Selecting the right objective measure for association analysis

Article *in* Information Systems · June 2004

DOI: 10.1016/S0306-4379(03)00072-3 · Source: DBLP

Vipin Kumar
University of California, San Diego

Jaideep Srivastava
University of Minnesota Twin Cities

Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

ELSEVIER

Information Systems 29 (2004) 293–313

www.elsevier.com/locate/infosys

Selecting the right objective measure for association analysis☆

Pang-Ning Tan*, Vipin Kumar, Jaideep Srivastava

Department of Computer Science, University of Minnesota, 200 Union Street SE, Minneapolis, MN 55455, USA

Abstract

Objective measures such as support, confidence, interest factor, correlation, and entropy are often used to evaluate the interestingness of association patterns. However, in many situations, these measures may provide conflicting information about the interestingness of a pattern. Data mining practitioners also tend to apply an objective measure without realizing that there may be better alternatives available for their application. In this paper, we describe several key properties one should examine in order to select the right measure for a given application. A comparative study of these properties is made using twenty-one measures that were originally developed in diverse fields such as statistics, social science, machine learning, and data mining. We show that depending on its properties, each measure is useful for some application, but not for others. We also demonstrate two scenarios in which many existing measures become consistent with each other, namely, when support-based pruning and a technique known as table standardization are applied. Finally, we present an algorithm for selecting a small set of patterns such that domain experts can find a measure that best fits their requirements by ranking this small set of patterns.

1. Introduction

The analysis of relationships between variables is a fundamental task at the heart of many data mining problems. For example, the central task of association analysis [1,2] is to discover sets of binary variables (called items) that *co-occur* together frequently in a transaction database, while the goal of feature selection is to identify groups of variables that are highly *correlated* with each other, or with respect to a specific target variable. Regardless of how the relationships are defined, such analyses often require a suitable measure to evaluate the dependencies between variables. For example, objective measures such as support, confidence, interest factor, correlation, and entropy have been used extensively to evaluate the interestingness of association patterns—the stronger is the dependence relationship, the more interesting is the pattern. These objective measures are defined in terms of the frequency counts tabulated in a 2×2 contingency table, as shown in Table 1.

Although there are numerous measures available for evaluating association patterns, a significant number of them provide conflicting information about the interestingness of a pattern.

☆Recommended by Nick Koudas. This work was partially supported by NSF grant number ACI-9982274, DOE contract number DOE/LLNL W-7045-ENG-48 and by the Army High Performance Computing Research Center contract number DAAD19-01-2-0014. The content of this work does not necessarily reflect the position or policy of the government and no official endorsement should be inferred. Access to computing facilities was provided by AHPCRC and the Minnesota Supercomputing Institute.

*Corresponding author. Tel.: +1-612-626-8083.

E-mail addresses: ptan@cs.umn.edu (P.-N. Tan), kumar@cs.umn.edu (V. Kumar), srivasta@cs.umn.edu (J. Srivastava).

0306-4379/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0306-4379(03)00072-3

Table 1
A 2 × 2 contingency table for items A and B

	B	$\overline{B}$	
A	f_{11}	f_{10}	f_{1+}
$\overline{A}$	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	N

Table 2
Ten examples of contingency tables

Contingency table	f_{11}	f_{10}	f_{01}	f_{00}
E1	8123	83	424	1370
E2	8330	2	622	1046
E3	9481	94	127	298
E4	3954	3080	5	2961
E5	2886	1363	1320	4431
E6	1500	2000	500	6000
E7	4000	2000	1000	3000
E8	4000	2000	2000	2000
E9	1720	7121	5	1154
E10	61	2483	4	7452

To illustrate this, consider the 10 contingency tables, E1–E10, shown in Table 2. Table 3 shows the ranking of these tables according to 21 different measures developed in diverse fields such as statistics, social science, machine learning, and data mining.[1] The table also shows that different measures can lead to substantially different rankings of contingency tables. For example, E10 is ranked highest according to the I measure but lowest according to the ϕ-coefficient; while $E3$ is ranked lowest by the AV measure but highest by the *IS* measure. Thus, selecting the right measure for a given application poses a dilemma because many measures may disagree with each other.

To understand why some of the measures are inconsistent, we need to examine the properties of each measure. In this paper, we present several key properties one should consider in order to select the right measure for a given application. Some of these properties are well-known to the data mining community, while others, which are as important, have received less attention. One such property is the invariance of a measure under row and column scaling operations. We illustrate this with the following classic example by Mosteller [3].

Table 4(a) and (b) illustrates the relationship between the gender of a student and the grade obtained for a particular course for two different samples. Note that the sample used in Table 4(b) contains twice the number of male students in Table 4(a) and 10 times the number of female students. However, the relative performance of male students is the same for both samples and the same applies to the female students. Mosteller concluded that the dependencies in both tables are equivalent because the underlying association between gender and grade should be independent of the relative number of male and female students in the samples [3]. Yet, as we show later, many intuitively appealing measures, such as the ϕ-coefficient, mutual information, Gini index or cosine measure, are sensitive to scaling of rows and columns of the table. Although there are measures that consider the association in both tables to be equivalent (e.g., odds ratio [3]), they have properties that make them unsuitable for other applications.

In this paper, we perform a comparative study of the properties for 21 existing objective measures. Despite the general lack of agreement among many of these measures, there are two situations in which they become consistent with each other. First, we show that the rankings produced by many measures become highly correlated when support-based pruning is used. Support-based pruning also tends to eliminate mostly uncorrelated and poorly correlated patterns. Second, we show that a technique known as table standardization [3,4] can also be used to make the measures consistent with each other.

An alternative way to find a desirable measure is by comparing how well the rankings produced by each measure agree with the expectations of domain experts. This would require the domain experts to manually rank all the contingency tables extracted from data, which is quite a laborious task. Instead, we show that it is possible to identify a small set of “well-separated” contingency tables such that finding the most suitable measure using this small set of tables is almost equivalent to finding the best measure using the entire data set.

[1] A complete definition of these measures is given in Section 2.

Table 3
Rankings of contingency tables using various objective measures. (lower number means higher rank)

Contingency Table	ϕ	λ	α	Q	Y	κ	M	J	G	s	c	L	V	I	IS	PS	F	AV	S	ζ	K
E1	1	1	3	3	3	1	2	2	1	3	5	5	4	6	2	2	4	6	1	2	5
E2	2	2	1	1	1	2	1	3	2	2	1	1	1	8	3	5	1	8	2	3	6
E3	3	3	4	4	4	3	3	8	7	1	4	4	6	10	1	8	6	10	3	1	10
E4	4	7	2	2	2	5	4	1	3	6	2	2	2	4	4	1	2	3	4	5	1
E5	5	4	8	8	8	4	7	5	4	7	9	9	9	3	6	3	9	4	5	6	3
E6	6	6	7	7	7	7	6	4	6	9	8	8	7	2	8	6	7	2	7	8	2
E7	7	5	9	9	9	6	8	6	5	4	7	7	8	5	5	4	8	5	6	4	4
E8	8	9	10	10	10	8	10	10	8	4	10	10	10	9	7	7	10	9	8	7	9
E9	9	9	5	5	5	9	9	7	9	8	3	3	3	7	9	9	3	7	9	9	8
E10	10	8	6	6	6	10	5	9	10	10	6	6	5	1	10	10	5	1	10	10	7

Table 4
The grade-gender example

	Male	Female	
High	2	3	5
Low	1	4	5
	3	7	10

(a)

	Male	Female	
High	4	30	34
Low	2	40	42
	6	70	76

(b)

1.1. Paper contribution

The specific contributions of this paper are as follows.

1. We present an overview of 21 objective measures that were proposed in the statistics, social science, machine learning, and data mining literature. We show that application of different measures may lead to substantially differing orderings of patterns.
2. We present several key properties that will help analysts to select the right measure for a given application. A comparative study of these properties is made using the twenty-one existing measures. Our results suggest that we can identify several groups of consistent measures having similar properties.
3. We illustrate two situations in which most of the measures become consistent with each other, namely, when support-based pruning and a technique known as table standardization are used. We also demonstrate the utility of support-based pruning in terms of eliminating uncorrelated and poorly correlated patterns.
4. We present an algorithm for selecting a small set of tables such that domain experts can determine the most suitable measure by looking at their rankings for this small set of tables.

1.2. Related work

The problem of analyzing objective measures used by data mining algorithms has attracted considerable attention in recent years [5–11]. Piatetsky-Shapiro proposed three principles that must be satisfied by any reasonable objective measures. Our current work analyzes the properties of existing measures using these principles as well as several additional properties.

Bayardo et al. [9] compared the optimal rules selected by various objective measures. They showed that given a collection of rules $A \rightarrow B$, where B is fixed, the most interesting rules selected by many well-known measures reside along the support-confidence border. There is an intuitive reason for this observation. Because the rule consequent is fixed, the objective measure is a function of only two parameters, $P(A,B)$ and $P(A)$, or equivalently, the rule support $P(A,B)$ and rule confidence $P(B|A)$. More importantly, Bayardo et al. showed that many well-known measures are monotone functions of support and confidence, which explains the reason for the optimal rules to be located along the support-confidence border. Our work is quite different because our analysis is not limited to rules that have identical consequents. In addition, we focus on understanding the properties of existing measures under certain transformations (e.g., support-based pruning and scaling of rows or columns of contingency tables).

Hilderman et al. [6,7] compared the various *diversity* measures used for ranking data summaries. Each summary is a relational table containing a set of attribute-domain pairs and a derived attribute called Count, which indicates the number of objects aggregated by each tuple in the summary table. Diversity measures are defined according to the distribution of Count attribute values. In [7], the authors proposed five principles a good measure must satisfy to be considered useful for ranking summaries. Some of these principles are similar to the ones proposed by Piatetsky-Shapiro, while others may not be applicable to association analysis because they assume that the Count attribute values are in certain sorted order (such ordering is less intuitive for contingency tables).

Kononenko [8] investigated the properties of measures used in the construction of decision trees. The purpose of his work is to illustrate the effect of the number of classes and attribute values on the value of a measure. For example, he showed that the values for measures such as Gini index and J-measure increase linearly with the number of attribute values. In contrast, the focus of our work is to study the general properties of objective measures for binary-valued variables. Gavrilov et al. [10] and Zhao et al. [11] compared the various objective functions used by clustering algorithms. In both of these methods, it was assumed that the ground truth, i.e., the ideal cluster composition, is known a priori. Such an assumption is not needed in our approach for analyzing the properties of objective measures. However, they might be useful for validating whether the selected measure agrees with the expectation of domain experts.

