1.5 Interestingness Measures

The goal of mining association patterns is to extract rules or itemsets that are interesting to the data analysts. In the original association rule mining framework[1], support and confidence are used as the criteria to evaluate the quality (or interestingness) of the patterns. Over the years, many alternative measures have been proposed, in the context of association rules as well as other types of logical rules such as classification rules [21], probabilistic rules [22], strong rules [23], etc. These measures can be applied during preprocessing, mining or post-processing steps of the knowledge discovery process:

1. Interestingness measures are used during preprocessing to eliminate items or features that are uninteresting or irrelevant to the knowledge discovery task. For instance, the support measure can be used to remove items that occur very frequently in the transaction database since patterns containing these items are usually uninteresting. Examples of such highly frequent items include stopwords (for the text mining domain), shipping and handling cost (a common line item in the e-commerce domain) and homepages (which have very high access rate in the Web Mining domain).

2. Interestingness measures can be used during the mining process to reduce the combinatorial complexity of the search problem. However, not all interestingness measures can be used effectively. For example, support, and to certain extent, confidence, are effective measures because they possess anti-monotonicity properties that can help to prune the search space.

3. Measures that do not possess any monotonicity properties can be applied during post-processing to filter or rank the discovered patterns.

In general, interestingness measures can be divided into two broad categories: objective and subjective measures [24]. Objective measures rate the patterns based on some statistics computed from the observed data. This include measures such as support, confidence, chi-square, interest, correlation, etc. In contrast, subjective measures depend not only on the statistics of the data, but also rely on user's interpretation. Among the subjective measures that have been proposed include generalized impressions [25], neighborhood-based unexpectedness [26], etc.

1.5.1 Objective Measures

Objective measure is a data-driven approach to assess the interestingness of a pattern. It is domain-independent and requires minimal user participation (aside from specifying the quality measure threshold). Some objective measures are symmetric with respect to permutation of the items while other are not. From an association rule mining perspective, symmetric measures are often used for itemsets whereas asymmetric measures are applied to rules.

Principles of Objective Measures

For any given pattern, in the form of a rule $X \Longrightarrow Y$ or itemset $X \cup Y$, the objective measure (OM) of the pattern is usually computed as a function of the support counts $\sigma(X)$, $\sigma(Y)$, $\sigma(X \cup Y)$ and the total number of transactions, $|T|$. In [23], Piatetsky-Shapiro had outlined three basic principles obeyed by any good objective measure:

1. $OM = 0$ if X and Y are statistically independent.

2. OM monotonically increases with $\sigma(X \cup Y)$ when all other parameters remain the same.

3. OM monotonically decreases with $\sigma(X)$ (or $\sigma(Y)$) when all other parameters remain the same.

We will illustrate the motivation behind these principles with several examples. Let $X = 1$ (or $X = 0$) denotes the presence (or absence) of item X from a given transaction. We can summarize the relationship between two items, X and Y, by using a 2×2 *contingency table* as shown in Table 1.2. The four cells in this table represent all the possible combinations of X and Y values. Let f_{ij} denotes the support count for

	$Y = 1$	$Y = 0$	
$X = 1$	f_{11}	f_{10}	f_{1+}
$X = 0$	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	$\|T\|$

Table 1.2: A 2×2 contingency table between X and Y.

	$Y = 1$	$Y = 0$	
$X = 1$	60	80	140
$X = 0$	120	160	280
	180	240	420

Table 1.3: Contingency table for Example 11.

$X = i$ and $Y = j$. Also, $f_{i+} = f_{i0} + f_{i1}$ is the support count for $X = i$ while f_{+j} is the support count for $Y = j$.

The first principle is based on the assumption that a pattern is considered to be interesting only if items within the pattern are highly dependent on each other. This is the principle which relates the notion of objective interestingness to statistical dependence. A well-known fact from statistics states that a pair of variables, X and Y, is independent if and only if $P(Y|X) = P(Y)$, or equivalently, $P(X \cup Y) = P(X)P(Y)$. Therefore, the first principle of objective measure suggests that the least interesting pattern is the one involving statistically independent items.

