Interestingness Measures for Association Patterns : A Perspective

Pang-Ning Tan[1] and Vipin Kumar[1]

Department of Computer Science,
University of Minnesota,
200 Union Street SE,
Minneapolis, MN 55455.
{ptan,kumar}@cs.umn.edu

Abstract. Association rules are valuable patterns because they offer useful insight into the types of dependencies that exist between attributes of a data set. Due to the completeness nature of algorithms for mining association-type patterns (such as Apriori), the number of patterns extracted are often very large. Therefore, there is a need to prune or rank the discovered patterns according to some measure of interestingness. In this paper, we will examine the various interestingness measures that arise from statistics, machine learning and data mining literature. We will investigate how close these measures reflect the statistical notion of correlation. We will show that support-based pruning is appropriate because it removes mostly uncorrelated and negatively correlated patterns. Another useful measure is the χ^2 statistic, which is often used to test whether there is sufficient evidence in the data samples to reject the hypothesis that items in a pattern are independent. Our experimental results verified that many of the intuitive measures (such as Piatetsky-Shapiro's rule-interest, confidence, laplace, entropy gain, etc.) are very similar in nature to correlation coefficient (in the region of support values typically encountered in practice). Finally, we will introduce a new metric, called the IS measure, and show that it is highly linear with respect to correlation coefficient for many interesting association patterns.

1 Introduction

Association rules [3, 2] are valuable patterns that can be derived from large database of transactions. They offer useful insight into the dependence relationships that exist among attributes of the data set. Conceptually, an association rule indicates that the presence of a set of items, or itemset, in a transaction would imply the occurence of other items in the same transaction.

The association rule discovery problem is often decomposed into two separate tasks : (1) to discover all itemsets having support above a user-defined threshold, and (2) to generate rules from these frequent itemsets. The first task can be very expensive, because it requires a lot of I/O operations. Over the years, many algorithms, such as Apriori and other level-wise algorithms [3, 2, 4, 17, 9, 1], have been developed to efficiently discover these itemsets.

On the other hand, rule generation is less I/O intensive. Nevertheless, there are two major problems with association rule generation : (1) too many rules are being generated (rule quantity problem), and (2) not all of the rules are interesting (rule quality problem). Both problems are not entirely independent. For example, knowledge about the quality of a rule can be used to reduce the number of rules presented to an analyst.

There has been various research effort aimed at mitigating both problems. The rule quantity problem can be handled by pruning or summarizing the discovered rules. Toivonen et al.[21] proposed the idea of using structural rule covers to remove redundant rules and clustering as a means for grouping (summarizing) together related rule covers. Liu et al. [14] used the standard χ^2 test to prune insignificant rules and introduced the concept of direction setting rules to summarize the discovered pattern. Other researchers adopt a completely different view on how to handle the rule quantity problem. Srikant et al. [20] and Ng et al. [16] used the constraints provided by a user to limit the number of rules that are produced.

Solution to the rule quality problem relies on specification of an interestingness measure to represent the novelty, utility or significance of a pattern. By ordering the discovered rules according to their degree of interestingness, one can ensure that only good quality rules are presented to an analyst. Some of the measures are applicable to itemsets as well as rules. [1] In such cases, they can be incorporated into the itemset generation step for early pruning of uninteresting itemsets.

In the original formulation of association rule discovery problem, support and confidence are two of the interestingness measures proposed. Support is necessary because it represents the statistical significance of a pattern. From the marketing perspective, support of an itemset in retail sales data justifies the feasibility of promoting the items together. Support is also good for pruning the search space since it possesses a nice downward closure property (Figure 1). Beyond that, it may not serve as a reliable interestingness measure. For example, rules with

[1] We will use the term association pattern to refer to both an association rule and the itemset from which the rule is generated.

high support quite often correspond to obvious knowledge about the domain. The rule $Bread \Longrightarrow Milk$, for instance, may not be interesting to a data analyst simply because it does not reveal any surprising information, despite gathering sufficiently high support. In Fig. 1, any itemsets that lie outside the frequent

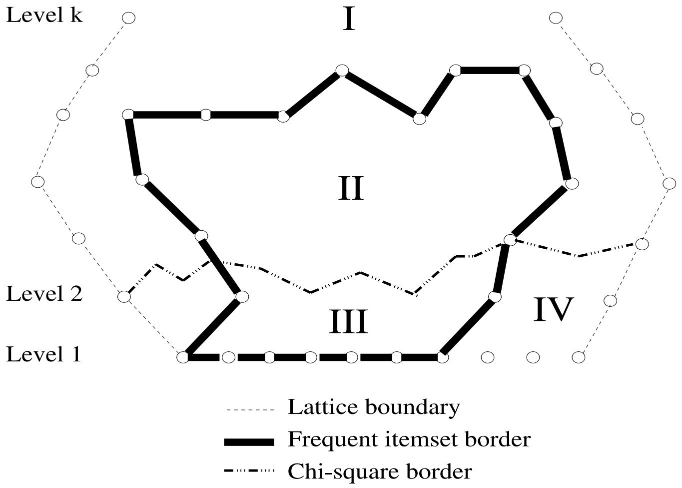

Fig. 1. Lattice structure for all itemsets. This structure can be divided into several regions : I. Infrequent and dependent itemsets, II. frequent and dependent itemsets, III. frequent and independent itemsets, IV. infrequent and independent itemsets.

itemset border can be ruled out as uninteresting. However, one still have to face the problem of combinatorial explosion due to the large number of frequent itemsets and rules that can be potentially generated.

In theory, confidence measures the conditional probability of events associated with a particular rule. For example, if a rule $X \longrightarrow Y$ has a confidence c, it means that $c\%$ of all transactions that contain X will also contain Y. Unfortunately, confidence can be misleading in many practical situations, as shown by Brin et al. in [7]. [7] offerred an alternative to evaluate the significance of association patterns using standard hypothesis testing techniques. Specifically, they proposed the use of χ^2 statistic to evaluate the dependence between items in an itemset. This measure is desirable because it will rule out itemsets that occur by chance. [7, 19] also showed that χ^2 possess an upward closure property that can be used to further prune the search space. This property will allow us to look for a border between dependent and independent itemsets (Fig. 1. However, we will argue that the alternative proposed in [7, 19] may still be unsatisfactory.

This paper intends to follow-up on the earlier work done by Brin et al in [7]. The main contributions of this paper are as follows :

1. We investigate the possibility of using various measures from statistics, machine learning and data mining literature to rank the association patterns.

2. We show that support-based pruning is useful for removing uncorrelated and negatively correlated itemsets.
3. We combine support-based pruning with χ^2 pruning to reduce the complexity of mining interesting association patterns. Specifically, we examine the applicability of various interestingness measures to region II of Fig. 1.
4. We introduce a new measure, called the IS measure, which takes into account both the interestingness and support of a pattern.
5. We evaluate how well the various interestingness measures can capture the notion of statistical correlation. In fact, our empirical results show that many of these measures are capable of representing statistical correlation within certain range of support values.

The rest of the paper is organized as follows. In section 2, we describe several statistical techniques for measuring the degree of association between variables. We then explore some of the interestingness measures that arise from data mining in section 3. Experimental results comparing the various metrics to standard statistical correlation are also presented. In Section 4, we look at how these measures can be applied to find interesting patterns in region II of Fig. 1. Finally, we conclude by summarizing our results in Section 5.

2 Statistical Measures of Dependency

Inferring dependencies between variables in data is a well-studied area in statistics. In this section, we will present several statistical methods for measuring the dependencies between categorical variables. The first class of methods, called the goodness of fit test, compares the actual distribution of a data set to its expected distribution under a null hypothesis. To test for item dependence, the null hypothesis assumes that a pattern consists of items that are independent of each other. Then, based on evidence provided by the observed data, one can determine whether to accept or reject the independence assumption. Though hypothesis testing is a well-established methodology in statistical inference, techniques that estimate directly the degree of dependence are often more desirable. These are called measures of association in statistical literature.