1.3. Paper organization

The remainder of this paper is organized as follows. In Section 2, we present an overview of the various measures examined in this paper. Section 3 describes a method to determine whether two measures are consistent with each other. Section 4 presents several key properties for analyzing and comparing objective measures. Section 5 describes the effect of applying support-based pruning while Section 6 describes the effect of table standardization. Section 7 presents an algorithm for selecting a small set of tables to be ranked by domain experts. Finally, we conclude with a summary and directions for future work.

2. Overview of objective measures

Table 5 provides the list of measures examined in this study. The definition for each measure is given in terms of the probabilities estimated from a 2×2 contingency table.

3. Consistency between measures

Let $T(D) = \{t_1, t_2, \ldots, t_N\}$ denote the set of 2×2 contingency tables derived from a data set D. Each table represents the relationship between a pair of binary variables. Also, let $\mathbf{M}$ be the set of objective measures available for our analysis. For each measure, $M_i \in \mathbf{M}$, we can construct an *interest vector* $M_i(T) = \{m_{i1}, m_{i2}, \ldots, m_{iN}\}$, where m_{ij} corresponds to the value of M_i for table t_j. Each interest vector can also be transformed into a *ranking vector* $O_i(T) = \{o_{i1}, o_{i2}, \ldots, o_{iN}\}$, where o_{ij} corresponds to the rank of m_{ij} and $\forall j, k : o_{ij} \leqslant o_{ik}$ if and only if $m_{ik} \geqslant m_{ij}$.

We can define the consistency between a pair of measures in terms of the similarity between their ranking vectors. For example, consider the pair of ranking vectors produced by ϕ and κ in Table 3. Since their rankings are very similar, we may conclude that both measures are highly consistent with each other, with respect to the data set shown in Table 2. In contrast, comparison between the ranking vectors produced by ϕ and I suggests that both measures are not consistent with each other.

There are several measures available for computing the similarity between a pair of ranking vectors. This include Spearman's rank coefficient, Pearson's correlation, cosine measure, and the

Table 5
Objective measures for association patterns

#	Measure	Definition
1	ϕ-coefficient	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
2	Goodman-Kruskal's (λ)	$\frac{\sum_j \max_k P(A_j,B_k)+\sum_k \max_j P(A_j,B_k)-\max_j P(A_j)-max_k P(B_k)}{2-\max_j P(A_j)-\max_k P(B_k)}$
3	Odds ratio (α)	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(A,\overline{B})P(\overline{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(\overline{AB})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{AB})+P(A,\overline{B})P(\overline{A},B)} = \frac{\alpha-1}{\alpha+1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\overline{AB})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{AB})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$
6	Kappa (κ)	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
7	Mutual Information (M)	$\frac{\sum_i \sum_j P(A_i,B_j) \log \frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i) \log P(A_i), -\sum_j P(B_j) \log P(B_j))}$
8	J-Measure (J)	$\max\big(P(A,B)\log(\frac{P(B\mid A)}{P(B)}) + P(A\overline{B})\log(\frac{P(\overline{B}\mid A)}{P(\overline{B})}),$ $P(A,B)\log(\frac{P(A\mid B)}{P(A)}) + P(\overline{A}B)\log(\frac{P(\overline{A}\mid B)}{P(\overline{A})})\big)$
9	Gini index (G)	$\max\big(P(A)[P(B\mid A)^2 + P(\overline{B}\mid A)^2] + P(\overline{A})[P(B\mid \overline{A})^2 + P(\overline{B}\mid \overline{A})^2]$ $-P(B)^2 - P(\overline{B})^2,$ $P(B)[P(A\mid B)^2 + P(\overline{A}\mid B)^2] + P(\overline{B})[P(A\mid \overline{B})^2 + P(\overline{A}\mid \overline{B})^2]$ $-P(A)^2 - P(\overline{A})^2\big)$
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B\mid A), P(A\mid B))$
12	Laplace (L)	$\max\big(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2}\big)$
13	Conviction (V)	$\max\big(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})}\big)$
14	Interest (I)	$\frac{P(A,B)}{P(A)P(B)}$
15	cosine (IS)	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
16	Piatetsky-Shapiro's (PS)	$P(A,B) - P(A)P(B)$
17	Certainty factor (F)	$\max\big(\frac{P(B\mid A)-P(B)}{1-P(B)}, \frac{P(A\mid B)-P(A)}{1-P(A)}\big)$
18	Added Value (AV)	$\max(P(B\mid A) - P(B), P(A\mid B) - P(A))$
19	Collective strength (S)	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})} \times \frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
21	Klosgen (K)	$\sqrt{P(A,B)}\max(P(B\mid A) - P(B), P(A\mid B) - P(A))$

A summary description for each measure:

ϕ-coefficient.[4]. This measure is analogous to Pearson's product–moment correlation coefficient for continuous variables. It is closely related to the χ^2 statistic since $\phi^2 = \chi^2/N$. Although the χ^2 statistic is often used for goodness of fit testing, it is seldom used as a measure of association because it depends on the size of the database [3].

λ-coefficient [12]. The λ coefficient, also known as the index of predictive association, was initially proposed by Goodman and Kruskal [12]. The intuition behind this measure is that if two variables are highly dependent on each other, then the error in predicting one of them would be small whenever the value of the other variable is known. λ is used to capture the amount of reduction in the prediction error.

Odds ratio [3]. This measure represents the odds for obtaining the different outcomes of a variable. For example, consider the frequency counts given in Table 3. If B is present, then the odds of finding A in the same transaction is f_{11}/f_{01}. On the other hand, if B is absent, then the odds for finding A is f_{10}/f_{00}. If there is no association between A and B, then the odds for finding A in a transaction should remain the same, regardless of whether B is present in the transaction. We can use the ratio of these odds, $(f_{11}f_{00}/f_{01}f_{10})$, to determine the degree to which A and B are associated with each other.

Yule's Q- [13] *and Y-coefficients* [14]. The value for odds ratio ranges from 0 (for perfect negative correlation) to ∞ (for perfect positive correlation). Yule's Q and Y coefficients are normalized variants of the odds ratio, defined in a way that they range from -1 to $+1$.

κ-coefficient [15]. This measure captures the degree of agreement between a pair of variables. If the variables agree with each other, then the values for $P(A,B)$ and $P(\bar{A},\bar{B})$ will be large, which in turn, results in a higher value for κ.

Entropy [16], *J-measure* [17], *and Gini* [18]. Entropy is related to the variance of a probability distribution. The entropy of a uniform distribution is large, whereas the entropy of a skewed distribution is small. Mutual information is an entropy-based measure for

Table 5
(*Continued*)

evaluating the dependencies among variables. It represents the amount of reduction in the entropy of a variable when the value of a second variable is known. If the two variables are strongly associated, then the amount of reduction in entropy, i.e., its mutual information, is high. Other measures defined according to the probability distribution of variables include J-Measure [17] and Gini index [18].

Support [1]. Support is often used to represent the significance of an association pattern [1,19]. It is also useful from a computational perspective because it has a nice downward closure property that allows us to prune the exponential search space of candidate patterns.

Confidence, *Laplace* [20], *and Conviction* [21]. Confidence is often used to measure the accuracy of a given rule. However, it can produce misleading results, especially when the support of the rule consequent is higher than the rule confidence [22]. Other variants of the confidence measure include the Laplace function [20] and conviction [21].

Interest factor [22–26]. This measure is used quite extensively in data mining for measuring deviation from statistical independence. However, it is sensitive to the support of the items (f_{1+} or f_{+1}). DuMouchel has recently proposed a statistical correction to I for small sample sizes, using an empirical Bayes technique [26]. Other variants of this measure include Piatetsky-Shapiro's rule-interest [5], certainty factor [27], collective strength [28] and added value [29].

IS measure [30]. This measure can be derived from the ϕ-coefficient [30]. It is the geometric mean between interest factor (I) and the support measure (s). The *IS* measure for pairs of items is also equivalent to the cosine measure, which is a widely-used similarity measure for vector-space models.

Jaccard [31] *and Klosgen measures* [32]. The Jaccard measure [31] is used extensively in information retrieval to measure the similarity between documents, while Klosgen K measure [32] was used by the Explora knowledge discovery system.

inverse of the L_2-norm. Our experimental results suggest that there is not much difference between using any one of the measures as our similarity function. In fact, if the values within each ranking vector is unique, we can prove that Pearson's correlation, cosine measure and the inverse of the L_2-norm are all monotonically related. Thus, we decide to use Pearson's correlation as our similarity measure.

Definition 1 (Consistency between measures). Two measures, M_1 and M_2, are consistent each other with respect to data set D if the *correlation* between $O_1(T)$ and $O_2(T)$ is greater than or equal to some positive threshold t.[2]

4. Properties of objective measures

In this section, we describe several important properties of an objective measure. While some of these properties have been extensively investigated in the data mining literature [5,33], others are not well-known.

4.1. Desired properties of a measure

Piatetsky-Shapiro [5] has proposed three key properties a good measure M should satisfy:

P1: $M = 0$ if A and B are statistically independent;
P2: M monotonically increases with $P(A, B)$ when $P(A)$ and $P(B)$ remain the same;
P3: M monotonically decreases with $P(A)$ (or $P(B)$) when the rest of the parameters ($P(A, B)$ and $P(B)$ or $P(A)$) remain unchanged.

These properties are well-known and have been extended by many authors [7,33]. Table 6 illustrates the extent to which each of the existing measure satisfies the above properties.

4.2. Other properties of a measure

There are other properties that deserve further investigation. These properties can be described using a matrix formulation. In this formulation, each 2×2 contingency table is represented by a *contingency matrix*, $\mathbf{M} = [\mathbf{f}_{11}\mathbf{f}_{10}; \mathbf{f}_{01}\mathbf{f}_{00}]$ while each objective measure is a matrix operator, O, that maps the matrix $\mathbf{M}$ into a scalar value, k, i.e., $O\mathbf{M} = k$. For instance, the ϕ coefficient is equivalent to a normalized form of the determinant operator, where $Det(\mathbf{M}) = f_{11}f_{00} - f_{01}f_{10}$. Thus, statistical independence is represented by a singular

[2] The choice for t can be tied to the desired significance level of correlation. The critical value for correlation depends on the number of independent tables available and the confidence level desired. For example, at 99% confidence level and 50 degrees of freedom, any correlation above 0.35 is statistically significant.