Example 11 : Let us turn to the example shown in Table 1.3. There are 420 transactions in the database and item Y occurs in 180 of them. Without any prior knowledge, the probability that item Y is present in a transaction is $P(Y) = \sigma(Y)/|T| = 180/420 = 3/7$. Next, suppose we know that item X is present in a particular transaction. Will this affect the likelihood of observing item Y in the same transaction? If it does, then we say that X and Y are correlated; otherwise, they are independent of each other. In this example, the probability of observing Y given X can be computed using Bayes theorem, i.e. $P(Y|X) = \sigma(X \cup Y)/\sigma(X) = 60/140 = 3/7$, which co-incides with $P(Y)$. Thus, X and Y are independent of each other because knowing the presence of X does not influence the probability for predicting the occurrence of Y.

The second principle is also quite intuitive because it suggests that a pattern with large support is more interesting than one having a low support, when everything else stays the same.

Example 12 : Consider the database summary given in Table 1.4. The support counts for items X and Y are identical to the example given in Table 1.3. However, the values of f_{11} and f_{00} of this table is significantly larger than their respective values in the previous example. Intuitively, a larger f_{11} means that whenever X occurs, the odds of Y being present in the same transaction is high. Conversely, a larger f_{00} means that the absence of X is a good indicator for the absence of Y. Hence, if both f_{11} and f_{00} are large, then we may conclude that X and Y are positively correlated with each other. The second principle acknowledges this fact by suggesting that an objective measure should rate patterns with large support (f_{11}) more favorably because they tend to correspond to positively correlated items.

	$Y = 1$	$Y = 0$	
$X = 1$	100	40	140
$X = 0$	80	200	280
	180	240	420

Table 1.4: Contingency table for Example 12.

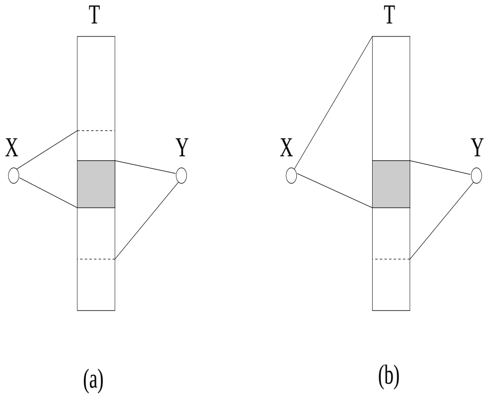

Figure 1.9: A comparison between the significance of $P(X \cup Y)$ (i.e. the shaded region). X and Y are items while T is the set of all transactions. Intuitively, the interestingness measure for (a) should be larger than (b) because the shaded region for (a) is more significant.

	$Y = 1$	$Y = 0$	
$X = 1$	100	120	220
$X = 0$	80	120	200
	180	240	420

Table 1.5: Contingency table for Example 13.

The third principle can be understood by examining the significance of transactions that contain all items in a pattern. In figure 1.9, the circles represent items (or itemsets) while the middle rectangle represents the set of all transactions. From a statistical perspective, the shaded region of figure 1.9(b) is less significant compare to figure 1.9(a) because it requires a larger coverage for X. The third principle suggests that an objective measure should rate the pattern exhibited by the left hand side figure more favorably because its overlapping region is more significant.

Example 13 : In Table 1.5, we have increased the support count of X while maintaining the support counts of Y and $X \cup Y$. This increment has resulted in a smaller value of f_{00}. Since the entries of the table has become more uniform, it is plausible to surmise that the association between X and Y has grown weaker compare to example 12. This observation is consistent with the third principle of objective measure.

Besides these basic principles, Piatetsky-Shapiro have suggested to incorporate other parameters such as pattern complexity when evaluating the significance of a pattern. In the meantime, several other authors have also attempted to extend the basic principles beyond their current form. For instance, Major and Mangano [27] proposed a fourth principle which states that OM should monotonically increase with $\sigma(X)$ when $\sigma(X \cup Y)/\sigma(X)$ and $\sigma(Y)$ are unchanged. Kamber and Singhal [28] introduced a fifth principle to distinguish between two types of inductive rules, called discriminant and characteristic rules. Despite these extensions, the essential properties for a good objective measure are sufficiently captured by the first three principles alone.