Our discussion will focus primarily on pairs of dichotomous variables, even though some of the techniques described here can be generalized to larger sized patterns. Let A and B denote a pair of binary variables. The dataset that contains these variables can be summarized into a 2×2 contingency table as shown in Table 1. Each cell represents the four possible combinations of A and B values. f_{ij} corresponds to the frequency (or support count) for each cell; while f_{i+} and f_{+j} are the marginal sums for row i and column j respectively. For example, $f_{i+} = f_{i1} + f_{i0}$. Also, N refers to the size of the database.

2.1 Goodness of fit test

Goodness of fit tests are often used for comparing probability distributions. For testing variable dependencies, a null hypothesis is initially proposed, stating that

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

Table 1. A 2×2 contingency table for binary variables.

the variables are independent of each other. A test procedure is then formulated to measure the discrepancy between the observed sample from its expected distribution under the null hypothesis. Pearson's χ^2 statistic is often used for this purpose :

$$\chi^2 = \sum_{j,k} \frac{(f_{jk} - E(f_{jk}))^2}{E(f_{jk})} \tag{1}$$

where the sum ranges over all the cells in the contingency table. The expected frequency of a particular cell is $E(f_{jk}) = N \times \frac{f_{j+}}{N} \times \frac{f_{+k}}{N}$. Briefly, χ^2 measures the normalized deviation between the observed frequencies, f_{jk}, from their expected frequencies, $E(f_{jk})$, under the null hypothesis. The larger the deviation, the more evidence we have to reject the independence hypothesis.

For the 2×2 contingency table shown in Table 1, its χ^2 value can be simplified into the following expression :

$$\begin{aligned}\chi^2 &= \frac{(f_{11} - f_{1+}f_{+1}/N)^2}{f_{1+}f_{+1}/N} + \frac{(f_{10} - f_{1+}f_{+0}/N)^2}{f_{1+}f_{+0}/N} + \frac{(f_{01} - f_{0+}f_{+1}/N)^2}{f_{0+}f_{+1}/N} \\ &\quad + \frac{(f_{00} - f_{0+}f_{+0}/N)^2}{f_{0+}f_{+0}/N} \\ &= \frac{N(f_{11}f_{00} - f_{01}f_{10})^2}{f_{1+}f_{0+}f_{+1}f_{+0}}\end{aligned} \tag{2}$$

Since this statistic is calculated using all the cells from the contingency table, its value can be large if the variables exhibit strong positive (or negative) correlation with each other. [7, 19] have used this property to find both positive and negatively correlated association patterns. They also showed that χ^2 is upward closed, a property can be exploited to prune the search space of the problem. However, this test may not be the ultimate answer due to the following reasons:

1. As stated in [7], χ^2 does not tell us the strength of correlation between items in an association pattern. Instead, it will only help us to decide whether items in the pattern are independent of each other. Thus, it cannot be used for ranking purposes.
2. The upward closure property of χ^2 ensures that all itemsets above the χ^2 border are statistically dependent. In reality, some itemsets above the χ^2 border will be more interesting than others. Therefore, just knowing the border alone is insufficient.

3. The χ^2 statistic depends on the total number of transactions. On the other hand, the χ^2 cutoff value depends only on the degrees of freedom of the attributes (which is 1 for binary attributes) and the significance level desired. For example, the rejection region for binary attributes at 0.05 significance level is 3.84. When the number of transactions are large, the cutoff value can be exceeded by a very large number of itemsets.

Another form of goodness of fit test is based on the likelihood ratio test. However, this approach requires specification of a parametric model in the order to compute the likelihood ratio. More details can be found in statistical references such as [13]. For large sample size, it can be shown that this test is equivalent to Pearson χ^2 test.

2.2 Measures of Association

Goodness of fit tests can only tell us whether a pattern contains independent items. For ranking purposes, a direct measure of association between variables is a preferable scheme. We will present three such measures in this section. They are Pearson's ϕ-coefficient, Goodman and Kruskal's λ coefficient and uncertainty measures from information theory.

In order to compare the similarity between these metrics, a series of experiments have been conducted using a synthetic dataset. This dataset contains 10000 randomly generated data points. Each data point is a 4-tuple $(f_{11}, f_{10}, f_{01}, f_{00})$, generated according to the following conditions :

1. $\frac{f_{11}}{N} < 1$, $\frac{f_{10}}{N} < 1$ and $\frac{f_{01}}{N} < 1$ and
2. $f_{11} + f_{10} + f_{01} < N$.

This data set will be used to compute the various interestingness measures introduced throughout this paper. We then choose one of the statistical measures of association as the reference metric, and plot the graphs of the remaining measures with respect to this reference metric.

2.2.1 Correlation coefficient Correlation coefficient measures the degree of linearity between a pair of random variables. Theoretically, it is defined as the covariance between two variables, divided by their standard deviations:

$$\rho_{AB} = \frac{\mathrm{Cov}(A,B)}{\sigma_A \sigma_B} \tag{3}$$

where $\mathrm{Cov}(A,B) = E(AB) - E(A)E(B)$. The range of ρ_{AB} is between -1 and $+1$. If the two variables are independent, then $\rho_{AB} = 0$. However, the converse is not necessarily true. It is possible that $\rho_{AB} = 0$ even when the variables are not independent. This occurs when both $E(AB)$ and $E(A)$ are zero, thus satisfying $\rho_{AB} = 0$ trivially. Fortunately, this is not a problem for binary variables. If $E(AB)$ and $E(A)$ are zero, then both $P(A,B)$ and $P(A,\overline{B})$ have to be zero, thus rendering A to be a meaningless attribute.

For binary variables, $\sigma_A = \sqrt{p(1-p)}$ where $p \equiv P(A) = f_{1+}/N$. The correlation coefficient between A and B can be written as

$$\begin{aligned}\rho_{AB} &= \frac{1 \cdot 1\frac{f_{11}}{N} - \frac{f_{1+}}{N} \cdot \frac{f_{+1}}{N}}{\sqrt{\frac{f_{1+}}{N}(1-\frac{f_{1+}}{N})}\sqrt{\frac{f_{+1}}{N}(1-\frac{f_{+1}}{N})}} \\ &= \frac{Nf_{11} - f_{1+} \cdot f_{+1}}{\sqrt{f_{1+}(N-f_{1+})f_{+1}(N-f_{+1})}} \\ &= \frac{Nf_{11} - f_{1+} \cdot f_{+1}}{\sqrt{f_{1+}f_{0+}f_{+1}f_{+0}}} \qquad (4)\end{aligned}$$

The numerator of this equation can be simplified into the following :

$$\begin{aligned}Nf_{11} - f_{1+} \cdot f_{+1} &= Nf_{11} - (f_{11} + f_{10})(f_{11} + f_{01}) \\ &= Nf_{11} - f_{11}^2 - f_{11}f_{10} - f_{11}f_{01} - f_{10}f_{01} \\ &= f_{11}(N - f_{11} - f_{10} - f_{01}) - f_{10}f_{01} \\ &= f_{11}f_{00} - f_{10}f_{01} \qquad (5)\end{aligned}$$

Hence, equation 4 becomes

$$\rho_{AB} = \frac{f_{11}f_{00} - f_{10}f_{01}}{\sqrt{f_{1+}f_{0+}f_{+1}f_{+0}}} \qquad (6)$$

The above equation is obtained assuming that the contingency table is constructed using frequencies of an entire population. For finite samples, Pearson introduces the phi-coefficient, ϕ, which is defined to be :

$$\phi = \frac{\hat{f}_{11}\hat{f}_{00} - \hat{f}_{10}\hat{f}_{01}}{\sqrt{\hat{f}_{1+}\hat{f}_{0+}\hat{f}_{+1}\hat{f}_{+0}}} \qquad (7)$$

where $\hat{f}_{ij}$'s are the observed frequencies from samples of the population. Notice the similarity between this coefficient and equation 6. We will use the term correlation coefficient and ϕ-coefficient interchangeably for the rest of the paper.

