Selecting the Right Interestingness Measure for Association Patterns

Pang-Ning Tan
Department of Computer Science and Engineering
University of Minnesota
200 Union Street SE
Minneapolis, MN 55455
ptan@cs.umn.edu

Vipin Kumar
Department of Computer Science and Engineering
University of Minnesota
200 Union Street SE
Minneapolis, MN 55455
kumar@cs.umn.edu

Jaideep Srivastava
Department of Computer Science and Engineering
University of Minnesota
200 Union Street SE
Minneapolis, MN 55455
srivasta@cs.umn.edu

ABSTRACT

Many techniques for association rule mining and feature selection require a suitable metric to capture the dependencies among variables in a data set. For example, metrics such as support, confidence, lift, correlation, and collective strength are often used to determine the interestingness of association patterns. However, many such measures provide conflicting information about the interestingness of a pattern, and the best metric to use for a given application domain is rarely known. In this paper, we present an overview of various measures proposed in the statistics, machine learning and data mining literature. We describe several key properties one should examine in order to select the right measure for a given application domain. A comparative study of these properties is made using twenty one of the existing measures. We show that each measure has different properties which make them useful for some application domains, but not for others. We also present two scenarios in which most of the existing measures agree with each other, namely, support-based pruning and table standardization. Finally, we present an algorithm to select a small set of tables such that an expert can select a desirable measure by looking at just this small set of tables.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*

Keywords

Interestingness Measure, Contingency tables, Associations

1. INTRODUCTION

The analysis of relationships among variables is a fundamental task at the heart of many data mining problems. For instance, the central task of association rule mining [2] is to find sets of binary variables that *co-occur* together frequently in a transaction database, while the goal of feature selection problems is to identify groups of variables that are strongly *correlated* with each other or with a specific target variable. Regardless of how the relationships are defined, such analysis often requires a suitable metric to capture the dependencies among variables. For example, metrics such as support, confidence, lift, correlation, and collective strength have been used extensively to evaluate the interestingness of association patterns [9, 14, 1, 15, 11]. These metrics are defined in terms of the frequency counts tabulated in a 2×2 contingency table as shown in Table 1. Unfortunately, many such metrics provide conflicting information about the interestingness of a pattern, and the best metric to use for a given application domain is rarely known.

Table 1: A 2×2 contingency table for variables 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: Example of contingency tables.

Example	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

In this paper, we show that not all measures are equally good at capturing the dependencies among variables. Furthermore, there is no measure that is consistently better than others in all application domains. This is because each measure has its own selection bias that justifies the rationale for preferring a set of tables over another. To illustrate this, consider the ten example contingency tables, E1 - E10,

Table 3: Rankings of contingency tables using various interestingness measures.

Example	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

given in Table 2. We compute the association in each example by using several well-known measures such as the ϕ-coefficient[1], interest factor, mutual information, J-Measure, *etc.* (A complete list and definitions of these metrics are given in Table 5.) Each example is then ranked according to its measure in decreasing order of magnitude, as shown in Table 3. The results of this table indicate that different measures can lead to substantially different orderings of contingency tables. For example, $E10$ is ranked highest by the I measure but lowest according to the ϕ-coefficient, while $E3$ is ranked lowest by the AV measure but highest according to the IS measure. Thus, selecting the right measure can be a tricky problem because one has to recognize the intrinsic properties of the existing measures.

There are several properties that need to be considered when we analyze a measure. Some of these properties are well-known to the data mining community, while others, which are equally important, deserve more attention. One important property is the sensitivity of a measure to row and column scaling operations. We illustrate this with the following classic example by Mosteller [12]:

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)

The table above illustrates the relationship between the gender of a student and the grade obtained for a particular course. Table 4(b) is obtained by doubling the number of male students and multiplying the number of female students by a factor of 10. However, on average, the performance of male students for the particular course is no better than it was before, and the same applies to the female students. Mosteller concluded that 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 [12]. Yet, many intuitively appealing measures, such as ϕ, mutual information, gini index and cosine measure, are sensitive to scaling of rows and columns of the table. Although measures that are invariant to this operation do exist, *e.g.*, odds ratio, they have other properties that make them unsuitable for many applications.