Table 6
Properties of objective measures. Note that none of the measures satisfies all the properties

Symbol	Measure	Range	P1	P2	P3	O1	O2	O3	O3'	O4
ϕ	ϕ-coefficient	$-1\cdots 0\cdots 1$	Yes	Yes	Yes	Yes	No	Yes	Yes	No
λ	Goodman-Kruskal's	$0\cdots 1$	Yes	No	No	Yes	No	No*	Yes	No
α	odds ratio	$0\cdots 1\cdots\infty$	Yes*	Yes	Yes	Yes	Yes	Yes*	Yes	No
Q	Yule's Q	$-1\cdots 0\cdots 1$	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Y	Yule's Y	$-1\cdots 0\cdots 1$	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
κ	Cohen's	$-1\cdots 0\cdots 1$	Yes	Yes	Yes	Yes	No	No	Yes	No
M	Mutual information	$0\cdots 1$	Yes	Yes	Yes	No**	No	No*	Yes	No
J	J-Measure	$0\cdots 1$	Yes	No	No	No**	No	No	No	No
G	Gini index	$0\cdots 1$	Yes	No	No	No**	No	No*	Yes	No
s	Support	$0\cdots 1$	No	Yes	No	Yes	No	No	No	No
c	Confidence	$0\cdots 1$	No	Yes	No	No**	No	No	No	Yes
L	Laplace	$0\cdots 1$	No	Yes	No	No**	No	No	No	No
V	Conviction	$0.5\cdots 1\cdots\infty$	No	Yes	No	No**	No	No	Yes	No
I	Interest	$0\cdots 1\cdots\infty$	Yes*	Yes	Yes	Yes	No	No	No	No
IS	Cosine	$0\cdots\sqrt{P(A,B)}\cdots 1$	No	Yes	Yes	Yes	No	No	No	Yes
PS	Piatetsky-Shapiro's	$-0.25\cdots 0\cdots 0.25$	Yes	Yes	Yes	Yes	No	Yes	Yes	No
F	Certainty factor	$-1\cdots 0\cdots 1$	Yes	Yes	Yes	No**	No	No	Yes	No
AV	Added value	$-0.5\cdots 0\cdots 1$	Yes	Yes	Yes	No**	No	No	No	No
S	Collective strength	$0\cdots 1\cdots\infty$	No	Yes	Yes	Yes	No	Yes*	Yes	No
ζ	Jaccard	$0\cdots 1$	No	Yes	Yes	Yes	No	No	No	Yes
K	Klosgen's	$(\frac{2}{\sqrt{3}}-1)^{1/2}[2-\sqrt{3}-\frac{1}{\sqrt{3}}]\cdots 0\cdots\frac{2}{3\sqrt{3}}$	Yes	Yes	Yes	No**	No	No	No	No

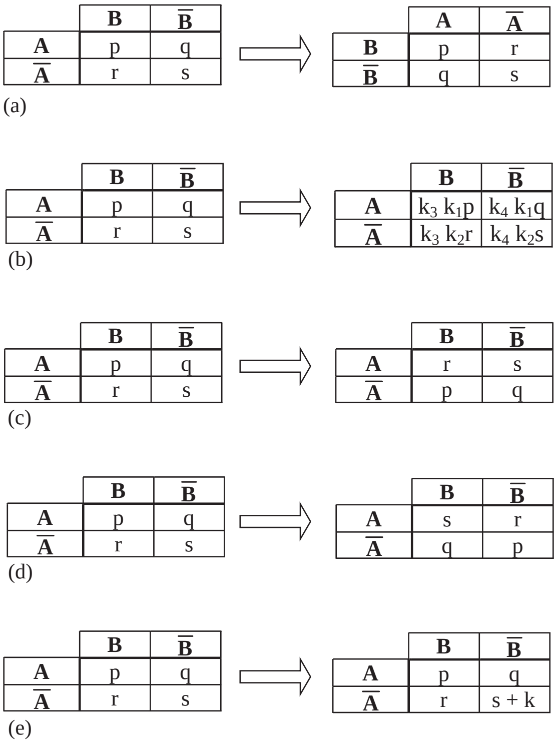
where: P1: $O(\mathbf{M}) = 0$ if $det(\mathbf{M}) = 0$, i.e., whenever A and B are statistically independent.
P2: $O(\mathbf{M_2}) > O(\mathbf{M_1})$ if $\mathbf{M_2} = \mathbf{M_1}+\ [k\ \ -k;\ \ -k\ k]$.
P3: $O(\mathbf{M_2}) < O(\mathbf{M_1})$ if $\mathbf{M_2} = \mathbf{M_1}+\ [0\ k;\ \ 0\ \ -k]$ or $\mathbf{M_2} = \mathbf{M_1} + [0\ 0;\ \ k\ \ -k]$.
O1: Property 1: Symmetry under variable permutation.
O2: Property 2: Row and Column scaling invariance.
O3: Property 3: Antisymmetry under row or column permutation.
O3': Property 4: Inversion invariance.
O4: Property 5: Null invariance.
Yes*: Yes if measure is normalized.
No*: Symmetry under row or column permutation.
No**: No unless the measure is symmetrized by taking $\max(M(A,B), M(B,A))$.

matrix **M** whose determinant is equal to zero. The underlying properties of a measure can be analyzed by performing various operations on the contingency tables as depicted in Fig. 1.

Property 1 (Symmetry under variable permutation). *A measure O is symmetric under variable permutation* (*Fig.* 1(*a*)), $A \leftrightarrow B$, *if* $O(\mathbf{M}^T) = O(\mathbf{M})$ *for all contingency matrices* **M**. *Otherwise*, *it is called an asymmetric measure.*

The asymmetric measures investigated in this study include confidence, laplace, J-Measure, conviction, added value, Gini index, mutual information, and Klosgen measure. Examples of symmetric measures are ϕ-coefficient, cosine (IS), interest factor (I) and odds ratio (α). In practice, asymmetric measures are used for implication rules, where there is a need to distinguish between the strength of the rule $A \rightarrow B$ from $B \rightarrow A$. Since every contingency matrix produces two values when we apply an asymmetric measure, we use the maximum of these two values to be its overall value when we compare the properties of symmetric and asymmetric measures.

Property 2 (Row/column scaling invariance). *Let* $\mathbf{R} = \mathbf{C} = [k_1\ 0;\ 0\ k_2]$ *be a* 2×2 *square matrix*,



	B	$\bar{B}$
A	p	q
$\bar{A}$	r	s

	A	$\bar{A}$
B	p	r
$\bar{B}$	q	s

(a)

	B	$\bar{B}$
A	p	q
$\bar{A}$	r	s

	B	$\bar{B}$
A	$k_3 k_1 p$	$k_4 k_1 q$
$\bar{A}$	$k_3 k_2 r$	$k_4 k_2 s$

(b)

	B	$\bar{B}$
A	p	q
$\bar{A}$	r	s

	B	$\bar{B}$
A	r	s
$\bar{A}$	p	q

(c)

	B	$\bar{B}$
A	p	q
$\bar{A}$	r	s

	B	$\bar{B}$
A	s	r
$\bar{A}$	q	p

(d)

	B	$\bar{B}$
A	p	q
$\bar{A}$	r	s

	B	$\bar{B}$
A	p	q
$\bar{A}$	r	s + k

(e)

Fig. 1. Operations on a contingency table. (a) Variable permutation operation. (b) Row and column scaling operation. (c) Row and column permutation operation. (d) Inversion operation. (e) Null addition operation.

where k_1 and k_2 are positive constants. The product $\mathbf{R} \times \mathbf{M}$ *corresponds to scaling the first row of matrix* $\mathbf{M}$ *by k_1 and the second row by k_2, while the product* $\mathbf{M} \times \mathbf{C}$ *corresponds to scaling the first column of* $\mathbf{M}$ *by k_1 and the second column by k_2 (Fig. 1(b)). A measure O is invariant under row and column scaling if* $O(\mathbf{RM}) = O(\mathbf{M})$ *and* $O(\mathbf{MC}) = O(\mathbf{M})$ *for all contingency matrices*, $\mathbf{M}$.

Odds ratio (α) along with Yule's Q and Y coefficients are the only measures in Table 6 that are invariant under the row and column scaling operations. This property is useful for data sets containing nominal variables such as Mosteller's grade-gender example in Section 1.

Property 3 (Antisymmetry under row/column permutation). Let $\mathbf{S} = [0\ 1;\ 1\ 0]$ *be a* 2×2 *permutation matrix. A normalized*[3] *measure O is antisymmetric under the row permutation operation if* $O(\mathbf{SM}) = -O(\mathbf{M})$, *and antisymmetric under the column permutation operation if* $O(\mathbf{MS}) = -O(\mathbf{M})$ *for all contingency matrices* $\mathbf{M}$ (Fig. 1(c)).

The ϕ-coefficient, PS, Q and Y are examples of antisymmetric measures under the row and column permutation operations while mutual information and Gini index are examples of symmetric measures. Asymmetric measures under this operation include support, confidence, IS and interest factor. Measures that are symmetric under the row and column permutation operations do not distinguish between positive and negative correlations of a table. One should be careful when using them to evaluate the interestingness of a pattern.

Property 4 (Inversion invariance). *Let* $\mathbf{S} = [0\ 1;\ 1\ 0]$ *be a* 2×2 *permutation matrix. A measure O is invariant under the inversion operation* (*Fig.* 1(d)) *if* $O(\mathbf{SMS}) = O(\mathbf{M})$ *for all contingency matrices* $\mathbf{M}$.

Inversion is a special case of the row/column permutation where both rows and columns are swapped simultaneously. We can think of the inversion operation as flipping the 0's (absence) to become 1's (presence), and vice versa. This property allows us to distinguish between symmetric binary measures, which are invariant under the inversion operation, from asymmetric binary measures. Examples of symmetric binary measures include ϕ, odds ratio, κ and collective strength, while the examples for asymmetric binary measures include I, IS, PS and Jaccard measure.