Example 1.

	$B=1$	$B=0$	
$A=1$	4	3	7
$A=0$	1	2	3
	5	5	10

Table 2. Simplified Example

The ϕ-coefficient for Table 2 is $(8-3)/\sqrt{5 \times 5 \times 7 \times 3} = 1/\sqrt{21}$.

The ϕ-coefficient is closely related to χ^2 statistic. Upon comparing equation 7 with equation 2, we would obtain $\phi^2 = \chi^2/N$. ϕ^2 is also called the coefficient of determination in statistics literature. We will use the ϕ-coefficient as the reference metric for comparison with other interestingness measures.

2.2.2 Predictive Association Another statistical measure of association, developed by Goodman and Kruskal, is called the index of predictive association. This measure is based upon the following observation : if two variables are highly dependent, then the error in predicting one of the variables would be smaller whenever the other variable is known. To be more precise, the index of predictive association for a variable, say A, is defined to be :

$$\lambda_A = \frac{P(\epsilon_A) - P(\epsilon_A|B)}{P(\epsilon_A)} \tag{8}$$

where ϵ_A is the error in predicting A. λ_A is a non-negative real number between 0 and 1. If no other information is available, the best guess we can make regarding the value of A is the value $\hat{A}$ that has the largest probability, ie. $\hat{A} = \arg(\max_k P(A_k))$. The error in using this estimate is $P(\epsilon_A) = 1 - P(\hat{A}) = 1 - \max_k P(A_k)$.

Now, suppose we observe $B = B_1$. With this extra information, the best estimate of A is the value that maximizes the conditional probability $\hat{A} = \arg(\max_k P(A_k|B_1))$. The error associated with this estimator is $P(\epsilon_A|B_1) = 1 - \max_k P(A_k|B_1)$. The average prediction error for A given B can be computed by averaging over the entire range of B values :

$$\begin{aligned} P(\epsilon_A|B) &= P(\epsilon_A|B_1)P(B_1) + \cdots + P(\epsilon_A|B_m)P(B_m) \\ &= (1 - \max_k P(A_k|B_1))P(B_1) + \cdots + (1 - \max_k P(A_k|B_m))P(B_m) \\ &= \sum_j P(B_j) - \sum_j \max_k P(A_k|B_j)P(B_j) \\ &= 1 - \sum_j \max_k P(A_k, B_j) \end{aligned} \tag{9}$$

where $\sum_j P(B_j) = 1$ and $P(A_k|B_j)P(B_j) = P(A_k, B_j)$.

Equation 8 can now be written as :

$$\begin{aligned} \lambda_A &= \frac{\sum_j \max_k P(A_k, B_j) - \max_k P(A_k)}{1 - \max_k P(A_k)} \\ &= \frac{\sum_j \max_k f_{jk} - \max_k f_{+k}}{N - \max_k f_{+k}} \end{aligned} \tag{10}$$

This equation can be used as a measure for the implication rule $B \Longrightarrow A$. For the itemset $\{A, B\}$, we may prefer a symmetric version of this coefficient :

$$\lambda_{AB} = \frac{\sum_j \max_k f_{jk} + \sum_k \max_j f_{jk} - \max_k f_{+k} - \max_j f_{j+}}{2N - \max_k f_{+k} - \max_j f_{j+}} \tag{11}$$

Example 2. For the example given in Table 2, $\lambda_{AB} = (4+2+4+3-5-7)/(20-5-7) = 1/8$.

The relationship between ϕ and λ is shown in Figure 2. The graphs in this figure are generated using synthetic data. They show the effect of changing both upper and lower support thresholds on itempairs of the dataset. In practice, many real-world datasets may involve tens of thousands of variables and millions of tuples (rows). For such datasets, support for interesting variables are quite low, say less than 30% or 40%. Association rule mining on these datasets often require a minimum support threshold at 0.1%, 0.5% or even as high as 5% (if too many patterns are generated). These graphs show that a strong linear relationship exists between λ and the ϕ-coefficient within the above range of low support values. Further discussion about the effect of changing support thresholds will be presented in Section 3. Also, one can observe that the predictive association measure treats both negative and positive correlations in the same way.

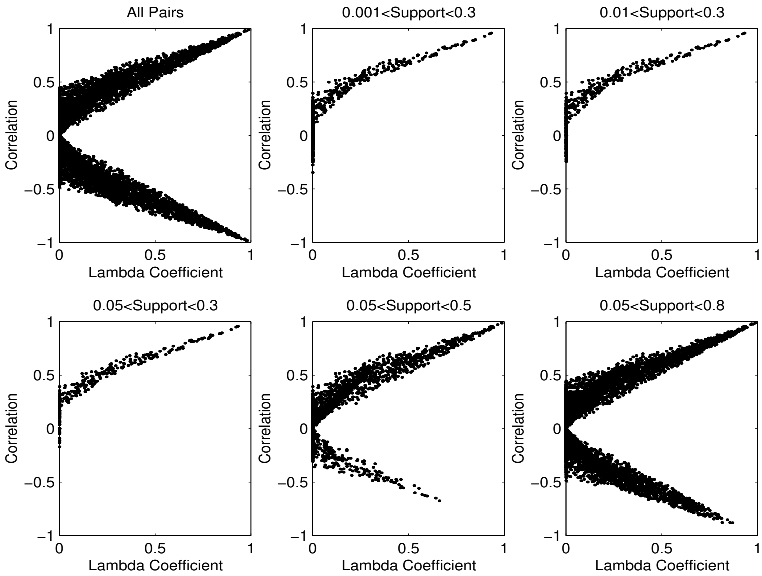

Fig. 2. Correlation coefficient versus Predictive association plot.

2.2.3 Uncertainty A third measure of association originates from information theory. Uncertainty (or entropy) of a variable X is given by $H(A) = -\sum_k P(A_k) \log P(A_k)$. In a sense, this measure is analogous to statistical variance. When the probabilities are evenly spread among the values of a variable, then its entropy will be large; and vice-versa. We can extend this concept to more than one variable. For example, the joint entropy between two variables is defined to be $H(A,B) = -\sum_j \sum_k P(A_k, B_j) \log P(A_k, B_j)$ while its conditional entropy is $H(A|B) = H(A,B) - H(B)$. One way to specify the degree of association between two variables is in terms of a quantity called mutual information

:

$$\begin{aligned} I(A,B) &= H(A) - H(A|B) \\ &= H(A) + H(B) - H(A,B) \\ &= \sum_i \sum_j P(A_i,B_j) \log \frac{P(A_i,B_j)}{P(A_i)P(B_j)} \end{aligned} \tag{12}$$

Mutual information specifies the amount of reduction in uncertainty of a variable A when a second variable B is known. This measure is symmetric with respect to both A and B. The overall measure of association between A and B can be expressed in terms of the maximum mutual information ratio :

$$S(A,B) = \frac{H(A) + H(B) - H(A,B)}{\min[H(A), H(B)]} \tag{13}$$

The relationship between the ϕ-coefficient and S is shown in Figure 3. The graphs look rather similar to those between predictive association and ϕ and are symmetric about the horizontal axis when all itempairs are considered. In the region of low support values, there is a strong linear relationship between the two measures. This indicates that any one of the three measures can be used as a statistical measure of association between dichotomous variables, within this region of interest.