[1]The ϕ-coefficient is analogous to Pearson's correlation coefficient for continuous variables

Nevertheless, there are situations in which many of the existing measures become consistent with each other. First, the measures may become highly correlated when support-based pruning is used. Support-based pruning also tends to eliminate uncorrelated and poorly correlated patterns. Second, after standardizing the contingency tables to have uniform margins[12, 3], many of the well-known measures become equivalent to each other.

If both situations do not hold, we can find the most appropriate measure by comparing how well each measure agrees with the expectations of domain experts. This would require the domain experts to manually rank all the patterns or contingency tables extracted from the data. However, we show that it is possible to select a small set of "well-separated" contingency tables such that finding the most appropriate measure using this small set of tables is almost equivalent to finding the best measure using the entire data set.

The problem of evaluating objective measures used by data mining algorithms has attracted considerable attention in recent years [7, 6, 10]. For example, Kononenko *et al.* [10] have examined the use of different *impurity functions* for top-down inductive decision trees while Hilderman *et al.* [7, 6] have conducted extensive studies on the behavior of various *diversity measures* for ranking data summaries generated by attribute-oriented generalization methods.

The specific contributions of this paper are:

- We present an overview of various measures proposed in the statistics, machine learning and data mining literature.

- We describe several key properties one should examine in order to select the right measure for a given application domain. A comparative study of these properties is made using twenty one of the existing measures.

- We present two scenarios in which most of the existing measures agree with each other, namely, support-based pruning and table standardization.

- We present an algorithm to select a small set of tables such that an expert can select a desirable measure by looking at just this small set of tables.

2. PRELIMINARIES

Let $T(D) = \{t_1, t_2, \cdots t_N\}$ denotes the set of patterns, represented as contingency tables, derived from the data set D and $\mathbf{P}$ is the set of measures available to an analyst. Given an interestingness measure $M \in \mathbf{P}$, we can compute the vector $M(T) = \{m_1, m_2, \cdots, m_N\}$, which corresponds to the values of M for each contingency table that belongs to

Table 5: Interestingness Measures for Association Patterns.

#	Measure	Formula
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)	$\max\big(P(A,B)\log(\frac{P(B\|A)}{P(B)}) + P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),$ $P(A,B)\log(\frac{P(A\|B)}{P(A)}) + P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})\big)$
9	Gini index (G)	$\max\big(P(A)[P(B\|A)^2 + P(\overline{B}\|A)^2] + P(\overline{A})[P(B\|\overline{A})^2 + P(\overline{B}\|\overline{A})^2]$ $-P(B)^2 - P(\overline{B})^2,$ $P(B)[P(A\|B)^2 + P(\overline{A}\|B)^2] + P(\overline{B})[P(A\|\overline{B})^2 + P(\overline{A}\|\overline{B})^2]$ $-P(A)^2 - P(\overline{A})^2\big)$
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B\|A), P(A\|B))$
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)	$\max\big(\frac{P(B\|A)-P(B)}{1-P(B)}, \frac{P(A\|B)-P(A)}{1-P(A)}\big)$
18	Added Value (AV)	$\max(P(B\|A) - P(B), P(A\|B) - P(A))$
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)	$\sqrt{P(A,B)}\max(P(B\|A) - P(B), P(A\|B) - P(A))$

$T(D)$. $M(T)$ can also be transformed into a ranking vector $O_M(T) = \{o_1, o_2, \cdots, o_N\}$, whose components correspond to the rank order of each interestingness value, m_i. With this representation, the similarity between any two measures M_1 and M_2 can be evaluated by a similarity measure between vectors $O_{M_1}(T)$ and $O_{M_2}(T)$. If the values within two ranking vectors are unique, one can show that Pearson's correlation, cosine measure and an inverse of the L_2-norm are monotonically related. For simplicity, we choose one of them, Pearson's correlation, as our similarity measure.

Definition 1. [**Similarity between Measures**] Two measures of association, M_1 and M_2, are similar to each other with respect to the data set D if the *correlation* between $O_{M_1}(T)$ and $O_{M_2}(T)$ is greater than or equal to some positive threshold t.