Objective Measure from Statistics

In statistics, a standard way to measure the degree of association between binary variables is to use Pearson's ϕ-coefficient, where:

$$\phi = \frac{f_{11}f_{00} - f_{10}f_{01}}{\sqrt{f_{1+}f_{0+}f_{+1}f_{+0}}} \tag{1.1}$$

The value of this coefficient ranges from -1 to $+1$. [3] Moreover, for any 2×2 contingency tables, $P(X \cup Y) - P(X)P(Y) = (f_{11}f_{00} - f_{10}f_{01})/|T|$. If the two variables are independent, then $P(X \cup Y) = P(X)P(Y)$, or $f_{11}f_{00} = f_{10}f_{01}$. Hence, the ϕ-coefficient vanishes; which agrees with the first principle of objective measures. One can also verify that the rest of the principles are satisfied by this measure. This will be left as an exercise for the reader.

Example 14 : The ϕ-coefficient for the example given in Table 1.4 is

$$(200 \times 100 - 80 \times 40)/\sqrt{180 \times 240 \times 140 \times 280} = .408$$

which is relatively larger than zero. This agrees with our earlier observation in Example 12 that X and Y are positively correlated with one another.

χ^2 is another widely-used statistical measure to test for independence between variables. Brin et al. [36] have used this measure to determine whether items in a given itemset are statistically independent of each other. The basic idea of the χ^2 test is to assume that the variables (items) obey a hypothetical probability distribution. By calculating the deviation between the observed data distribution from its expected (hypothetical) distribution, one can decide whether to accept or reject the initial hypothesis. The χ^2 statistic is computed in the following way:

$$\chi^2 = \sum_{j,k} \frac{(f^o_{jk} - f^e_{jk})^2}{f^e_{jk}} \tag{1.2}$$

This statistic measures the sum of the normalized deviation between the observed support count, f^o_{jk}, of each cell in the contingency table from its expected support count, f^e_{jk}.

To test for variable dependencies, we first hypothesize that the variables are independent of each other. In this case, the expected support count for each contingency table cell entry is $f^e_{jk} = N \times \frac{f^o_{j+}}{N} \times \frac{f^o_{+k}}{N}$. If the computed χ^2 value is large, then we have more evidence to reject the independence hypothesis. This begs the question of how large should the χ^2 value be in order to conclude with high confidence that the variables are not independent? This question can be answered if we know the actual distribution of the χ^2 statistic. For large enough $|T|$, it can be shown that this statistic has a chi-square probability distribution with 1 degree of freedom (for binary variables). By looking up the standard probability tables for χ^2 distribution, we can obtain a cutoff value for which the independence assumption can be rejected at any significance level, α. For example, the cutoff value for $\alpha = 0.05$ is 3.84.

Example 15 : For the dataset shown in Table 1.4, $\chi^2 = 70$. Since this value is greater than 3.84, we can reject the hypothesis that X and Y are independent of each other at 0.05 significance level.

For the 2×2 contingency table shown in Table 1.2, the χ^2 value can be simplified into following expression:

$$\chi^2_{\text{binary}} = \frac{|T|(f_{11}f_{00} - f_{01}f_{10})^2}{f_{1+}f_{0+}f_{+1}f_{+0}} \tag{1.3}$$

Note the similarity between this expression and Pearson's ϕ coefficient. In fact, these two statistics are related to each other via the following equation :

$$\phi^2 = \frac{\chi^2_{\text{binary}}}{|T|} \tag{1.4}$$

So far, we have shown that both ϕ and χ^2 can be used to estimate the degree of association between two variables. The question is, are these measures sufficient to capture the objective interestingness of a pattern? If it does, then we would not have to rely on other measures such as support or confidence to

[3] This coefficient can be derived from the standard definition for linear correlation coefficient. We will use the terminologies ϕ and correlation coefficients interchangeably.