We illustrate the importance of inversion invariance with an example depicted in Fig. 2. In this figure, each column vector is a vector of transactions for a particular item. It is intuitively clear that the first pair of vectors, A and B, have very little association between them. The second pair of vectors, C and D, are inverted versions of vectors A and B. Despite the fact that both C and D co-occur together more frequently, their ϕ

[3] A measure is normalized if its value ranges between -1 and $+1$. An unnormalized measure U that ranges between 0 and $+\infty$ can be normalized via transformation functions such as $(U-1)/(U+1)$ or $(\tan^{-1}\log(U))/\pi/2$.

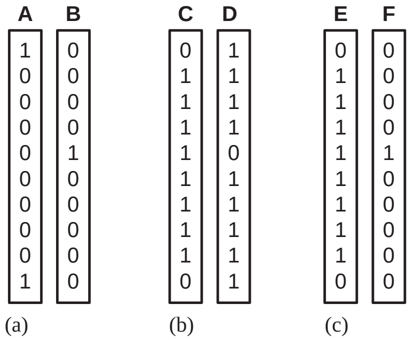


Fig. 2. Comparison between the ϕ-coefficients for three pairs of vectors. The ϕ values for (a), (b) and (c) are −0.1667, −0.1667 and 0.1667, respectively.

coefficient are still the same as before. In fact, it is smaller than the ϕ-coefficient of the third pair of vectors, E and F, for which $E = C$ and $F = B$. This example demonstrates the drawback of using ϕ-coefficient and other symmetric binary measures for applications that require unequal treatments of the binary values of a variable, such as market basket analysis [34].

Other matrix operations, such as matrix addition, can also be applied to a contingency matrix. For example, the second property, P2, proposed by Piatetsky-Shapiro is equivalent to adding the matrix $\mathbf{M}$ with $[k\ \ -k;\ \ -k\ \ k]$, while the third property, $P3$, is equivalent to adding $[0\ \ k;\ \ 0\ \ -k]$ or $[0\ \ 0;\ \ k\ \ -k]$ to $\mathbf{M}$.

Property 5 (Null invariance). *A measure is null-invariant if* $O(\mathbf{M} + \mathbf{C}) = O(\mathbf{M})$ *where* $C = [0\ \ 0;\ \ 0\ \ k]$ *and k is a positive constant.*

For binary variables, this operation corresponds to adding more records that do not contain the two variables under consideration, as shown in Fig. 1(e). Some of the null-invariant measures include IS (cosine) and the Jaccard similarity measure, ζ. This property is useful for domains having sparse data sets, where co-presence of items is more important than co-absence. Examples include market-basket data and text documents.

4.3. Summary

The discussion in this section suggests that there is no measure that is consistently better than others in all application domains. This is because different measures have different intrinsic properties, some of which may be desirable for certain applications but not for others. Thus, in order to find the right measure, we need to match the desired properties of an application against the properties of the existing measures. This can be done by computing the similarity between a property vector that represents the desired properties of the application with the property vectors that represent the intrinsic properties of existing measures. Each component of the property vector corresponds to one of the columns given in Table 6. Since property P1 can be satisfied trivially by rescaling some of the measures, it is not included in the property vector. Each vector component can also be weighted according to its level of importance to the application.

Fig. 3 shows the correlation between the property vectors of various measures. Observe that there are several groups of measures with very similar properties, as shown in Table 7. Some of these groupings are quite obvious, e.g., Groups 1 and 2, while others are quite unexpected, e.g., Groups 3, 6, and 7. In the latter case, since the properties listed in Table 6 are not necessarily comprehensive, we do not expect to distinguish all the available measures using these properties.

5. Effect of support-based pruning

Support-based pruning is often used as a pre-filter prior to the application of other objective measures such as confidence, ϕ-coefficient, interest factor, etc. Because of its anti-monotone property, support allows us to effectively prune the exponential number of candidate patterns. Beyond this, little else is known about the advantages of applying this strategy. The purpose of this section is to discuss two additional effects it has on the rest of the objective measures.

5.1. Elimination of poorly correlated contingency tables

First, we will analyze the quality of patterns eliminated by support-based pruning. Ideally, we

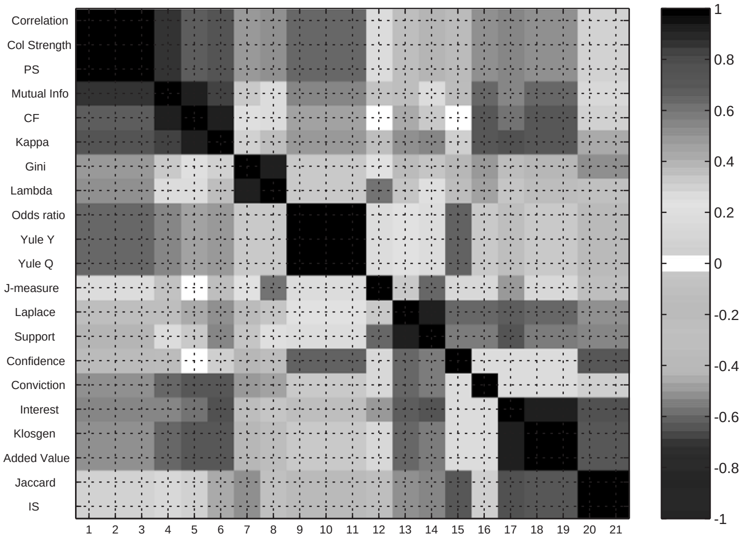


Fig. 3. Correlation between measures based on their property vector. Note that the column labels are the same as the row labels.

Table 7
Groups of objective measures with similar properties

Group	Objective measures
1	Odds ratio, Yule's Q, and Yule's Y.
2	Cosine (IS) and Jaccard.
3	Support and Laplace.
4	ϕ-coefficient, collective strength, and PS.
5	Gini index and λ.
6	Interest factor, added value, and Klosgen K.
7	Mutual information, certainty factor, and κ.

prefer to eliminate only patterns that are poorly correlated. Otherwise, we may end up missing too many interesting patterns.

To study this effect, we have created a synthetic data set that contains $100,000$ 2×2 contingency tables. Each table contains randomly populated f_{ij} values subjected to the constraint $\sum_{i,j} f_{ij} = 1$. The support and ϕ-coefficient for each table can be computed using the formula shown in Table 5. By examining the distribution of ϕ-coefficient values, we can determine whether there are any highly correlated patterns inadvertently removed as a result of support-based pruning.

For this analysis, we apply two kinds of support-based pruning strategies. The first strategy is to impose a minimum support threshold on the value of f_{11}. This approach is identical to the support-based pruning strategy employed by most of the association analysis algorithms. The second strategy is to impose a maximum support threshold on both f_{1+} and f_{+1}. This strategy is equivalent to removing the most frequent items from a data set (e.g., staple products such as sugar, bread, and milk). The results for both of these experiments are illustrated in Figs. 4(a) and (b).

For the entire data set of 100,000 tables, the ϕ-coefficients are normally distributed around $\phi = 0$, as depicted in the upper left-hand corner of both graphs. When a maximum support threshold is imposed, the ϕ-coefficient of the eliminated tables follows a bell-shaped distribution, as shown in Fig. 4(a). In other words, imposing a maximum support threshold tends to eliminate uncorrelated, positively correlated, and negatively correlated tables at equal proportions. This observation can be explained by the nature of the synthetic data—since the frequency counts of the contingency tables are generated randomly, the eliminated tables also have a very similar distribution as the ϕ-coefficient distribution for the entire data.

On the other hand, if a lower bound of support is specified (Fig. 4(b)), most of the contingency tables removed are either uncorrelated ($\phi = 0$) or negatively correlated ($\phi < 0$). This observation is

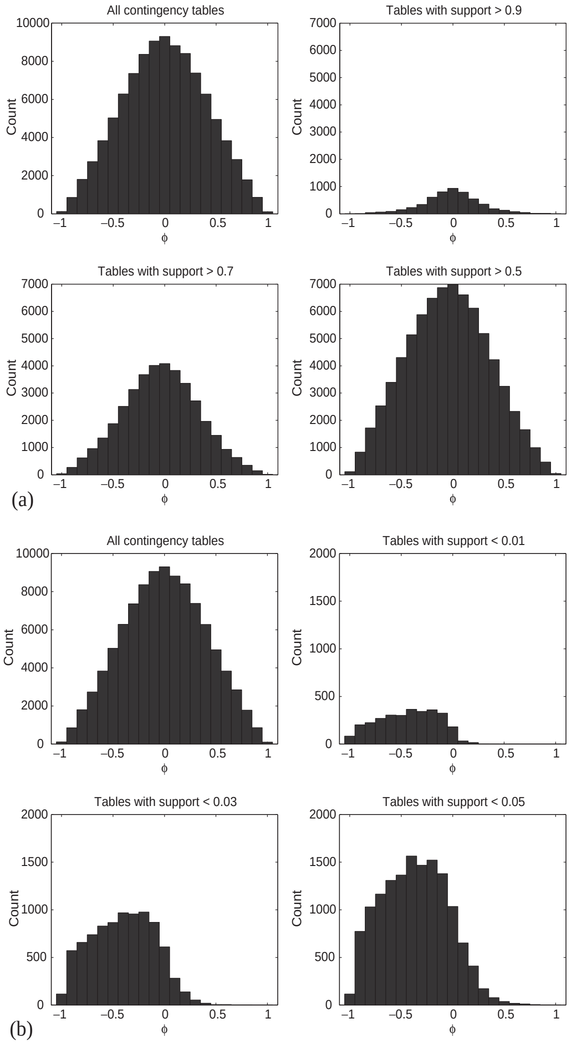


Fig. 4. Effect of support pruning on contingency tables. (a) Distribution of ϕ-coefficient for contingency tables that are removed by applying a maximum support threshold. (b) Distribution of ϕ-coefficient for contingency tables that are removed by applying a minimum support threshold.

quite intuitive because, for a contingency table with low support, at least one of the values for f_{10}, f_{01} or f_{00} must be relatively high to compensate for the low frequency count in f_{11}. Such tables tend to be uncorrelated or negatively correlated unless their f_{00} values are extremely high. This observation is also consistent with the property $P2$ described in Section 4.2, which states that an objective measure should increase as the support count increases.

Support-based pruning is a viable technique as long as only positively correlated tables are of interest to the data mining application. One such situation arises in market basket analysis where such a pruning strategy is used extensively.