3. PROPERTIES OF A MEASURE

In this section, we describe several key properties of a measure. While some of these properties have been extensively investigated in the data mining literature [13, 8], others are not that well-known. A complete listing of the measures examined in this study is given in Table 5.

3.1 Desired Properties of a Measure

Piatetsky-Shapiro [13] 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 [8, 6]. Table 6 illustrates the extent to which each of the existing measure satisfies the above properties.

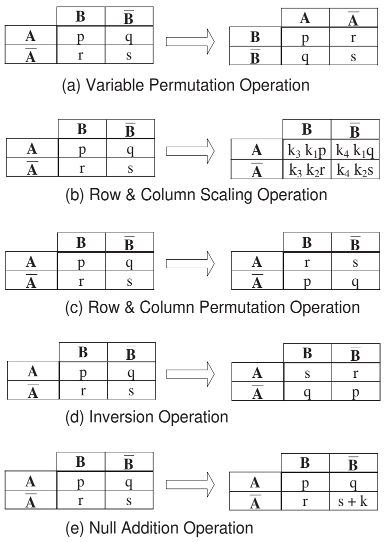
3.2 Other Properties of a Measure

There are other properties that deserve further investigation. These properties can be described using a matrix formulation. In this formulation, every 2×2 contingency table is represented as a contingency matrix, $\mathbf{M} = [\mathbf{f_{11} f_{10}; f_{01} f_{00}}]$ while every interestingness 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 matrix $\mathbf{M}$ 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 Figure 1.

Table 6: Properties of interestingness 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	$-1\cdots0\cdots1$	Yes	Yes	Yes	Yes	No	Yes	Yes	No
λ	Goodman-Kruskal's	$0\cdots1$	Yes	No	No	Yes	No	No*	Yes	No
α	odds ratio	$0\cdots1\cdots\infty$	Yes*	Yes	Yes	Yes	Yes	Yes*	Yes	No
Q	Yule's Q	$-1\cdots0\cdots1$	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Y	Yule's Y	$-1\cdots0\cdots1$	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
κ	Cohen's	$-1\cdots0\cdots1$	Yes	Yes	Yes	Yes	No	No	Yes	No
M	Mutual Information	$0\cdots1$	Yes	Yes	Yes	No**	No	No*	Yes	No
J	J-Measure	$0\cdots1$	Yes	No	No	No**	No	No	No	No
G	Gini index	$0\cdots1$	Yes	No	No	No**	No	No*	Yes	No
s	Support	$0\cdots1$	No	Yes	No	Yes	No	No	No	No
c	Confidence	$0\cdots1$	No	Yes	No	No**	No	No	No	Yes
L	Laplace	$0\cdots1$	No	Yes	No	No**	No	No	No	No
V	Conviction	$0.5\cdots1\cdots\infty$	No	Yes	No	No**	No	No	Yes	No
I	Interest	$0\cdots1\cdots\infty$	Yes*	Yes	Yes	Yes	No	No	No	No
IS	Cosine	$0\cdots\sqrt{P(A,B)}\cdots1$	No	Yes	Yes	Yes	No	No	No	Yes
PS	Piatetsky-Shapiro's	$-0.25\cdots0\cdots0.25$	Yes	Yes	Yes	Yes	No	Yes	Yes	No
F	Certainty factor	$-1\cdots0\cdots1$	Yes	Yes	Yes	No**	No	No	Yes	No
AV	Added value	$-0.5\cdots0\cdots1$	Yes	Yes	Yes	No**	No	No	No	No
S	Collective strength	$0\cdots1\cdots\infty$	No	Yes	Yes	Yes	No	Yes*	Yes	No
ζ	Jaccard	$0\cdots1$	No	Yes	Yes	Yes	No	No	No	Yes
K	Klosgen's	$(\frac{2}{\sqrt{3}}-1)^{1/2}[2-\sqrt{3}-\frac{1}{\sqrt{3}}]\cdots0\cdots\frac{2}{3\sqrt{3}}$	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 $\max(M(A,B),M(B,A))$.



(a) Variable Permutation Operation
(b) Row & Column Scaling Operation
(c) Row & Column Permutation Operation
(d) Inversion Operation
(e) Null Addition Operation

Figure 1: Operations on a contingency table.