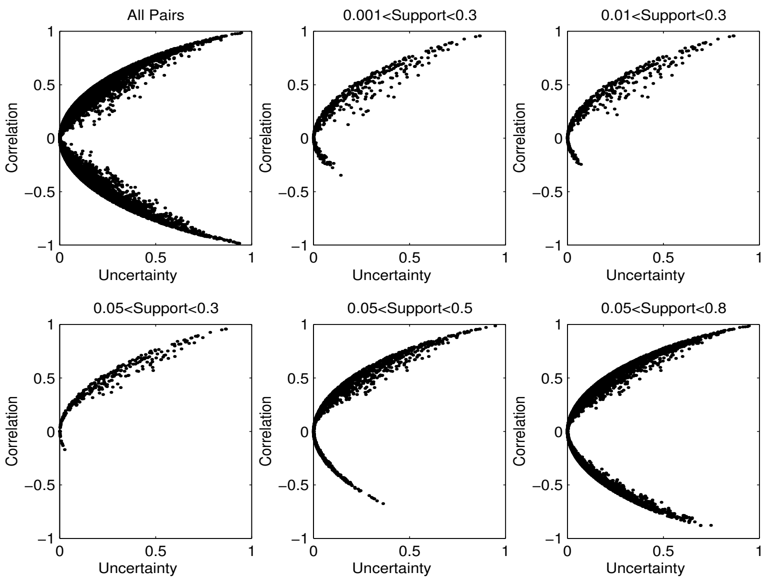

Fig. 3. Correlation coefficient versus Uncertainty.

3 Measures of Dependency from Data Mining

In recent years, various data mining techniques have been proposed to extract interesting patterns from large databases. Due to the large number of discovered patterns, numerous objective measures were introduced to quantify the interestingness of these patterns. Most of the measures can be adapted to association-type patterns (such as frequent itemset and association rules). In this section, we will describe some of the measures and compare their characteristics with respect to statistical correlation. We start by presenting the fundamental principles underlying a good, intuitive objective measure.

3.1 Principles of Objective Measures of Interestingness

Objective measure is a data-driven, domain-independent approach to evaluate the interestingness of discovered patterns [2]. [18] have outlined three basic principles that must be obeyed by any intuitive objective measure, F :

1. $F = 0$ if A and B are statistically independent; i.e. whenever $P(A, B) = P(A)P(B)$.
2. F monotonically increases with $P(A, B)$ when all other parameters remain the same.
3. F monotonically decreases with $P(A)$ (or $P(B)$) when the rest of the parameters stay the same.

The first principle ensures that patterns that occur by chance have zero interest value. The second principle states that the interest value should be large for patterns that have higher statistical significance (where statistical significance is expressed in terms of the support of the pattern, $P(A, B)$). The third principle is used to compare the interest values of patterns with equivalent statistical significance. The pattern that requires a larger coverage ($P(A)$ or or $P(B)$) in order to attain the same degree of significance, is less interesting than the pattern with lower coverage.

Many other authors have attempted to extend and refine these fundamental principles. For example, [15] have added a fourth principle saying that F should be monotonically increasing with $P(A)$ while $P(A, B)/P(A)$ and $P(B)$ remain unchanged. [12] introduced a fifth principle to distinguish between two types of inductive rules, called discriminant rules and characteristic rules. Although these are useful extensions, we believe that the essential properties of a good objective measure are sufficiently captured by the first three principles.

These principles may serve as a guideline for evaluating different objective measures. In practice, however, there are many good measures that do not satisfy all of the above principles. For example, the first principle is too rigid because it specifies what the absolute interest value should be for independent items. One way to circumvent this is by requiring $|F|$ to be equal to some constant c_{min} when both A and B are independent, or by rescaling F to $F' \equiv F - c_{\min}$.

[2] In contrast to subjective measures which are more user-driven and domain-dependent.

3.2 Support and Confidence

A rule $A \rightarrow B$ has support s if $s\%$ of all the transactions contains both A and B. The rule has a confidence c if $c\%$ of all transactions that contain A also contain B. In terms of the 2×2 contingency table, $s = f_{11}/N$ and $c = f_{11}/f_{1+}$.

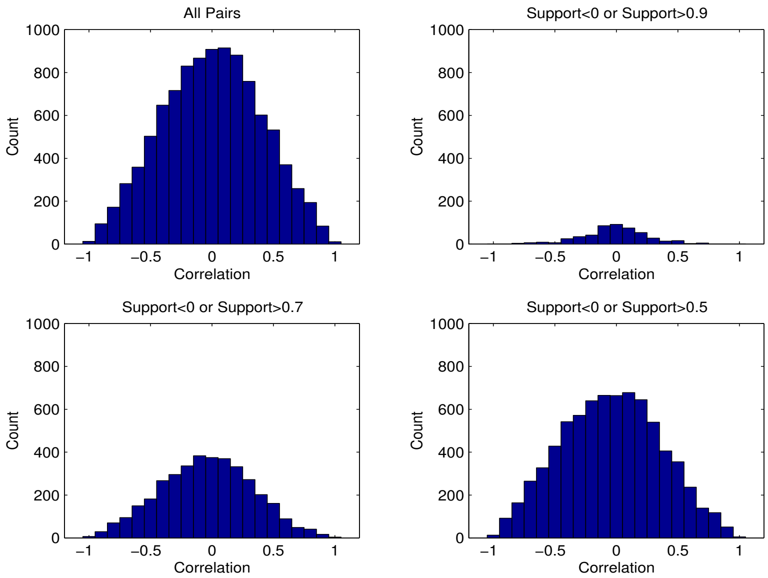

Fig. 4. Number of itempairs removed by applying upper support threshold.

Earlier, it was mentioned that support is necessary because it measures the statistical significance of a pattern. Since the choice of an appropriate support threshold is often ad-hoc, we need to ensure that support-based pruning will not remove many of the interesting patterns. In this paper, we assume that only positively correlated itemsets are of interest to a data analyst. Figures 4 and 5 show the effect of applying various support thresholds on the synthetic dataset. The upper-left graph in both figures depict the histograms of ϕ-coefficients for every itempairs in the dataset. These histograms appear to be very similar to a Gaussian distribution. The rest of the histograms show the itempairs that are removed when various support thresholds are imposed.

Figure 4 shows the number of itempairs that were removed when the upper support threshold is decreased. The distribution of the removed itempairs is similar to a Gaussian distribution. This result indicates that by placing a maximum support threshold, one will end up pruning uncorrelated, positively correlated and negatively correlated itempairs in equal proportions to their initial distribution. In contrast, if a lower bound of support is specified, most of the itempairs removed are either uncorrelated or negatively correlated. This result makes sense

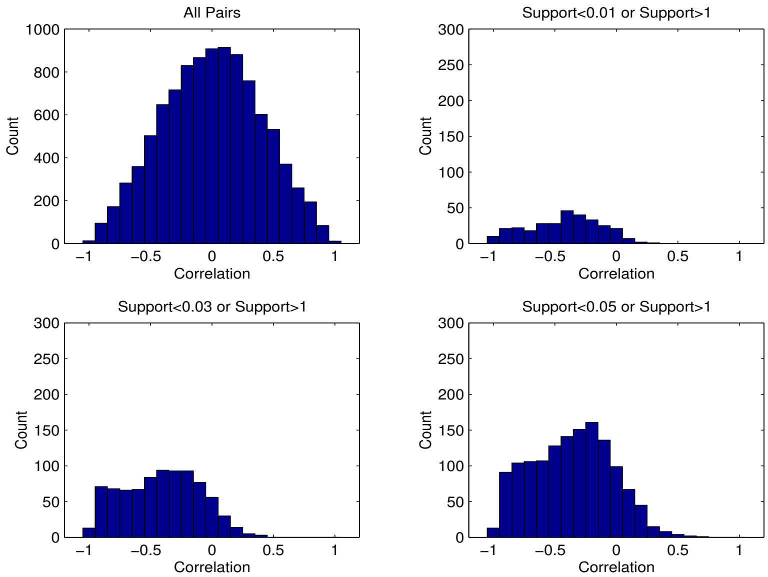

Fig. 5. Number of itempairs removed by applying lower support threshold.

because itempairs with low support tend to have large values in f_{10}, f_{01} or f_{00} cells of their contingency tables. This would often correspond to uncorrelated or negatively correlated itemsets.