5.2. *Consistency of measures under support constraints*

Support-based pruning also affects the issue of consistency among objective measures. To illustrate this, consider the diagram shown in Fig. 5. The figure is obtained by generating a synthetic data set similar to the previous section except that the contingency tables are non-negatively correlated. Convex measures such as mutual information, Gini index, J-measure, and λ assign positive values to their negatively-correlated tables. Thus, they tend to prefer negatively correlated tables over uncorrelated ones, unlike measures such as ϕ-coefficient, Yule's Q and Y, PS, etc. To avoid such complication, our synthetic data set for this experiment is restricted only to uncorrelated and positively correlated tables.

Using Definition 1, we can determine the consistency between every pair of measures by computing the correlation between their ranking vectors. Fig. 5 depicts the pair-wise correlation when various support bounds are imposed. We have re-ordered the correlation matrix using the reverse Cuthill-McKee algorithm [35] so that the darker cells are moved as close as possible to the main diagonal. The darker cells indicate that the correlation between the pair of measures is approximately greater than 0.8.

Initially, without support pruning, we observe that many of the highly correlated measures agree with the seven groups of measures identified in Section 4.3. For Groups 1–5, the pairwise correlations between measures from the same group are greater than 0.94. For Group 6, the correlation between interest factor and added value is 0.948; interest factor and K is 0.740; and K and added value is 0.873. For Group 7, the correlation between mutual information and κ is 0.936; mutual information and certainty factor is 0.790;

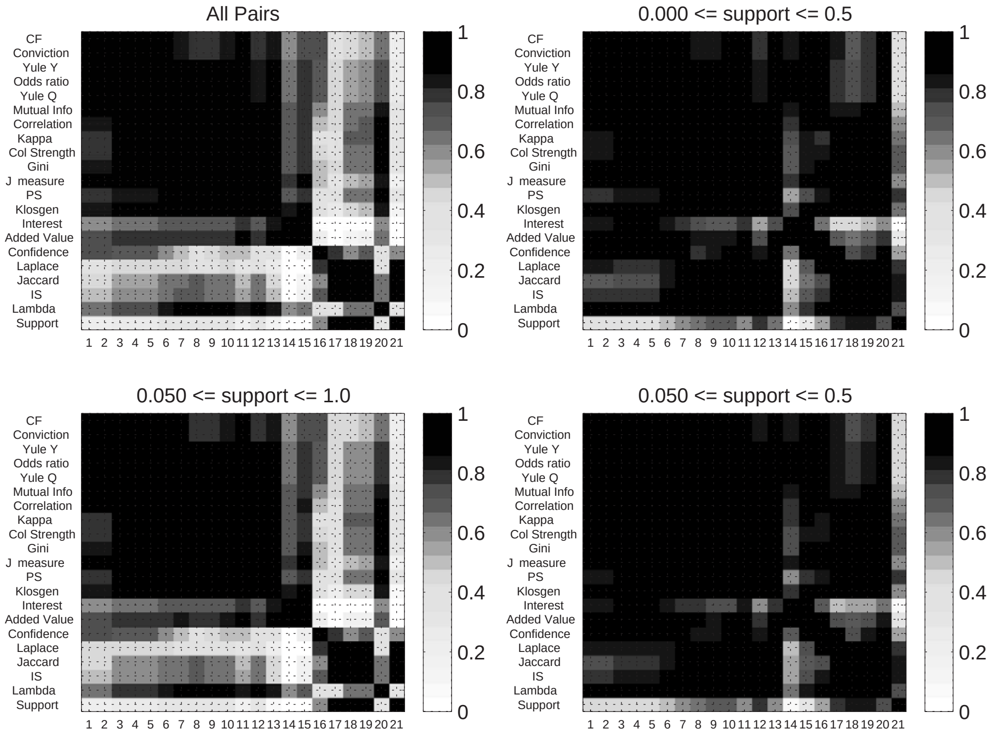


Fig. 5. Similarity between measures at various ranges of support values. Note that the column labels are the same as the row labels.

and κ and certainty factor is 0.747. These results suggest that the properties defined in Table 6 may explain most of the high correlations in the upper left-hand diagram shown in Fig. 5.

Next, we examine the effect of applying a maximum support threshold to the contingency tables. The result is shown in the upper right-hand diagram. Notice the growing region of dark cells compared to the previous case, indicating that more measures are becoming highly correlated with each other. Without support-based pruning, nearly 40% of the pairs have correlation above 0.85. With maximum support pruning, this percentage increases to more than 68%. For example, interest factor, which is quite inconsistent with almost all other measures except for added value, have become more consistent when high-support items are removed. This observation can be explained as an artifact of interest factor. Consider the contingency tables shown in Table 8, where A and B correspond to a pair of uncorrelated items, while C and D correspond to a pair of perfectly correlated items. However, because the support for item C is very high, $I(C,D) = 1.0112$, which is close to the value for statistical independence. By eliminating the high support items, we may resolve this type of inconsistency between interest factor and other objective measures.

Table 8
Effect of high-support items on interest factor

	B	$\overline{B}$	
A	100	200	300
$\overline{A}$	200	400	600
	300	600	900

(a) $I(A,B) = 1$,
$\phi(A,B) = 0$.

	D	$\overline{D}$	
C	890	0	890
$\overline{C}$	0	10	10
	890	10	900

(b) $I(C,D) = 1.012$,
$\phi(C,D) = 1$.

Our result also suggests that imposing a minimum support threshold does not seem to improve the consistency among measures. However, when it is used along with a maximum support threshold, the correlations among measures do show some slight improvements compared to applying the maximum support threshold alone—more than 71% of the pairs have correlation above 0.85. This analysis suggests that imposing a tighter bound on the support of association patterns may force many measures become highly correlated with each other.

6. Table standardization

Standardization is a widely used technique in statistics, political science, and social science studies to handle contingency tables that have different marginals. Mosteller suggested that standardization is needed to get a better idea of the underlying association between variables [3], by transforming an existing table so that their marginals become equal, i.e., $f^*_{1+} = f^*_{0+} = f^*_{+1} = f^*_{+0} = N/2$ (see Table 9). A standardized table is useful because it provides a visual depiction of how the joint distribution of two variables would look like after eliminating biases due to non-uniform marginals.

6.1. *Effect of non-uniform marginals*

Standardization is important because some measures can be affected by differences in the marginal totals. To illustrate this point, consider a pair of contingency tables, $X = [a\ b;\ c\ d]$ and $Y = [p\ q;\ r\ s]$. We can compute the difference between the ϕ-coefficients for both tables as follows.

$$\log(\phi_X) = \log(ad - bc) - \tfrac{1}{2}[\log(a+b) + \log(a+c) + \log(b+c) + \log(b+d)], \quad (1)$$

$$\log(\phi_Y) = \log(pq - rs) - \tfrac{1}{2}[\log(p+q) + \log(p+r) + \log(q+s) + \log(r+s)], \quad (2)$$

where the ϕ-coefficient is expressed as a logarithmic value to simplify the calculations. The difference between the two coefficients can be written as

$$\log(\phi_X) - \log(\phi_Y) = \Delta_1 - 0.5\Delta_2,$$

where

$$\Delta_1 = \log(ad - bc) - \log(pq - rs)$$

and

$$\Delta_2 = \log(a+b)(a+c)(b+c)(b+d) - \log(p+q)(p+r)(q+s)(r+s).$$

If the marginal totals for both tables are identical, then any observed difference between $\log(\phi_X)$ and $\log(\phi_Y)$ comes from the first term, Δ_1. Conversely, if the marginals are not identical, then the observed difference in ϕ can be caused by either Δ_1, Δ_2, or both.

The problem of non-uniform marginals is somewhat analogous to using accuracy for evaluating the performance of classification models. If a data set contains 99% examples of class 0 and 1% examples of class 1, then a classifier that produces models that classify every test example to be class 0 would have a high accuracy, despite performing miserably on class 1 examples. Thus, accuracy is not a reliable measure because it can be easily obscured by differences in the class distribution. One way to overcome this problem is by stratifying the data set so that both classes have equal representation during model building. A similar "stratification" strategy can be used to handle contingency tables with non-uniform support, i.e., by standardizing the frequency counts of a contingency table.

Table 9
Table standardization

$$\begin{array}{c|c|c|c} & B & \overline{B} & \\ \hline A & f_{11} & f_{10} & f_{1+} \\ \hline \overline{A} & f_{01} & f_{00} & f_{0+} \\ \hline & f_{+1} & f_{+0} & N \end{array} \longrightarrow \begin{array}{c|c|c|c} & B & \overline{B} & \\ \hline A & f^*_{11} & f^*_{10} & f^*_{1+} \\ \hline \overline{A} & f^*_{01} & f^*_{00} & f^*_{0+} \\ \hline & f^*_{+1} & f^*_{+0} & N \end{array} \longrightarrow \begin{array}{c|c|c|c} & B & \overline{B} & \\ \hline A & x & N/2 - x & N/2 \\ \hline \overline{A} & N/2 - x & x & N/2 \\ \hline & N/2 & N/2 & N \end{array}$$

6.2. IPF standardization

Mosteller presented the following iterative standardization procedure, which is called the Iterative Proportional Fitting algorithm or IPF [4], for adjusting the cell frequencies of a table until the desired margins, f^*_{i+} and f^*_{+j}, are obtained:

Row scaling:

$$f^{(k)}_{ij} = f^{(k-1)}_{ij} \times \frac{f^*_{i+}}{f^{(k-1)}_{i+}}, \tag{3}$$

Column scaling:

$$f^{(k+1)}_{ij} = f^{(k)}_{ij} \times \frac{f^*_{+j}}{f^{(k)}_{+j}}. \tag{4}$$

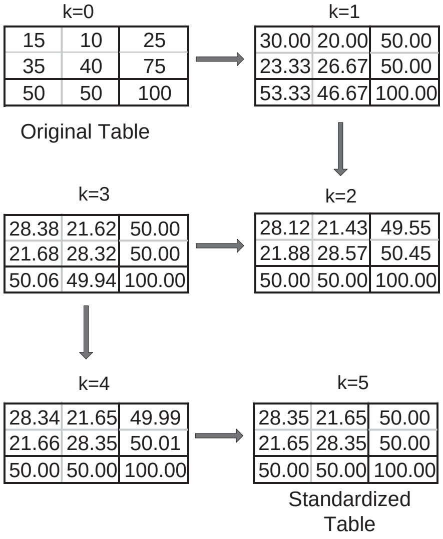
An example of the IPF standardization procedure is demonstrated in Fig. 6.