Property 1. **[Symmetry Under Variable Permutation]** A measure O is symmetric under variable permutation (Figure 1(a)), $A \leftrightarrow B$, if $O(\mathbf{M}^T)=O(\mathbf{M})$ for all contingency matrices $\mathbf{M}$. 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's evaluation function. 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 $A \longrightarrow B$ from $B \longrightarrow A$. 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, 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 (Figure 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 [2] measure O is antisymmetric under

[2] A measure is normalized if its value ranges between -1 and

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 $\mathbf{M}$ (Figure 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 (Figure 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.

A	B	C	D	E	F
1	0	0	1	0	0
0	0	1	1	1	0
0	0	1	1	1	0
0	0	1	1	1	0
0	1	1	0	1	1
0	0	1	1	1	0
0	0	1	1	1	0
0	0	1	1	1	0
0	0	1	1	1	0
1	0	0	1	0	0
(a)		(b)		(c)	

Figure 2: Comparison between the ϕ-coefficients for 3 pairs of vectors. The ϕ values for (a), (b) and (c) are -0.1667, -0.1667 and 0.1667, respectively.

We illustrate the importance of inversion invariance with an example depicted in Figure 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 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 [5].

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 binary measure of association 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 Figure 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.

3.3 Summary

The discussion in this section suggests that there is no measure that is 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, one must match the desired properties of an application against the properties of the existing measures.

4. EFFECT OF SUPPORT-BASED PRUNING

Support is a widely-used measure in association rule mining because it represents the statistical significance of a pattern. Due to its anti-monotonicity property, the support measure has been used extensively to develop efficient algorithms for mining such patterns. We now describe two additional consequences of using the support measure.

4.1 Equivalence of Measures under Support Constraints

First, we show that many of the measures are highly correlated with each other under certain support constraints. To illustrate this, we randomly generated a synthetic data set that contains 10,000 contingency tables and ranked the tables according to all the available measures. Using Definition 1, we can compute the similarity between every pair of measures for the synthetic data set. Figure 3 depicts the pair-wise similarity when various support bounds are imposed. The dark cells indicate that the similarity, *i.e.*, correlation, between the two measures is greater than 0.85 while the lighter cells indicate otherwise. We have re-ordered the similarity matrix using the reverse Cuthill-McKee algorithm [4] so that the darker cells are moved as close as possible to the main diagonal. Our results show that by imposing a tighter bound on the support of the patterns, many of the measures become highly correlated with each other. This is shown by the growing region of dark cells as the support bounds are tightened. In fact, the majority of the pair-wise correlation between measures is greater than 0.85 when the support values are between 0.5% and 30% (the bottom-right figure), which is a quite reasonable range of support values for many practical domains.

4.2 Elimination of Poorly Correlated Tables using Support-based Pruning

Many association rule algorithms allow an analyst to specify a minimum support threshold to prune out the low-support patterns. Since the choice of minimum support

+1. An unnormalized measure U that ranges between 0 and $+\infty$ can be normalized via transformation functions such as $\frac{U-1}{U+1}$ or $\frac{\tan^{-1}\log(U)}{\phi/2}$.