	$Y = 1$	$Y = 0$	
$X = 1$	50	20	70
$X = 0$	20	10	30
	70	30	100

	$Y = 1$	$Y = 0$	
$X = 1$	10	20	30
$X = 0$	20	50	70
	30	70	100

Table 1.6: Symmetry of the ϕ-coefficients. Both tables have $\phi = 5/21$, but the table on the left should be more interesting it because it has larger support.

determine the interestingness of a pattern. Unfortunately, there are several drawbacks of using χ^2 and the ϕ coefficient. χ^2 is not sufficient because it is meaningless to compare the χ^2 value between different sets of items. A more critical problem is that χ^2 depends on $|T|$. If the size of the database is large enough, many of the discovered patterns will easily pass the χ^2 test. Thus, for large databases, χ^2 may not be useful for filtering patterns. The ϕ coefficient is desirable because it relates to statistical correlation, a widely accepted measure in statistics. However, it is also symmetric in terms of exchanging $f_{11} \leftrightarrow f_{00}$ and $f_{10} \leftrightarrow f_{01}$. This could be a problem because it cannot distinguish between large f_{00} from large f_{11} as shown in Table 1.6. Intuitively, the table on the left should correspond to a more interesting pattern compare to the one on the right, yet their ϕ coefficients stay the same.

Other measures of association from statistical literature include Goodman and Kruskal's λ and τ coefficients, Pearson's coefficient of contingency, uncertainty coefficients, etc. Interested readers should consult references such as [29, 30, 31].

Objective Measures from Data Mining

We now turn to objective measures of interestingness that have been proposed in various knowledge discovery applications. A brief introduction to some of these measures can be found in the survey paper by Hamilton et al. [32]. First, we will describe support and confidence, the two fundamental measures used in Apriori and other association rule mining algorithms.

1. **Support and Confidence**

   Quite often, the support and confidence thresholds are chosen so as to restrict the number of discovered patterns. If the support threshold is too high, it tends to produce patterns that are obvious to most analysts. On the other hand, a low support threshold may results in an unmanageable number of patterns. Furthermore, some of the most interesting patterns may have very low support, e.g. patterns involving expensive jewelries. Currently, there has been some attempts to rectify this problem by using multiple minimum supports[33], assigning weights to different items and generating rules without support pruning.

   Confidence can also be problematic as pointed out in Section 1.4.2. It may produce misleading rules generated from negatively correlated itemsets. There are other confidence-like measures such as the laplace function, which is typically used to measure the expected accuracy of classification rules [34]. The laplace measure for a pair of items is defined as [35] :

$$\text{laplace} = \frac{\sigma(X \cup Y) + 1}{\sigma(X) + 2} \tag{1.5}$$

   This function asymptotically approaches the confidence measure when the support counts for $X \cup Y$ and X are large.

   Example 16 : The laplace measure for the rule $X \Longrightarrow Y$ using the data given in Table 1.4 is $101/142 = .711$.

2. **Interest factor**

Interest factor is another measure of association between items [37, 38, 39]. It is defined as the ratio between the joint probability of two items to their marginal probabilities, i.e:

$$I(X,Y) \quad = \quad \frac{P(X,Y)}{P(X)P(Y)} = \frac{f_{11}N}{f_{1+}f_{+1}} \tag{1.6}$$

This ratio can range anywhere between 0 and $+\infty$. An interest factor of 1 corresponds to total independence between two items. Positively (or negatively) correlated items will have an interest factor greater (less) than 1. A linear transformation of this measure, $I \mapsto I' = I - 1$, obeys all three basic principles of objective measures.

Example 17 : The interest factor for the data given in Table 1.4 is $100 \times 420/180 \times 140 = 1.667$.

Unfortunately this measure can be close to one (independence) even though the two items are highly dependent on each other, as illustrated in the next example.

Example 18 : Consider the following situation. Let $P(X,Y) = 0.1, P(X) = 0.1$ and $P(Y) = 0.1$. Note that the probabilities for each cell in the contingency table can be determined uniquely once these three parameters are given. For example, $P(\neg X, \neg Y)$ must be 0.9 since $P(X, \neg Y) = 0$ and $P(\neg Y) = 1 - P(Y) = 0.9$. The interest factor is large i.e., $1/0.1 = 10$, while its ϕ-coefficient is equal to 1 (perfect positive correlation). Now, if $P(X,Y) = 0.9, P(X) = 0.9$ and $P(Y) = 0.9$, once again the ϕ-coefficient is equal to 1 since both $P(X, \neg Y)$ and $P(\neg X, Y)$ are zero. However, the interest factor has dropped dramatically from 10 to $1/0.9 = 1.11$, which is very close to the independence situation!