Theorem 1. *The IPF standardization procedure is equivalent to multiplying the contingency matrix* $\mathbf{M} = [a\ b;\ c\ d]$ *with*

$$\begin{bmatrix} k_1 & 0 \\ 0 & k_2 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} k_3 & 0 \\ 0 & k_4 \end{bmatrix},$$

where k_1, k_2, k_3 and k_4 are products of the row and column scaling factors.

Proof. The following lemma is needed to prove the above theorem.



k=0

15	10	25
35	40	75
50	50	100

Original Table

k=1

30.00	20.00	50.00
23.33	26.67	50.00
53.33	46.67	100.00

k=2

28.12	21.43	49.55
21.88	28.57	50.45
50.00	50.00	100.00

k=3

28.38	21.62	50.00
21.68	28.32	50.00
50.06	49.94	100.00

k=4

28.34	21.65	49.99
21.66	28.35	50.01
50.00	50.00	100.00

k=5

28.35	21.65	50.00
21.65	28.35	50.00
50.00	50.00	100.00

Standardized Table

Fig. 6. Example of IPF standardization.

Lemma 1. *The product of two diagonal matrices is also a diagonal matrix.*

This lemma can be proved in the following way. Let $M_1 = [f_1\ 0;\ 0\ f_2]$ and $M_2 = [f_3\ 0;\ 0\ f_4]$. Then, $M_1 \times M_2 = [(f_1 f_3)\ 0;\ 0\ (f_2 f_4)]$, which is also a diagonal matrix.

To prove Theorem 1, we also need to use Definition 2, which states that scaling the row and column elements of a contingency table is equivalent to multiplying the contingency matrix by a scaling matrix $[k_1\ 0;\ 0\ k_2]$. For IPF, during the kth iteration, the rows are scaled by $f^*_{i+}/f^{(k-1)}_{i+}$, which is equivalent to multiplying the matrix by $[f^*_{1+}/f^{(k-1)}_{1+}\ 0;\ 0\ f^*_{0+}/f^{(k-1)}_{0+}]$ on the left. Meanwhile, during the $(k+1)$th iteration, the columns are scaled by $f^*_{+j}/f^{(k)}_{+j}$, which is equivalent to multiplying the matrix by $[f^*_{+1}/f^{(k)}_{+1}\ 0;\ 0\ f^*_{+0}/f^{(k)}_{+0}]$ on the right. Using Lemma 1, we can show that the result of multiplying the row and column scaling matrices is equivalent to

$$\begin{bmatrix} f^*_{1+}/f^{(m)}_{1+} \cdots f^*_{1+}/f^{(0)}_{1+} & 0 \\ 0 & f^*_{0+}/f^{(m)}_{0+} \cdots f^*_{0+}/f^{(0)}_{0+} \end{bmatrix} \\ \times \begin{bmatrix} a & b \\ c & d \end{bmatrix} \\ \times \begin{bmatrix} f^*_{+1}/f^{(m+1)}_{+1} \cdots f^*_{+1}/f^{(1)}_{+1} & 0 \\ 0 & f^*_{+0}/f^{(m+1)}_{+0} \cdots f^*_{+0}/f^{(1)}_{+0} \end{bmatrix}$$

thus, proving Theorem 1.

The above theorem also suggests that the iterative steps of IPF can be replaced by a single matrix multiplication operation if the scaling factors k_1, k_2, k_3 and k_4 are known. In Section 6, we will provide a non-iterative solution for k_1, k_2, k_3 and k_4.

6.3. Consistency of measures under table standardization

Interestingly, the consequence of doing standardization goes beyond ensuring uniform margins in a contingency table. More importantly, if we apply different measures from Table 5 on the standardized, positively correlated tables, their rankings become identical. To the best of our knowledge, this fact has not been observed by anyone else before. As an illustration, Table 10 shows the results of ranking the standardized contingency tables for each example given in Table 3. Observe that the rankings are identical for all the measures. This observation can be explained in the following way. After standardization, the contingency matrix has the following form $[x\ y;\ y\ x]$, where $x = f^*_{11}$ and $y = N/2 - x$. The rankings are the same because many measures of association (specifically, all 21 considered in this paper) are monotonically increasing functions of x when applied to the standardized, positively correlated tables. We illustrate this with the following example.

Example 1. The ϕ-coefficient of a standardized table is

$$\phi = \frac{x^2 - (N/2 - x)^2}{(N/2)^2} = \frac{4x}{N} - 1. \tag{5}$$

For a fixed N, ϕ is a monotonically increasing function of x. Similarly, we can show that other measures such as α, I, IS, PS, etc., are also monotonically increasing functions of x.

The only exceptions to this are λ, Gini index, mutual information, J-measure, and Klosgen's K, which are convex functions of x. Nevertheless, these measures are monotonically increasing when we consider only the values of x between $N/4$ and $N/2$, which correspond to non-negatively correlated tables. Since the examples given in Table 3 are positively correlated, all 21 measures given in this paper produce identical ordering for their standardized tables.

6.4. Generalized standardization procedure

Since each iterative step in IPF corresponds to either a row or column scaling operation, odds ratio is preserved throughout the transformation (Table 6). In other words, the final rankings on the standardized tables for any measure are consistent with the rankings produced by odds ratio on the original tables. For this reason, a casual observer may think that odds ratio is perhaps the best measure to use. This is not true because there are other ways to standardize a contingency table. To illustrate other standardization schemes, we first show how to obtain the exact solutions for f^*_{ij}s using a direct approach. If we fix the standardized table to have equal margins, this forces the f^*_{ij}s to satisfy the following equations:

$$f^*_{11} = f^*_{00};\quad f^*_{10} = f^*_{01};\quad f^*_{11} + f^*_{10} = N/2. \tag{6}$$

Since there are only three equations in (6), we have the freedom of choosing a fourth equation that will provide a unique solution to the table standardization problem. In Mosteller's approach, the fourth equation is used to ensure that the odds ratio of the original table is the same as the odds ratio of the standardized table. This leads to the following conservation equation:

$$\frac{f_{11}f_{00}}{f_{10}f_{01}} = \frac{f^*_{11}f^*_{00}}{f^*_{10}f^*_{01}}. \tag{7}$$

Table 10
Rankings of contingency tables after IPF standardization

Example	ϕ	λ	α	Q	Y	κ	M	J	G	s	c	L	V	I	IS	PS	F	AV	S	ζ	K
E1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
E2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
E3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
E4	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
E5	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
E6	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
E7	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
E8	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
E9	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
E10	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6

After combining Eqs. (6) and (7), the following solutions are obtained:

$$f_{11}^{*} = f_{00}^{*} = \frac{N\sqrt{f_{11}f_{00}}}{2(\sqrt{f_{11}f_{00}} + \sqrt{f_{10}f_{01}})}, \tag{8}$$

$$f_{10}^{*} = f_{01}^{*} = \frac{N\sqrt{f_{10}f_{01}}}{2(\sqrt{f_{11}f_{00}} + \sqrt{f_{10}f_{01}})}. \tag{9}$$

The above analysis suggests the possibility of using other standardization schemes for preserving measures besides the odds ratio. For example, the fourth equation could be chosen to preserve the invariance of *IS* (cosine measure). This would lead to the following conservation equation:

$$\frac{f_{11}}{\sqrt{(f_{11}+f_{10})(f_{11}+f_{01})}} = \frac{f_{11}^{*}}{\sqrt{(f_{11}^{*}+f_{10}^{*})(f_{11}^{*}+f_{01}^{*})}}, \tag{10}$$

whose solutions are:

$$f_{11}^{*} = f_{00}^{*} = \frac{Nf_{11}}{2\sqrt{(f_{11}+f_{10})(f_{11}+f_{01})}}, \tag{11}$$

$$f_{10}^{*} = f_{01}^{*} = \frac{N}{2}\,\frac{\sqrt{(f_{11}+f_{10})(f_{11}+f_{01})} - f_{11}}{\sqrt{(f_{11}+f_{10})(f_{11}+f_{01})}}. \tag{12}$$

Thus, each standardization scheme is closely tied to a specific invariant measure. If IPF standardization is natural for a given application, then odds ratio is the right measure to use. In other applications, a standardization scheme that preserves some other measure may be more appropriate.

6.5. General solution of standardization procedure

In Theorem 1, we showed that the IPF procedure can be formulated in terms of a matrix multiplication operation. Furthermore, the left and right multiplication matrices are equivalent to scaling the row and column elements of the original matrix by some constant factors k_1, k_2, k_3 and k_4. Note that one of these factors is actually redundant; theorem 1 can be stated in terms of three parameters, k_1', k_2' and k_3', i.e.,

$$\begin{bmatrix} k_1 & 0 \\ 0 & k_2 \end{bmatrix}\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} k_3 & 0 \\ 0 & k_4 \end{bmatrix} = \begin{bmatrix} k_1' & 0 \\ 0 & k_2' \end{bmatrix}\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} k_3' & 0 \\ 0 & 1 \end{bmatrix}.$$

Suppose $M = [a\ b;\ c\ d]$ is the original contingency table and $M_s = [x\ y;\ y\ x]$ is the standardized table. We can show that any generalized standardization procedure can be expressed in terms of three basic operations: row scaling, column scaling, and addition of null values.[4]

$$\begin{bmatrix} k_1 & 0 \\ 0 & k_2 \end{bmatrix}\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} k_3 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & k_4 \end{bmatrix} = \begin{bmatrix} x & y \\ y & x \end{bmatrix}.$$

This matrix equation can be easily solved to obtain

$$k_1 = \frac{y}{b}, \quad k_2 = \frac{y^2 a}{xbc}, \quad k_3 = \frac{xb}{ay},$$
$$k_4 = x\left(1 - \frac{ad/bc}{x^2/y^2}\right).$$

For IPF, since $ad/bc = x^2/y^2$, therefore $k_4 = 0$, and the entire standardization procedure can be expressed in terms of row and column scaling operations.