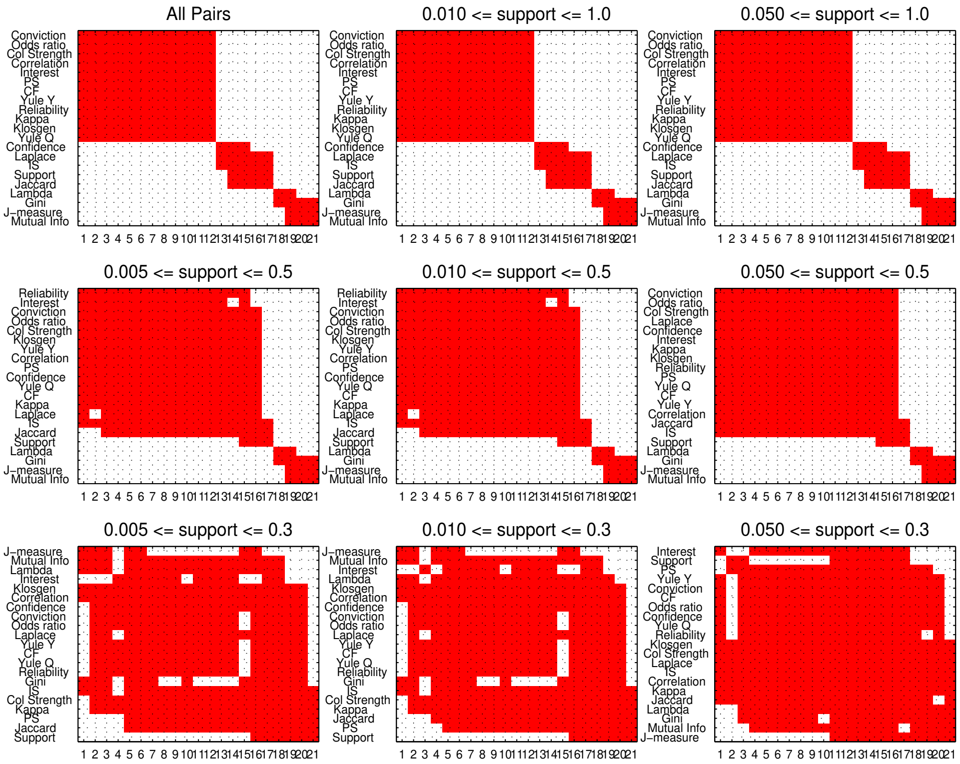


Figure 3: Similarity between measures at various ranges of support values. Note that the column labels are the same as the row labels.

threshold is somewhat arbitrary, we need to ensure that such a pruning strategy will not inadvertently remove many of the highly correlated patterns. To study the effect of support pruning, we examine the distribution of ϕ values for the contingency tables that are removed when various support thresholds are imposed on the synthetic data set. We use the ϕ-coefficient because it resembles Pearson's correlation coefficient for continuous variables. For this analysis, we impose the minimum support threshold on f_{11} and the maximum support threshold on both f_{1+} and f_{+1}. Without support constraints, the ϕ-coefficients for the entire tables are normally distributed around $\phi = 0$, as depicted in the top-left graph of figures 4(a) and (b). When a maximum support threshold is imposed, the ϕ values of the eliminated tables follow a bell-shaped distribution, as shown in figure 4(a). In other words, having a maximum support threshold will eliminate uncorrelated, positively correlated and negatively correlated tables at equal proportions.

On the other hand, if a lower bound of support is specified (Figure 4(b)), most of the contingency tables removed are either uncorrelated ($\phi = 0$) or negatively correlated ($\phi < 0$). This result makes sense because whenever a contingency table has a low support, the values of at least one of f_{10}, f_{01} or f_{00} must be relatively high to compensate for the low frequency count in f_{11}. This would correspond to poorly or negatively correlated contingency tables. The result is also consistent with the property $P2$ which states that a measure should increase as the support count increases.

Thus, support 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. TABLE STANDARDIZATION

Standardization is a widely-used technique in statistics, political science and social science studies to handle contingency tables that have non-uniform marginals. Mosteller suggested that standardization is needed to get a better idea of the underlying association between variables [12], by transforming an existing table so that their marginals are equal, *i.e.*, $f^*_{1+} = f^*_{0+} = f^*_{+1} = f^*_{+0} = N/2$ (see Table 7). 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.

Table 7: Table Standardization.

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

$\longrightarrow$

	B	$\overline{B}$	
A	f^*_{11}	f^*_{10}	f^*_{1+}
$\overline{A}$	f^*_{01}	f^*_{00}	f^*_{0+}
	f^*_{+1}	f^*_{+0}	N