Confidence was initially proposed to measure the strength of an implication rule. [7] showed that confidence may produce counter-intuitive rules especially when support of the itemset is high, as will be shown in the next example.

Example 3.

	$WindowsNT$	$\overline{WindowsNT}$	
$Linux$	20	10	30
$\overline{Linux}$	60	10	70
	80	20	100

Table 3. A 2×2 contingency table example.

Consider the 2×2 table shown in Table 3. It summarized the purchase of two brands of operating systems at a retail store within a certain time period. Each cell in this table corresponds to the frequency of events pertaining to the purchase of Windows NT or Linux operating systems. For instance, the upper right cell entry indicates that within the period under consideration, 10 customers bought the Linux operating systems but not Windows NT. Now, suppose the support and confidence thresholds were set at 5% and 50% respectively. The association rule $Linux \Longrightarrow WindowsNT$ would have a 20% support and 67% confidence.

Thus, it will pass both threshold conditions and eventually declared to be interesting. However, this information can be misleading. The prior probability that a customer purchases Windows NT is 80%. Once we know that the customer had purchased Linux, the conditional probability that he or she would buy Windows NT reduces to 75%. In other words, the rule $Linux \Longrightarrow WindowsNT$ does not make sense in this situation. This is why confidence may not be an appropriate measure.

The relationship between maximum confidence and correlation is shown in figure 6. Here, we assume that the overall confidence of an itemset is represented by the maximum confidence among the rules that can be generated from this itemset :

$$MaxConf = \max(\frac{f_{11}}{f_{1+}}, \frac{f_{11}}{f_{+1}}) \tag{14}$$

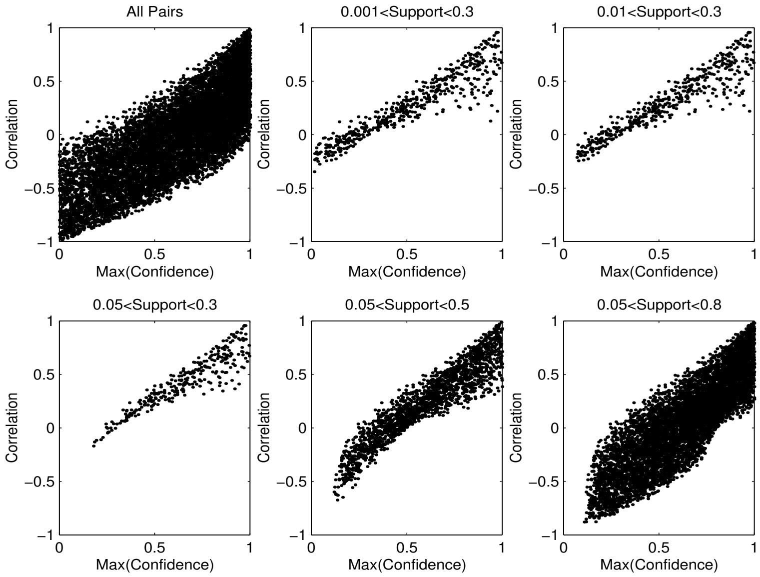

Fig. 6. Correlation coefficient versus Maximum Confidence.

Other confidence-like measures have been used in various data mining applications. For example, the laplace function [11]

$$\text{laplace} = \frac{\sigma(A \cup B) + 1}{\sigma(A) + 2} \tag{15}$$

is often used to measure the expected accuracy of classification rules. In this formula, $\sigma(\cdot)$ denotes the support count of an itemset. Both confidence and laplace function satisfy all but the first principle of objective measure.

3.3 Interest and *IS* Measure

As of late, interest factor is becoming a popular measure of interestingness for association-like patterns [7, 19, 6, 10]. This metric is defined to be the ratio between the joint probability of two variables with respect to their expected probabilities under the independence assumption.

$$I(A,B) = \frac{P(A,B)}{P(A)P(B)} \tag{16}$$

The interest factor can be any non-negative real number; with a value of 1 corresponding to statistical independence. This metric is desirable because it satisfies all three fundamental principles of objective measures.

Interest factor is closely related to the ϕ coefficient. For a 2×2 contingency table,

$$I(A,B) = \frac{f_{11}N}{f_{1+}f_{+1}} \tag{17}$$

If we re-arrange equation 4, we can obtain the following relationship between interest factor and the ϕ-coefficient :

$$\begin{aligned}
\phi &= \frac{Nf_{11} - f_{1+} \cdot f_{+1}}{\sqrt{f_{1+}f_{0+}f_{+1}f_{+0}}} \\
&= \frac{\frac{Nf_{11}}{f_{1+}f_{+1}} - 1 \cdot f_{1+}f_{+1}}{\sqrt{f_{1+}f_{0+}f_{+1}f_{+0}}} \\
&= \frac{(I-1) \cdot \sqrt{f_{1+}f_{+1}}}{\sqrt{f_{0+}f_{+0}}}
\end{aligned} \tag{18}$$

For large databases, the support count of a given itemset tends to be low, i.e. $\frac{f_{1+}}{N} \ll 1$ and $\frac{f_{+1}}{N} \ll 1$. In this scenario, both $\frac{f_{0+}}{N}$ and $\frac{f_{+0}}{N}$ are close to 1. Moreover, highly correlated items have $I \gg 1$. With this choice of approximation, equation 18 becomes :

$$\begin{aligned}
\phi &\approx I\sqrt{\frac{f_{1+}f_{+1}}{N^2}} \\
&= \frac{Nf_{11}}{f_{1+}f_{+1}}\sqrt{\frac{f_{1+}f_{+1}}{N^2}} \\
&= \sqrt{\frac{Nf_{11}}{f_{1+}f_{+1}} \cdot \frac{f_{11}}{N}} \\
&= \sqrt{I \times \frac{f_{11}}{N}}
\end{aligned} \tag{19}$$

This suggests that a better interestingness measure, derivable from statistical correlation, in the region of low support and high interest values is :

$$IS = \sqrt{I \times \frac{f_{11}}{N}} \tag{20}$$

IS has many desirable properties despite violating the first principle of objective measure. First of all, it contains a product of two important quantities, interest factor and support. In other words, this measure takes into account both interestingness and signifance of a pattern. Second, IS is equivalent to the geometric mean of confidence of rules that can be generated from the itempair i.e. $IS = \sqrt{Conf(A \Longrightarrow B) \times Conf(B \Longrightarrow A)}$. Another interpretation of this measure is as the cosine angle between two random vectors, i.e. $IS = P(A,B)/\sqrt{P(A,A)P(B,B)}$.

Figures 7 and 8 show the relationship between interest factor and IS with the ϕ-coefficient using the synthetic dataset. Note the high linearity exhibited by the IS measure, agreeing with the theoretical arguments above.