The above example reveals that different objective measures can have different views regarding the degree of association between items. If $P(X, \neg Y) = P(\neg X, Y) = 0$, the ϕ-coefficient will always be equal to 1, regardless of the actual probability for $P(X,Y)$ and $P(\neg X, \neg Y)$. Interest factor, on the other hand, prefers items with relatively small support, $P(X,Y) \ll 1$, when both $P(X, \neg Y)$ and $P(\neg X, Y)$ are zero. Let us now investigate the true relationship between these two measures.

For any pair of binary variables, one can show that their interest factor is related to the ϕ-coefficient according to the following equation [40] :

$$\phi \quad = \quad \frac{(I-1) \cdot \sqrt{f_{1+}f_{+1}}}{\sqrt{f_{0+}f_{+0}}} \tag{1.7}$$

Figure 1.10 illustrates such a relationship using a synthetic dataset with 10,000 data points. This dataset is generated by randomly populating a 2×2 contingency table. As mentioned earlier, it is sufficient to generate the three values $P(X,Y), P(X)$ and $P(Y)$, to uniquely determine the rest of the entries in the table. We then plotted a graph of ϕ-coefficient (correlation) versus interest factor for various support thresholds. The first graph (All Pairs) shows the values of ϕ and interest factor for all the data points. The next five graphs illustrate the effect of imposing maximum and minimum support thresholds on $P(X,Y)$. Note that as the range of allowable itemset support decreases, the graph becomes closer to a straight line, indicating that there is a strong correlation between the interest factor and ϕ. Another interesting observation is that as the minimum support threshold increases (for a fixed maximum threshold), we have less itemsets with negative correlation or ϕ. Imposing a higher minimum support threshold also reduces the amount of false positives (i.e. data points with large interest factor but low ϕ coefficient). Meanwhile, an upper support limit can reduce the number of false negatives (i.e, data points with low interest but high correlation). In many practical settings, where $|T|$ is large, we seldom have to worry about specifying an upper support limit because the support for most tems are not too large. If there indeed exists any item with very large support, we can simply remove it from our dataset because patterns involving this item are rarely interesting.

The concept of interest factor can be easily extended beyond two items. For example, the interest factor for a k-itemset can be written as :

$$I(x_1 \cdots x_k) = \frac{P(x_1, \cdots x_k)}{\Pi_{i=1}^{k} P(x_k)}.$$

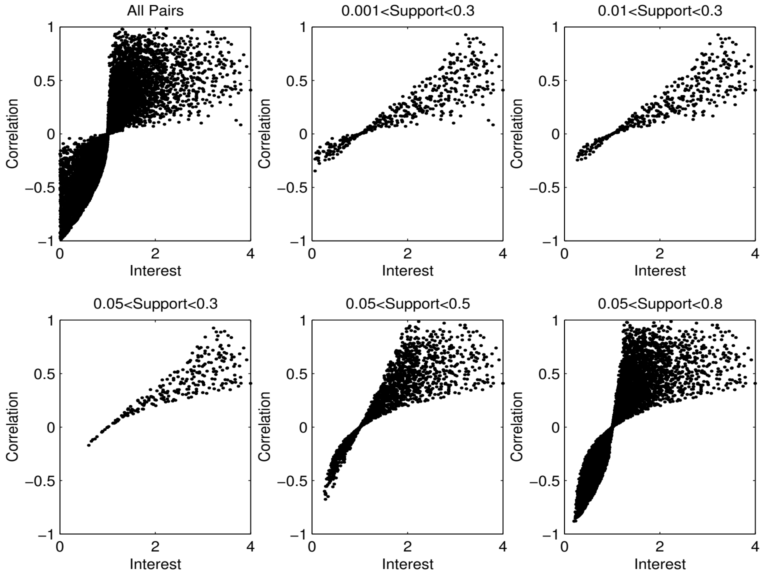

Figure 1.10: ϕ-coefficient versus Interest factor.