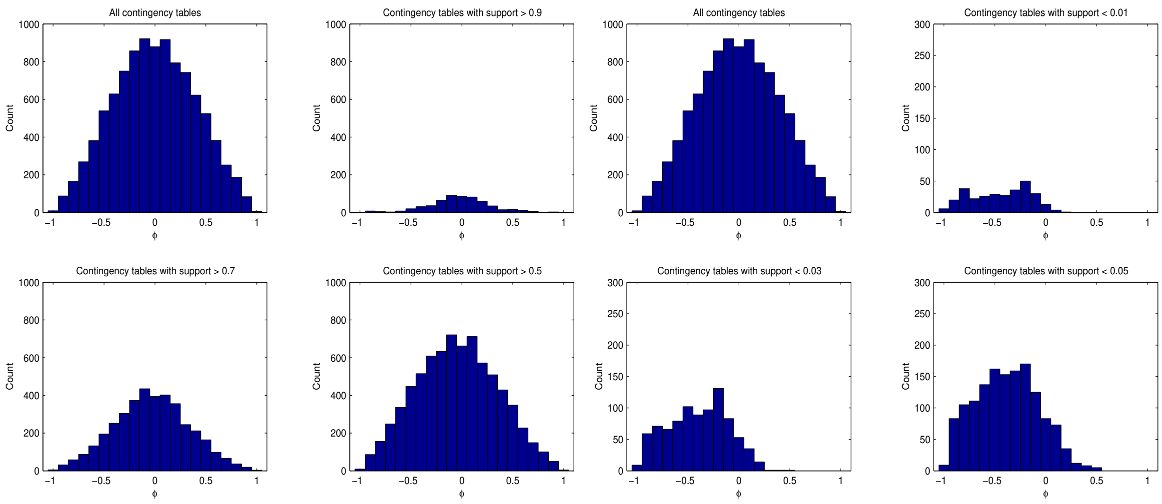
(a)

$\longrightarrow$

	B	$\overline{B}$	
A	x	$N/2 - x$	$N/2$
$\overline{A}$	$N/2 - x$	x	$N/2$
	$N/2$	$N/2$	N

(b)

Mosteller also gave the following iterative procedure, which is called the Iterative Proportional Fitting algorithm or IPF [3],



(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.

Figure 4: Effect of Support Pruning on Contingency tables.

for adjusting the cell frequencies of a table until the desired margins, f_{i+}^* and f_{+j}^*, are obtained:

$$\text{Row scaling :} \quad f_{ij}^{(k)} = f_{ij}^{(k-1)} \times \frac{f_{i+}^*}{f_{i+}^{(k-1)}} \tag{1}$$

$$\text{Column scaling :} \quad f_{ij}^{(k+1)} = f_{ij}^{(k)} \times \frac{f_{+j}^*}{f_{+j}^{(k)}} \tag{2}$$

An example of the IPF standardization procedure is demonstrated in Figure 5.

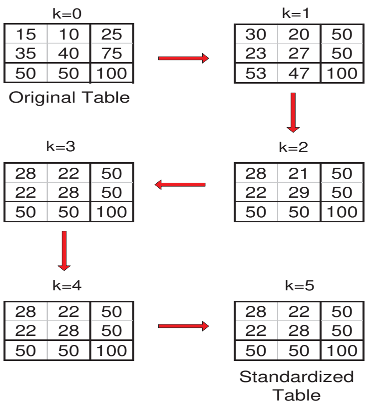


Figure 5: Example of IPF standardization.

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 8 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:

$$\phi = \frac{x^2 - (N/2 - x)^2}{(N/2)^2} = \frac{4x}{N} - 1 \tag{3}$$

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 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.

Note that 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

Table 8: 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

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:

$$f^*_{11} = f^*_{00}; \quad f^*_{10} = f^*_{01}; \quad f^*_{11} + f^*_{10} = N/2 \tag{4}$$

Since there are only three equations in (4), 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:

$$\frac{f_{11}f_{00}}{f_{10}f_{01}} = \frac{f^*_{11}f^*_{00}}{f^*_{10}f^*_{01}} \tag{5}$$

After combining equations 4 and 5, the following solutions are obtained:

$$f^*_{11} = f^*_{00} = \frac{N\sqrt{f_{11}f_{00}}}{2(\sqrt{f_{11}f_{00}} + \sqrt{f_{10}f_{01}})} \tag{6}$$

$$f^*_{10} = f^*_{01} = \frac{N\sqrt{f_{10}f_{01}}}{2(\sqrt{f_{11}f_{00}} + \sqrt{f_{10}f_{01}})} \tag{7}$$

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:

$$\frac{f_{11}}{\sqrt{(f_{11}+f_{10})(f_{11}+f_{01})}} = \frac{f^*_{11}}{\sqrt{(f^*_{11}+f^*_{10})(f^*_{11}+f^*_{01})}} \tag{8}$$

whose solutions are:

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

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

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. 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.