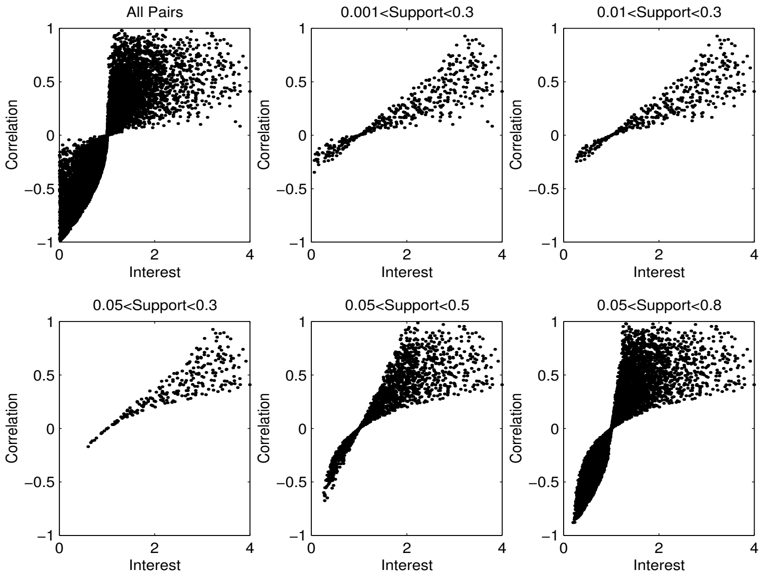

Fig. 7. Correlation coefficient versus Interest factor.

We have also repeated our experiments using real-world datasets. The first dataset is a subset of Reuters newswire articles[3]. This dataset contains 2886 attributes and 2005 documents. The second dataset is obtained from a large retail corporation. This dataset has 14462 attributes and 58565 transactions. The relationship between IS and ϕ-coefficient for these datasets are shown in Figure 9.

[3] available at http://www.research.att.com/˜lewis.

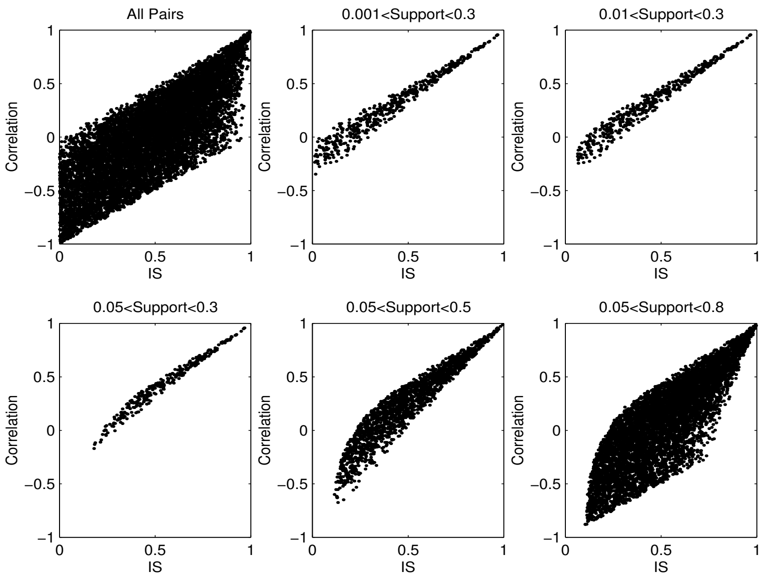

Fig. 8. Correlation coefficient versus IS measure.

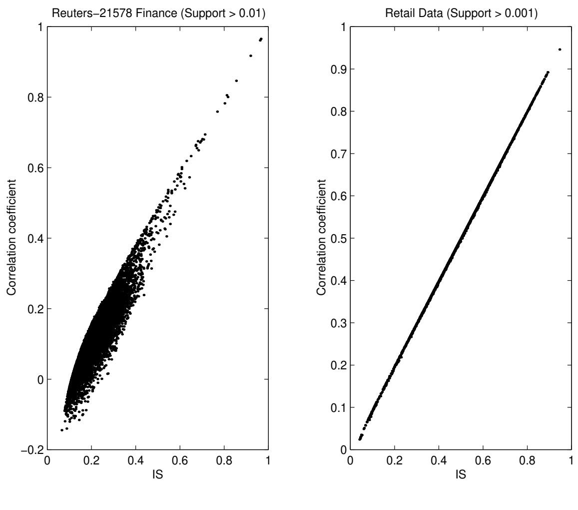

Fig. 9. Correlation coefficient versus IS measure for Reuters-21578 (Finance) and retail dataset).

3.4 Other Measures

We now describe three other interestingness measures that can be used for association-type patterns. They are the Gini index [5], Piatetsky-Shapiro's rule-interest [18] and conviction [8].

The Gini index for an association rule $A \Longrightarrow B$ is given by

$$\begin{aligned} Gini = \; & P(A)(P(B|A)^2 + P(\neg B|A)^2)) + P(\neg A)(P(B|\neg A)^2 + P(\neg B|\neg A)^2)) \\ & -P(B)^2 - P(\neg B)^2 \end{aligned} \tag{21}$$

This value may range from 0 (when A and B are completely independent) to 0.5 (for perfect correlation). Figure 10 shows the relationship between Gini index and correlation using the synthetic dataset. The symmetric nature of this plot indicates that Gini index treats both positive and negatively correlated rules in the same way. However, as the minimum support threshold increases, negatively correlated itempairs will be pruned, which agrees with our earlier observation.

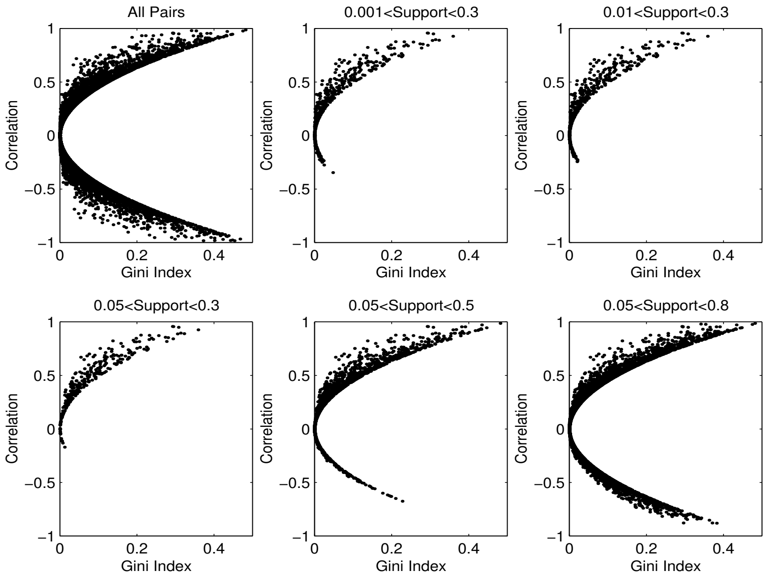

Fig. 10. Correlation coefficient versus Gini index.

The rule-interest function was introduced in [18] as a simple measure that satisfies all three fundamental principles of objective measures. This measure is defined to be

$$RI = P(A, B) - P(A)P(B) \tag{22}$$

The range of this function is between -0.25 and 0.25. Statistical independence occurs at $RI = 0$. The relationship between this measure and statistical correlation using synthetic dataset is shown in Figure 11.

