machine learning and data mining

Tan P, Kumar V, Srivastava J (2002) Selecting the right interestingness measure for association patterns. In: Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining, pp 32–41

Tan P, Kumar V, Srivastava J (2004) Selecting the right objective measure for association analysis. Inf Syst 29(4):293–313

Tatti N, Mampaey M (2010) Using background knowledge to rank itemsets. Data Min Knowl Discov 21(2):293–309

Vaillant B, Lenca P, Lallich S (2004) A clustering of interestingness measures. In: Proceedings of the 7th international conference on discovery science (LNAI 3245), pp 290–297

Verhein F, Chawla S (2007) Using significant, positively associated and relatively class correlated rules for associative classification of imbalanced datasets. In: Proceedings of the 7th IEEE international conference on data mining, pp 679–684

Webb G (2006) Discovery significant rule. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, pp 434–443

Webb G (2010) Self-sufficient itemsets: an approach to screening potentially interesting associations between items. ACM Trans Knowl Discov Data 4(1):3:1–3:20

Webb G (2011) Filtered-top-k association discovery. Wiley Interdiscip Rev Data Min Knowl Discov 1(3):183–192

Winitzki S (2003) Uniform approximation for transcendental functions. In: Proceedings of the international conference on computational science and its applications, part I (LNCS 2667), pp 780–789

Winitzki S (2008) A handy approximation for the error function and its inverse. http://www.scribd.com/doc/82414963/Winitzki-Approximation-to-Error-Function. Accessed 20 June 2012

Witten I, Eibe F (2000) Data mining: practical machine learning tools with Java implementations. Morgan Kaufmann, San Francisco

Wu T, Chen Y, Han J (2010) Re-examination of interestingness measures in pattern mining: a unified framework. Data Min Knowl Discov 21(3):371–397

Yao J, Liu H (1997) Searching multiple databases for interesting complexes. In: Proceedings of the 1st Pacific-Asia conference on knowledge discovery and data mining

Yao Y, Zhong N (1999) An analysis of quantitative measures associated with rules. In: Proceedings of the 3rd Pacific-Asia conference on knowledge discovery and data mining (LNCS 1574), pp 479–488

Yule G (1900) On the association of attributes in statistics: with illustrations from the material of the childhood society, &c. Philos Trans R Soc A 194:257–319

Yule G (1912) On the methods of measuring association between two attributes. J R Stat Soc 75(6):579–652

Zhang T (2000) Association rules. In: Proceedings of the 4th Pacific-Asia conference on knowledge discovery and data mining (LNAI 1805), pp 245–256