In this section, we describe a novel approach for finding the right measure based on the relative rankings provided by domain experts. Ideally, we would like the experts to rank all the contingency tables derived from the data. This would allow us to identify the most appropriate measure, consistent with the expectations of the experts. Since manual ordering of the contingency tables can be quite a laborious task, it is more desirable to provide a smaller set of contingency tables to the experts for ranking. We investigate two table selection algorithms in this paper:

- RANDOM: randomly select k out of the overall N tables and present them to the experts.
- DISJOINT: select k tables that are "furthest" apart according to their average rankings and would produce the largest amount of ranking conflicts, *i.e.*, large standard deviation in their ranking vector (see Table 9). A detailed explanation of this algorithm is given in [16].

Table 9: 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. $\sigma(t_i,t_j) = \sqrt{\sum_k (\Delta_{\mathbf{k}}(t_i,t_j) - \mu(t_i,t_j))^2/(|\mathbf{P}|-1)}$
   4d. $d(t_i,t_j) = \mu(t_i,t_j) + \sigma(t_i,t_j)$
5. while $|Z| < k$
   3a. Find $t \in T'$ that maximizes $\sum_j d(t,t_j)\ \forall t_j \in Z$
   3b. $Z = Z \cup \{t\}$ and $T' = T' - \{t\}$

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

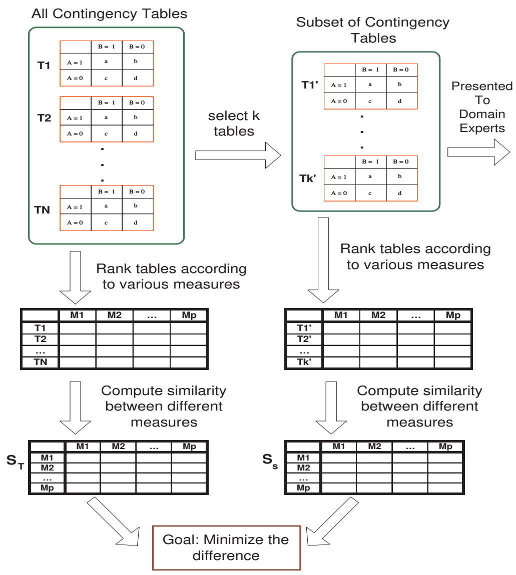


Figure 6: Sampling contingency tables.

To evaluate the effectiveness of our table selection algorithms, we use the approach shown in Figure 6. First, each contingency table is ranked according to the available measures. The similarity between various measures are then computed using Pearson's correlation. A good table selection scheme should minimize the difference between the similarity matrix computed from the samples, S_s, with the similarity matrix computed from the entire set of contingency tables, S_T. The following distance function is used to determine the difference between two similarity matrices:

$$D(S_s, S_T) = \max_{i,j} |S_T(i,j) - S_s(i,j)| \qquad (11)$$

We have conducted our experiments using the data sets shown in Table 10. For each data set, we randomly sample 100,000 pairs of binary variables[3] as 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. Figure 7 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 tables

[3]Only the frequent variables are considered, *i.e.*, those with support greater than a user-specified minimum support threshold.

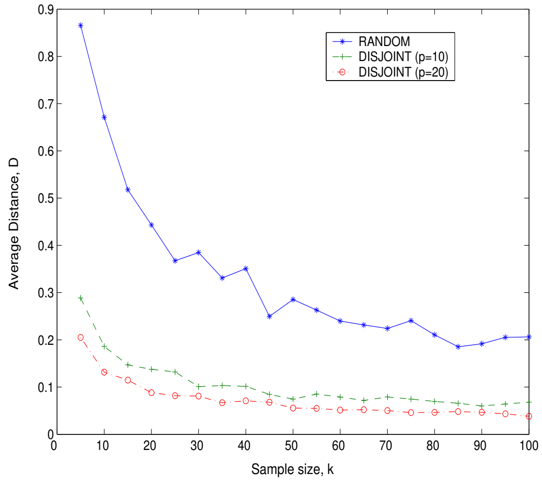


Figure 7: Average distance between similarity matrix computed from the samples (S_s) and the similarity matrix computed from the entire set of contingency tables (S_T) for the re0 data set.