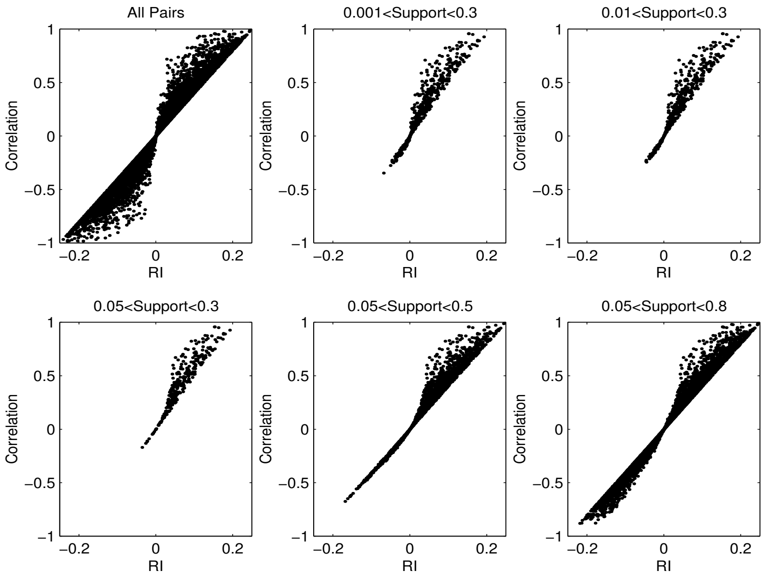

Fig. 11. Correlation coefficient versus Piatetsky's rule interest measure.

Conviction was introduced in [8] as an asymmetric version of the interest factor :

$$\text{conviction} = \frac{P(A)P(\neg B)}{P(A, \neg B)} \tag{23}$$

This measure is derived from interest factor in the following way. A rule $A \Longrightarrow B$ is logically equivalent to $\neg(A \wedge \neg B)$. Thus, equation (23) is an asymmetric way of testing the independence between A and $\neg B$.[4]. The ratio between $P(A, \neg B)$ and $P(A)P(\neg B)$ is inverted due to the negation symbol in the logical expression $\neg(A \wedge \neg B)$.

Conviction is different from confidence because it does not suffer from the same problem of producing misleading rules. Unlike interest factor, conviction will assign the value $+\infty$ if the confidence of the rule is 1 (regardless of what $P(A, B)$ is). Figure 12 shows the relationship between maximum conviction and the ϕ-coefficient.

4 Ranking of Association Patterns

In this section, we will show how the various measures described previously can be used for ranking the association patterns according to their degree of interestingness.

[4] If A and B are independent, so does $(A, \neg B)$, $(\neg A, B)$ and $(\neg A, \neg B)$.

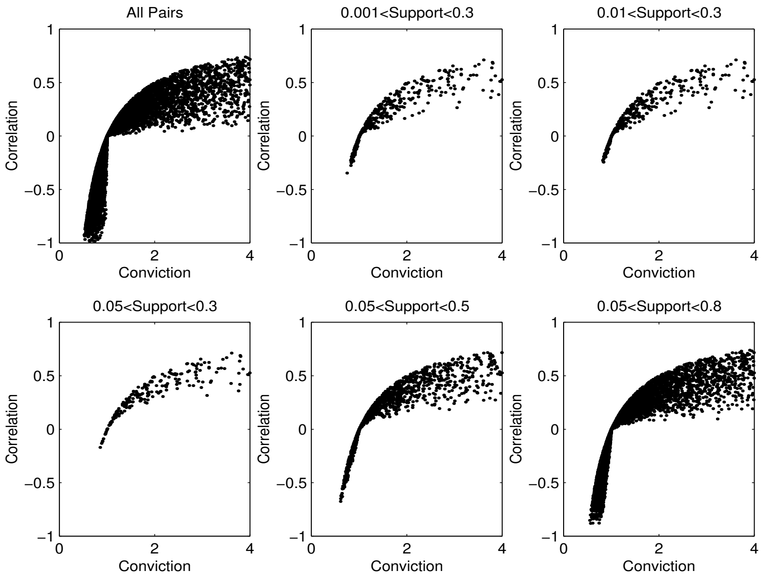

Fig. 12. ϕ-coefficient versus conviction.

Instead of ranking every itemsets, a good starting point would be to rank only itemsets that fall into region II of Fig. 1. Firstly, we need to determine the maximal frequent itemset border and χ^2 border using standard algorithms such as Apriori [4] and the Dependence Rules Algorithm[7]. The two borders can be used to remove all itemsets that are infrequent or independent. Next, we would compute the interest value for each remaining itemset according to an interestingness measure, F. If an analyst is only interested in itemsets, we can return the highly-ranked itemsets.

However, if an analyst is interested in rules rather than itemsets, one must define the corresponding objective measures for rules, F'. In many cases, the objective measures for itemsets may not be the same as that for association rules. Therefore, one must ensure that both F and F' are consistent with each other.

We will now illustrate an example of ranking itemsets and rules using the interest factor, I. Consider a large k-itemset $\{A_1, A_2 \cdots A_k\}$. There are 2^{k-2} ways to partition the itemset into rules. [5] The interest factor for the large k-itemset can be written as :

$$I(A_1, A_2, \cdots A_k) = \frac{P(A_1, A_2, \cdots A_k)}{P(A_1)P(A_2)\cdots P(A_k)} \quad . \tag{24}$$

Suppose we want to compute the corresponding interest value for the rule $A_1 A_2 \cdots A_j \longrightarrow A_{j+1} A_{j+2} \cdots A_k$. We can rewrite the above equation into the

[5] Here, due to the symmetry of the I factor, we assume that the rules $A \longrightarrow B$ and $B \longrightarrow A$ have the same interest value.

following form :

$$\begin{aligned}
I(A_1, A_2, \cdots A_k) &= \frac{P(A_1, A_2, \cdots A_j)\ P(A_{j+1}, A_{j+2}, \cdots A_k | A_1 \cdots A_j)}{P(A_1)P(A_2)\cdots P(A_k)} \\
&= \frac{P(A_1, A_2, \cdots A_j)}{P(A_1)P(A_2)\cdots P(A_j)}\ \frac{P(A_{j+1}, A_{j+2}, \cdots A_k | A_1 A_2 \cdots A_j)}{P(A_{j+1})P(A_{j+2})\cdots P(A_k)} \\
&= I(A_1, A_2, \cdots A_j)\ \frac{P(A_1, A_2, \cdots, A_n)}{P(A_{j+1})P(A_{j+2})\cdots P(A_k)P(A_1 A_2 \cdots A_j)} \\
&\quad \times \frac{P(A_{j+1}, A_{j+2}, \cdots A_k)}{P(A_{j+1}, A_{j+2}, \cdots A_k)} \\
&= I(A_1, A_2, \cdots A_j)\ I(A_{j+1}, A_{j+2}, \cdots A_k) \\
&\quad \times \frac{P(A_1, A_2, \cdots A_k)}{P(A_1, A_2, \cdots A_j)P(A_{j+1}, A_{j+2} \cdots A_k)}\ . \qquad (25)
\end{aligned}$$

The above equation allows us to define the interest factor for a rule in terms of the interest factor for the corresponding itemsets :

Definition 1. *The interest factor for the rule* $A_1 A_2 \cdots A_j \longrightarrow A_{j+1} A_{j+2} \cdots A_k$ *can be defined as :*

$$\begin{aligned}
I(A_1 \cdots A_j \longrightarrow A_{j+1} \cdots A_k) &= \frac{P(A_1, A_2, \cdots A_k)}{P(A_1, A_2, \cdots A_j)P(A_{j+1}, A_{j+2} \cdots A_k)} \\
&= \frac{I(A_1, A_2, \cdots A_k)}{I(A_1, A_2, \cdots A_j)\ I(A_{j+1}, A_{j+2}, \cdots A_k)}\ . \qquad (26)
\end{aligned}$$

The above definition is useful because it allows us to compute the interest factor for a rule using only the interest factors of the itemsets. Furthermore, it says that the best rule for a given itemset is the one that maximizes the difference between $I(A_1, A_2, \cdots A_k)$ and the product $I(A_1, A_2, \cdots A_j)\ I(A_{j+1}, A_{j+2}, \cdots A_k)$. This definition can also be used to define the interest part of the IS measure for an association rule.