7. Measure selection based on rankings by experts

Although the preceding sections describe two scenarios in which many of the measures become consistent with each other, such scenarios may not hold for all application domains. For example, support-based pruning may not be useful for domains containing nominal variables, while in other cases, one may not know the exact standardization scheme to follow. For such applications, an alternative approach is needed to find the best measure.

[4] Note that although the standardized table preserves the invariant measure, these intermediate steps of row or column scaling and addition of null values may not preserve the measure.

In this section, we describe a subjective approach for finding the right measure based on the relative rankings provided by domain experts. Ideally, we want the experts to rank all the contingency tables derived from the data. These rankings can help us identify the measure that is most consistent with the expectation of the experts. For example, we can compare the correlation between the rankings produced by the existing measures against the rankings provided by the experts and select the measure that produces the highest correlation.

Unfortunately, asking the experts to rank all the tables manually is often impractical. A more practical approach is to provide a smaller set of contingency tables to the experts for ranking and use this information to determine the most appropriate measure. To do this, we have to identify a small subset of contingency tables that optimizes the following criteria:

1. The subset must be small enough to allow domain experts to rank them manually. On the other hand, the subset must be large enough to ensure that choosing the best measure from the subset is almost equivalent to choosing the best measure when the rankings for all contingency tables are available.
2. The subset must be diverse enough to capture as much conflict of rankings as possible among the different measures.

The first criterion is usually determined by the experts because they are the ones who can decide the number of tables they are willing to rank. Therefore, the only criterion we can optimize algorithmically is the diversity of the subset. In this paper, we investigate two subset selection algorithms: RANDOM algorithm and DISJOINT algorithm.

RANDOM Algorithm. This algorithm randomly selects k of the N tables to be presented to the experts. We expect the RANDOM algorithm to work poorly when $k \ll N$. Nevertheless, the results obtained using this algorithm is still interesting because they can serve as a baseline reference.

DISJOINT Algorithm. This algorithm attempts to capture the diversity of the selected subset in terms of

1. Conflicts in the rankings produced by the existing measures. A contingency table whose rankings are $(1, 2, 3, 4, 5)$ according to five different measures have larger ranking conflicts compared to another table whose rankings are $(3, 2, 3, 2, 3)$. One way to capture the ranking conflicts is by computing the standard deviation of the ranking vector.
2. Range of rankings produced by the existing measures. Suppose there are five contingency tables whose rankings are given as follows.

Table t_1:	1	2	1	2	1
Table t_2:	10	11	10	11	10
Table t_3:	2000	2001	2000	2001	2000
Table t_4:	3090	3091	3090	3091	3090
Table t_5:	4000	4001	4000	4001	4000

The standard deviation of the rankings are identical for all the tables. If we are forced to choose three of the five tables, it is better to select t_1, t_3, and t_5 because they span a wide range of rankings. In other words, these tables are "furthest" apart in terms of their average rankings.

A high-level description of the algorithm is presented in Table 11. First, the algorithm computes the average and standard deviation of rankings for all the tables (step 2). It then adds the contingency table that has the largest amount of ranking conflicts into the result set Z (step 3). Next, the algorithm computes the "distance" between each pair of table in step 4. It then greedily tries to find k tables that are "furthest" apart according to their average rankings and produce the largest amount of ranking conflicts in terms of the standard deviation of their ranking vector (step 5a).

The DISJOINT algorithm can be quite expensive to implement because we need to compute the distance between all $(N \times (N-1))/2$ pairs of tables. To avoid this problem, we introduce an oversampling parameter, p, where $1 < p \ll \lceil N/k \rceil$, so that instead of sampling from the entire N

Table 11
The DISJOINT algorithm

```
Input: $T$: a set of $N$ contingency tables,
       $\mathbf{P}$: measures of association,
       $k$: the sample size,
       $p$: oversampling parameter

Output: $Z$: a set of $k$ contingency tables.

1. $T' \leftarrow$ randomly select $p \times k$ tables from $T$.
2. For each contingency table $t \in T'$,
    2a. $\forall M_i \in \mathbf{P}$, compute the rankings $O_{M_i}(t)$.
    2b. Compute mean and standard deviation of rankings:
        $\mu(t) = \sum_i O_{M_i}(t)/|\mathbf{P}|$
        $\sigma(t) = \sqrt{\sum_i (O_{M_i}(t) - \mu(t))^2/(|\mathbf{P}|-1)}$
3. $Z = \{t_m\}$ and $T' = T' - \{t_m\}$, where $t_m = \arg\max_t \sigma(t)$
4. For each $(t_i, t_j) \in T'$
    4a. $\forall M_k \in \mathbf{P}$, $\Delta_k(t_i,t_j) = O_{M_k}(t_i) - O_{M_k}(t_j)$
    4b. $\mu(t_i,t_j) = \sum_k \Delta_k(t_i,t_j)/|\mathbf{P}|$
    4c. $\sigma(t_i,t_j) = \sqrt{\sum_k (\mathbf{\Delta_k}(t_i,t_j) - \mu(t_i,t_j))^2/(|\mathbf{P}|-1)}$
    4d. $d(t_i,t_j) = \mu(t_i,t_j) + \sigma(t_i,t_j)$
5. while $|Z| < k$
    5a. Find $t \in T'$ that maximizes $\sum_j d(t,t_j)\ \forall t_j \in Z$
    5b. $Z = Z \cup \{t\}$ and $T' = T' - \{t\}$
```

tables, we select the k tables from a sub-population that contains only $k \times p$ tables. This reduces the complexity of the algorithm significantly to $(kp \times (kp-1))/2$ distance computations.

7.1. Experimental methodology

To evaluate the effectiveness of the subset selection algorithms, we use the approach shown in Fig. 7. Let T be the set of all contingency tables and S be the tables selected by a subset selection algorithm. Initially, we rank each contingency table according to all the available measures. The similarity between each pair of measure is then computed using Pearson's correlation coefficient. If the number of available measures is p, then a $p \times p$ similarity matrix will be created for each set, T and S. A good subset selection algorithm should minimize the difference between the similarity matrix computed from the subset, S_s, and the similarity matrix computed from the entire set of contingency tables, S_T. The following distance function is used to determine the difference between the two similarity matrices:

$$D(S_s, S_T) = \max_{i,j} |S_T(i,j) - S_s(i,j)|. \qquad (13)$$

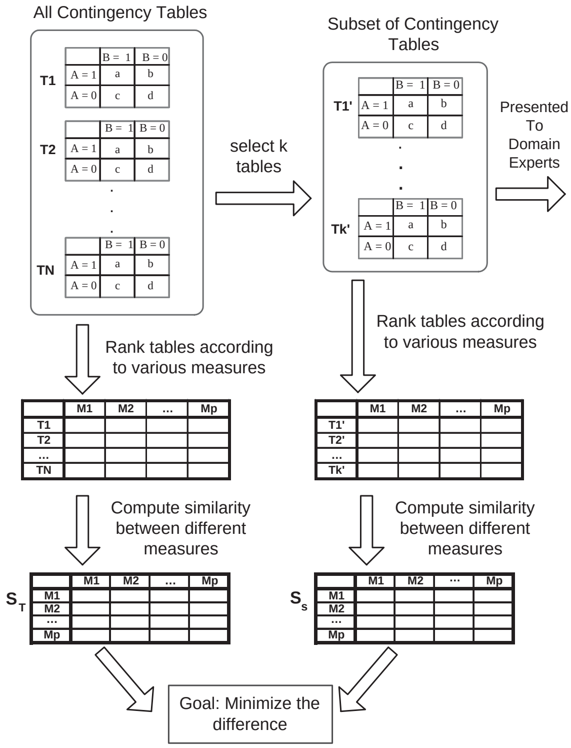

All Contingency Tables
T1
T2
TN
B = 1
B = 0
A = 1
A = 0
a
b
c
d
select k tables
Subset of Contingency Tables
T1'
Tk'
Presented To Domain Experts
Rank tables according to various measures
M1
M2
...
Mp
Compute similarity between different measures
S_T
S_s
Goal: Minimize the difference

Fig. 7. Evaluating the contingency tables selected by a subset selection algorithm.

If the distance is small, then we consider S as a good representative of the entire set of contingency tables T.

7.2. Experimental evaluation

We have conducted our experiments using the data sets shown in Table 12. For each data set, we randomly sample 100,000 pairs of binary items[5] as

Table 12
Data sets used in our experiments

Name	Description	Number of Variables
re0	Reuters-21578 articles	2886
la1	LA-times articles	31472
Product	Retail data	14462
S&P 500	Stock market data	976
E-Com	Web data	6664
Census	Survey data	59

[5] Only frequent items are considered, i.e., those with support greater than a user-specified minimum support threshold.

our initial set of contingency tables. We then apply the RANDOM and DISJOINT table selection algorithms on each data set and compare the distance function D at various sample sizes k. For each value of k, we repeat the procedure 20 times and compute the average distance D. Fig. 8 shows the relationships between the average distance D and sample size k for the re0 data set. As expected, our results indicate that the distance function D decreases with increasing sample size, mainly because the larger the sample size, the more similar it is to the entire data set. Furthermore, the DISJOINT algorithm does a substantially better job than random sampling in terms of choosing the right tables to be presented to the domain experts. This is because it tends to select

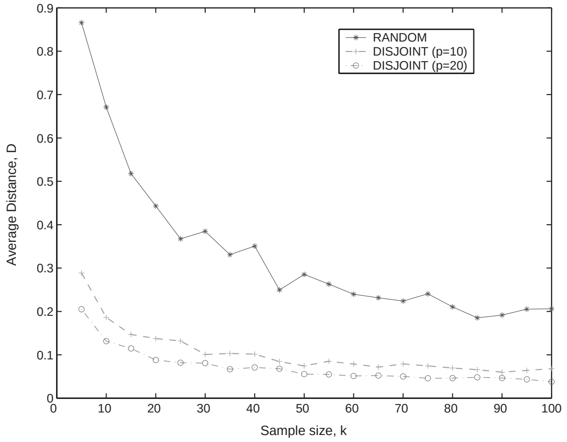


Fig. 8. Average distance between similarity matrix computed from the subset (S_s) and the similarity matrix computed from the entire set of contingency tables (S_T) for the re0 data set.