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.

Table 10: 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

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 Figure 8. 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.

7. CONCLUSIONS

In this paper, we have described several key properties one should consider before deciding what is the right measure to use for a given application domain. We show that there is no measure that is consistently better than others in all cases. Nevertheless, there are situations in which many of these measures are highly correlated with each other, *e.g.*, when support-based pruning or table standardization are used. If both situations do not hold, one should select the best

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.

Figure 8: Ordering of measures based on contingency tables selected by the DISJOINT algorithm.

measure by matching the properties of the existing measures against the expectations of the domain experts. We have presented an algorithm to select a small set of tables such that an expert can find the most appropriate measure by looking at just this small set of tables.

This work can be extended to k-way contingency tables. However, understanding the underlying association within a k-way table requires techniques to decompose the overall association into partial associations between the constituent variables. Log-linear models provide a good alternative for doing this. More research is also needed to understand the association between variables of mixed data types. A standard way to do this is by transforming the variables into similar data types (*e.g.*, by discretizing continuous variables or reducing the multiple categorical levels into binary levels) before applying the appropriate measure of association.

8. ACKNOWLEDGMENTS

This work was partially supported by NSF grant # ACI−9982274, DOE contract # DOE/LLNL W-7405-ENG-48 and by 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.

9. REFERENCES

[1] C. Aggarwal and P. Yu. A new framework for itemset generation. In *Proc. of the 17th Symposium on Principles of Database Systems*, pages 18–24, Seattle, WA, June 1998.

[2] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. of 1993 ACM-SIGMOD Int. Conf. on Management of Data*, pages 207–216, Washington, D.C., May 1993.

[3] A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, 1990.

[4] A. George and W. Liu. *Computer Solution of Large Sparse Positive Definite Systems*. Prentice-Hall series in computational mathematics, 1981.

[5] D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, 2001.

[6] R. Hilderman and H. Hamilton. Evaluation of interestingness measures for ranking discovered knowledge. In *Proc. of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'01)*, pages 247–259, Hong Kong, April 2001.

[7] R. Hilderman, H. Hamilton, and B. Barber. Ranking the interestingness of summaries from data mining systems. In *Proc. of the 12th International Florida Artificial Intelligence Research Symposium (FLAIRS'99)*, pages 100–106, Orlando, FL, May 1999.

[8] M. Kamber and R. Shinghal. Evaluating the interestingness of characteristic rules. In *Proc. of the Second Int'l Conference on Knowledge Discovery and Data Mining*, pages 263–266, Portland, Oregon, 1996.

[9] M. Klemettinen, H. Mannila, P. Ronkainen, T. Toivonen, and A. Verkamo. Finding interesting rules from large sets of discovered association rules. In *Proc. of the 3rd Int'l Conf. on Information and Knowledge Management (CIKM'94).*, pages 401–407, Gaithersburg, Maryland, November 1994.

[10] I. Kononenko. On biases in estimating multi-valued attributes. In *Proc. of the Fourteenth Int'l Joint Conf. on Artificial Intelligence (IJCAI'95)*, pages 1034–1040, Montreal, Canada, 1995.

[11] B. Liu, W. Hsu, and Y. Ma. Pruning and summarizing the discovered associations. In *Proc. of the Fifth Int'l Conference on Knowledge Discovery and Data Mining*, pages 125–134, San Diego, CA, August 1999.

[12] F. Mosteller. Association and estimation in contingency tables. *Journal of the American Statistical Association*, 63:1–28, 1968.

[13] G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In G. Piatetsky-Shapiro and W. Frawley, editors, *Knowledge Discovery in Databases*, pages 2299–248. MIT Press, Cambridge, MA, 1991.

[14] A. Silberschatz and A. Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Eng.*, 8(6):970–974, 1996.

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

[16] P. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. Technical Report 2002-112, Army High Performance Computing Research Center, 2002.