5 Conclusions

One way to compare the various measures presented in this paper is by determining their correlation with respect to the ϕ-coefficient. Table 4 illustrates the correlation values computed using the synthetic dataset, for various ranges of support values. Notice that RI works tremendously well in almost any situations. However, for low support regions, IS seems to be the best choice, which is not surprising considering it is derived from the correlation coefficient itself. On the other hand, conviction works poorly even for the low support region. This is because it has a very wide range of values (from 0 to ∞). Other measures such as the λ-coefficient and Gini index have very low correlation with ϕ when no support thresholds are imposed. This is because both measures are symmetric about zero (i.e. their values are non-negative). However, as the support region becomes smaller, the symmetry will be broken and the correlation values become larger.

The following conclusions can be made :

Support range	interest factor	IS	laplace	conviction	max conf	λ	entropy	RI	Gini index
$[0, 1]$	0.7057	0.7981	0.7855	0.0511	0.7854	-0.0027	-0.0065	0.9811	-0.0046
$0.005, 1]$	0.7055	0.7979	0.7862	0.0510	0.7861	0.0136	0.0220	0.9814	0.0151
$[0.01, 1]$	0.7135	0.7974	0.7846	0.0510	0.7845	0.0353	0.0541	0.9818	0.0388
$[0.05, 1]$	0.7393	0.7915	0.7659	0.0534	0.7659	0.2101	0.2577	0.9840	0.2263
$[0.005, 0.7]$	0.7293	0.8627	0.8856	0.0477	0.8854	0.0555	0.1011	0.9911	0.0511
$[0.01, 0.7]$	0.7391	0.8650	0.8879	0.0476	0.8878	0.0738	0.1327	0.9912	0.0746
$[0.05, 0.7]$	0.7725	0.8760	0.8929	0.0476	0.8928	0.2506	0.3566	0.9921	0.2855
$[0.005, 0.5]$	0.7315	0.9318	0.9298	0.0483	0.9296	0.5280	0.5571	0.9800	0.4722
$[0.01, 0.5]$	0.7433	0.9342	0.9313	0.0480	0.9311	0.5401	0.5831	0.9798	0.4920
$[0.05, 0.5]$	0.7835	0.9505	0.9350	0.0458	0.9349	0.6970	0.7601	0.9777	0.6914
$[0.005, 0.3]$	0.7057	0.9806	0.9317	0.3199	0.9311	0.8644	0.9023	0.9492	0.8426
$[0.01, 0.3]$	0.7280	0.9820	0.9340	0.3193	0.9336	0.8696	0.9101	0.9469	0.8482
$[0.05, 0.3]$	0.7704	0.9871	0.9273	0.3076	0.9271	0.9147	0.9452	0.9316	0.8897

Table 4. Correlation between different interestingness measures and ϕ-coefficient for various range of support values. These coefficients are computed for itempairs generated using the synthetic dataset.

- Support is a good measure because it represents how statistically significant a pattern is. Support-based pruning is effective because support is an anti-monotone function, and it allows us to prune mostly uncorrelated or negatively correlated patterns.
- χ^2 is appropriate to test whether there is sufficient evidence to show that items in a pattern are independent of each other. However, it does not quantify the strength of correlation among the items.
- Many of the measures (such as IS, laplace, maximum confidence, RI) behave similarly in the region of low support values (which typically occurs in large databases).

The above conclusions suggest that we can use any of the *appropriate* interestingness measures to rank the patterns of region II in Fig. 1. An appropriate measure should be highly correlated with statistical correlation and takes into account the support of the pattern. Finally, we have presented a consistent method for ranking the itemsets and rules according to their interest factor and IS measure.

For our future work, there are several directions we can pursue. So far, we have only compared interestingness measures for pairs of items. Most of the discussion presented here, such as the fundamental principles of objective measures and the various metrics, are applicable to larger itemsets. We have also discussed a method for ranking rules using interest and IS measure. It will be interesting to extend the method to other measures.

References

1. Ramesh C. Agarwal, Charu Aggarwal, and V. V. V. Prasad. A tree projection algorithm for generation of frequent itemsets. *Journal of Parallel and Distributed Computing (Special Issue on High Performance Data Mining)*, (Accepted for Publication) 2000.
2. R. Agrawal, T. Imielinski, and A. Swami. Database mining: a performance perspective. *IEEE Transactions on Knowledge and Data Engineering*, 5:914–925, 1993.
3. R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. ACM SIGMOD Intl. Conf. Management of Data*, pages 207–216, Washington D.C., USA, 1993.
4. R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. of the 20th VLDB Conference*, pages 487–499, Santiago, Chile, 1994.
5. L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Chapman & Hall, New York, 1984.
6. Tom Brijs, Gilbert Swinnen, Koen Vanhoof, and Geert Wets. "using association rules for product assortment decisions : A case study. In *KDD99*, pages 254–260, San Diego, Calif, August 1999.
7. S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *Proc. ACM SIGMOD Intl. Conf. Management of Data*, 1997.
8. S. Brin, R. Motwani, J. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proc. of 1997 ACM-SIGMOD Int. Conf. on Management of Data*, pages 255–264, Montreal, Canada, June 1997.
9. Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proc. of 1997 ACM-SIGMOD Int. Conf. on Management of Data*, pages 255–264, Montreal, Canada, June 1997.
10. Robert Cooley Chris Clifton. 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*, 1999.
11. Peter Clark and Robin Boswell. Rule induction with cn2 : Some recent improvements. In *Proceedings of the European Working Session on Learning EWSL-91*, pages 151–163, Porto, Portugal, 1991.
12. M. Kamber and R. Shinghal. Evaluating the interestingness of characteristic rules. In *KDD96*, pages 263–266, Portland, Oregon, 1996.
13. B.W. Lindgren. *Statistical Theory*. Chapman & Hall, fourth edition, 1998.
14. B. Liu, W. Hsu, and Y. Ma. Pruning and summarizing the discovered associations. In *Proc. of the fifth ACM SIGKDD Intl Conf on Knowledge Discovery and Data Mining*, pages 125–134, San Diego, CA, 1999.
15. John A Major and John A Mangano. Selecting among rules induced from a hurricane database. In *Proceedings of AAAI Workshop on Knowledge Discovery in Databases*, pages 30–31, 1993.
16. R. Ng, L. Lakshmanan, J. Han, and A. Pang. Exploratory mining and pruning optimizations of constrained association rules. In *Proc. of 1998 ACM-SIGMOD Int. Conf. on Management of Data*, pages 13–24, Seattle, WA, June 1998.
17. J.S. Park, M.S. Chen, and P.S. Yu. An effective hash-based algorithm for mining association rules. *SIGMOD Record*, 25(2):175–186, 1995.
18. Gregory Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In Gregory Piatetsky-Shapiro and William Frawley, editors, *Knowledge Discovery in Databases*, pages 2299–248. MIT Press, Cambridge, MA, 1991.

19. 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.
20. R. Srikant and R. Vu, Q.and Agrawal. Mining association rules with item constraints. In *Proc. of the Third Int'l Conference on Knowledge Discovery and Data Mining*, pages 67–73, Newport Beach, CA, August 1997.
21. H Toivonen, M. Klemettinen, P. Ronkainen, K. Hatonen, and H. Mannila. Pruning and grouping discovered association rules. In *ECML-95 Workshop on Statistics, Machine Learning and Knowledge Discovery in Databases*, pages 47–52, Heraklion, Greece, April 1995.