re0	Q	Y	κ	PS	F	AV	K	I	c	L	IS	ξ	s	S	λ	M	J	G	α	V
All tables	8	7	4	16	15	10	11	9	17	18	2	12	19	3	20	5	1	13	6	14
k=20	6	6	5	16	13	10	11	12	17	18	2	15	19	4	20	3	1	9	6	14

la1	Q	Y	κ	PS	F	AV	K	I	c	L	IS	ξ	s	S	λ	M	J	G	α	V
All tables	10	9	2	7	5	3	6	16	18	17	13	14	19	1	20	12	11	15	8	4
k=20	13	13	2	5	8	3	6	16	18	17	10	11	19	1	20	9	4	12	13	7

Product	Q	Y	κ	PS	F	AV	K	I	c	L	IS	ξ	s	S	λ	M	J	G	α	V
All tables	12	11	3	10	8	7	14	16	17	18	1	4	19	2	20	5	6	15	13	9
k=20	13	13	2	7	11	10	9	17	16	18	1	4	19	3	20	6	5	8	13	11

S&P500	Q	Y	κ	PS	F	AV	K	I	c	L	IS	ξ	s	S	λ	M	J	G	α	V
All tables	9	8	1	10	6	3	4	11	15	14	12	13	19	2	20	16	18	17	7	5
k=20	7	7	2	10	4	3	6	11	17	18	12	13	19	1	20	15	14	16	7	4

E-Com	Q	Y	κ	PS	F	AV	K	I	c	L	IS	ξ	s	S	λ	M	J	G	α	V
All tables	9	8	3	7	14	13	16	11	17	18	1	4	19	2	20	6	5	12	10	15
k=20	7	7	3	10	15	14	13	11	17	18	1	4	19	2	20	6	5	12	7	15

Census	Q	Y	κ	PS	F	AV	K	I	c	L	IS	ξ	s	S	λ	M	J	G	α	V
All tables	10	10	2	3	7	5	4	11	13	12	14	15	16	1	20	19	18	17	10	6
k=20	6	6	3	2	9	5	4	11	13	12	14	15	16	1	17	18	19	20	6	9

All tables: Rankings when all contingency tables are ordered.

k=20 : Rankings when 20 of the selected tables are ordered.

Fig. 9. Ordering of measures based on contingency tables selected by the DISJOINT algorithm.

tables that are furthest apart in terms of their relative rankings and tables that create a huge amount of ranking conflicts. Even at $k = 20$, there is little difference ($D < 0.15$) between the similarity matrices S_s and S_T.

We complement our evaluation above by showing that the ordering of measures produced by the DISJOINT algorithm on even a small sample of 20 tables is quite consistent with the ordering of measures if the entire tables are ranked by the domain experts. To do this, we assume that the rankings provided by the experts is identical to the rankings produced by one of the measures, say, the ϕ-coefficient. Next, we remove ϕ from the set of measures M considered by the DISJOINT algorithm and repeat the experiments above with $k = 20$ and $p = 10$. We compare the best measure selected by our algorithm against the best measure selected when the entire set of contingency tables is available. The results are depicted in Fig. 9. In nearly all cases, the difference in the ranking of a measure between the two (all tables versus a sample of 20 tables) is 0 or 1.

8. Conclusions

This paper presents several key properties for analyzing and comparing the various objective measures developed in the statistics, social science, machine learning, and data mining literature. Due to differences in some of their properties, a significant number of these measures may provide conflicting information about the interestingness of a pattern. However, we show that there are two situations in which the measures may become consistent with each other, namely, when support-based pruning or table standardization are used. We also show another advantage of using support in terms of eliminating uncorrelated and poorly correlated patterns. Finally, we develop an algorithm for selecting a small set of tables such that an expert can find a suitable measure by looking at just this small set of tables.

For future work, we plan to extend the analysis beyond two-way relationships. Only a handful of the measures shown in Table 5 (such as support, interest factor, and PS measure) can be generalized to multi-way relationships. Analyzing such relationships is much more cumbersome because the number of cells in a contingency table grows exponentially with k. New properties may also be needed to capture the utility of an objective measure in terms of analyzing k-way contingency tables. This is because a good objective measure must be able to distinguish between the direct association among k variables from their partial associations. More research is also needed to derive additional properties that can distinguish between some of the similar measures shown in Table 7. In addition, new properties or measures may be needed to analyze the relationship between variables of different types. A common approach for doing this is to transform one of the variables into the same type as the other. For example, given a pair of variables, consisting of one continuous and one categorical variable, we can discretize the continuous variable and map each interval into a discrete variable before applying an objective measure. In doing so, we may lose information about the relative ordering among the discretized intervals.

References

[1] R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases, in: Proceedings of 1993 ACM-SIGMOD International Conference on Management of Data, Washington, DC, May 1993, pp. 207–216.

[2] R. Agrawal, T. Imielinski, A. Swami, Database mining: a performance perspective, IEEE Trans. Knowledge Data Eng. 5 (6) (1993) 914–925.

[3] F. Mosteller, Association and estimation in contingency tables, J. Am. Stat. Assoc. 63 (1968) 1–28.

[4] A. Agresti, Categorical Data Analysis, Wiley, New York, 1990.

[5] G. Piatetsky-Shapiro, Discovery, analysis and presentation of strong rules, in: G. Piatetsky-Shapiro, W. Frawley (Eds.), Knowledge Discovery in Databases, MIT Press, Cambridge, MA, 1991, pp. 229–248.

[6] R.J. Hilderman, H.J. Hamilton, B. Barber, Ranking the interestingness of summaries from data mining systems, in: Proceedings of the 12th International Florida Artificial Intelligence Research Symposium (FLAIRS'99), Orlando, FL, May 1999, pp. 100–106.

[7] R.J. Hilderman, H.J. Hamilton, Knowledge Discovery and Measures of Interest, Kluwer Academic Publishers, Norwell, MA, 2001.

[8] I. Kononenko, On biases in estimating multi-valued attributes, in: Proceedings of the Fourteenth International

Joint Conference on Artificial Intelligence (IJCAI'95), Montreal, Canada, 1995, pp. 1034–1040.

[9] R. Bayardo, R. Agrawal, Mining the most interesting rules, in: Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining, San Diego, CA, August 1999, pp. 145–154.

[10] M. Gavrilov, D. Anguelov, P. Indyk, R. Motwani, Mining the stock market: which measure is best? in: Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining, Boston, MA, 2000.

[11] Y. Zhao, G. Karypis, Criterion functions for document clustering: experiments and analysis. Technical Report TR01-40, Department of Computer Science, University of Minnesota, 2001.

[12] L.A. Goodman, W.H. Kruskal, Measures of association for cross-classifications, J. Am. Stat. Assoc. 49 (1968) 732–764.

[13] G.U. Yule, On the association of attributes in statistics, Philos. Trans. R. Soc. A 194 (1900) 257–319.

[14] G.U. Yule, On the methods of measuring association between two attributes, J. R. Stat. Soc. 75 (1912) 579–642.

[15] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1960) 37–46.

[16] T. Cover, J. Thomas, Elements of Information Theory, New York, Wiley, 1991.

[17] P. Smyth, R.M. Goodman, Rule induction using information theory, in: Gregory Piatetsky-Shapiro, William Frawley (Eds.), Knowledge Discovery in Databases, MIT Press, Cambridge, MA, 1991, pp. 159–176.

[18] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Chapman & Hall, New York, 1984.

[19] R. Agrawal, R. Srikant, Fast algorithms for mining association rules in large databases, in: Proceedings of the 20th VLDB Conference, Santiago, Chile, September 1994, pp. 487–499.

[20] P. Clark, R. Boswell, Rule induction with cn2: some recent improvements, in: Proceedings of the European Working Session on Learning EWSL-91, Porto, Portugal, 1991, pp. 151–163.

[21] S. Brin, R. Motwani, J. Ullman, S. Tsur, Dynamic itemset counting and implication rules for market basket data, in: Proceedings of 1997 ACM-SIGMOD International Conference on Management of Data, Montreal, Canada, June 1997, pp. 255–264.

[22] S. Brin, R. Motwani, C. Silverstein, Beyond market baskets: generalizing association rules to correlations, in: Proceedings of 1997 ACM-SIGMOD International Conference on Management of Data, Tucson, Arizona, June 1997, pp. 255–264.

[23] C. Silverstein, S. Brin, R. Motwani, Beyond market baskets: generalizing association rules to dependence rules, Data Mining Knowledge Discovery 2 (1) (1998) 39–68.

[24] T. Brijs, G. Swinnen, K. Vanhoof, G. Wets, Using association rules for product assortment decisions: a case study, in: Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining, San Diego, CA, August 1999, pp. 254–260.

[25] C. Clifton, R. Cooley, Topcat: data mining for topic identification in a text corpus, in: Proceedings of the 3rd European Conference of Principles and Practice of Knowledge Discovery in Databases, Prague, Czech Republic, September 1999, pp. 174–183.

[26] W. DuMouchel, D. Pregibon, Empirical bayes screening for multi-item associations, in: Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining, 2001, pp. 67–76.

[27] E. Shortliffe, B. Buchanan, A model of inexact reasoning in medicine, Math. Biosci. 23 (1975) 351–379.

[28] C.C. Aggarwal, P.S. Yu, A new framework for itemset generation, in: Proceedings of the 17th Symposium on Principles of Database Systems, Seattle, WA, June 1998, pp. 18–24.

[29] S. Sahar, Y. Mansour, An empirical evaluation of objective interestingness criteria, in: SPIE Conference on Data Mining and Knowledge Discovery, Orlando, FL, April 1999, pp. 63–74.

[30] P.N. Tan, V. Kumar, Interestingness measures for association patterns: a perspective, in: KDD 2000 Workshop on Postprocessing in Machine Learning and Data Mining, Boston, MA, August 2000.

[31] C.J. van Rijsbergen, Information Retrieval, 2nd Edition, Butterworths, London, 1979.

[32] W. Klosgen, Problems for knowledge discovery in databases and their treatment in the statistics interpreter explora, Int. J. Intell. Systems 7 (7) (1992) 649–673.

[33] M. Kamber, R. Shinghal, Evaluating the interestingness of characteristic rules, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, 1996, pp. 263–266.

[34] D. Hand, H. Mannila, P. Smyth, Principles of Data Mining, MIT Press, Cambridge, MA, 2001.

[35] A. George, W.H. Liu, Computer Solution of Large Sparse Positive Definite Systems, Series in Computational Mathematics, Prentice-Hall, Englewood Cliffs, NJ, 1981.