3. IS **measure**

Previously, we have shown that if the range of support values is restricted, the interest factor will become increasingly correlated with the ϕ-coefficient. In practice, for large databases, the minimum support threshold is typically small, say around 0.1% or 0.5%. We now examine analytically the behavior of the interest-correlation plot within the restricted range of support values by imposing the following constraints on equation 1.7: $(f_{1+}/|T|) \ll 1, (f_{+1}/|T|) \ll 1$ and $I(X,Y) \gg 1$. The first two are support constraints on items X and Y while the last restriction ensures that we are only dealing with itemsets of high interest. After some manipulation, equation 1.7 becomes :

$$\phi \quad \approx \quad \sqrt{I(X,Y) \times f_{11}} = \sqrt{I(X,Y) \times P(X,Y)} \equiv IS(X,Y) \tag{1.8}$$

We shall define the right hand side of equation 1.8 to be the IS measure. The range of this measure is between 0 and +1 (for pairs of items). Figure 1.11 shows the relationship between the ϕ-coefficient and the IS-measure. The linearity of the graph for the restricted range of support values is not entirely surprising because the IS measure is originally derived from ϕ.

Example 19 : The IS measure for Table 1.4 is $\sqrt{100 \times 100/180/140} = 0.6299$.

Intuitively, IS seems to be a logical choice for an objective measure because it takes into account both the interestingness and support aspects of a rule or itemset. We have previously shown that the ϕ coefficient does not distinguish between the support (f_{11}) and absence (f_{00}) of an itemset. This is not a problem for the IS measure as shown in the next example:

Example 20 : Consider the example given in Table 1.6. Even though both contingency tables have the same ϕ-coefficients, the IS measure for the table on the left is larger (0.7143) than the one on the right (0.3333). This agrees with our observation.

Example 21 : Consider the example in the previous section where $P(X,Y) = 0.1, P(X) = 0.1$ and $P(Y) = 0.1$. Since $I = 10$, its IS measure is $\sqrt{10 \times 0.1} = 1$, which agrees with the ϕ-coefficient. Now, if $P(X,Y) = 0.9, P(X) = 0.9$ and $P(Y) = 0.9$, I will reduce to 10/9. However, the IS measure remains the same, $IS = \sqrt{0.9 \times 10/9} = 1$, which again is consistent with the ϕ-coefficient.

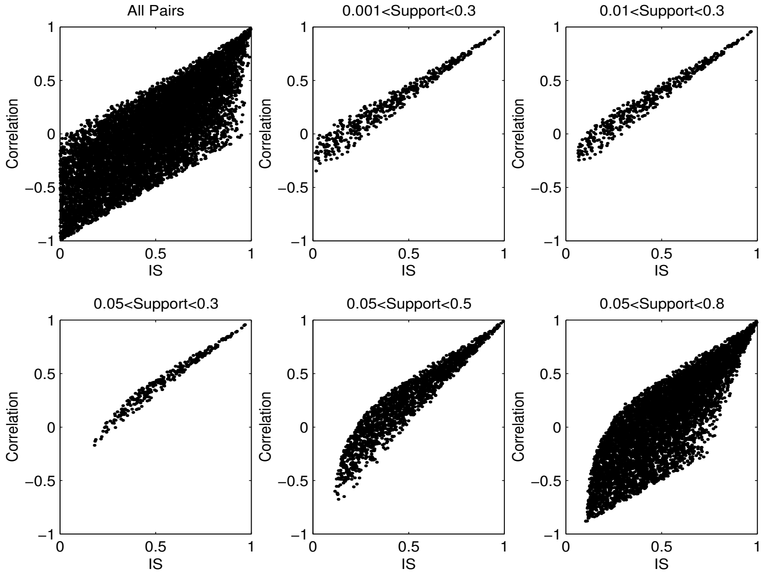

Figure 1.11: ϕ-coefficient versus IS measure.

The previous two examples show the strength of the IS measure in alleviating some of the existing problems with statistical correlation and interest factor. Another advantage of this measure is that it has a nice interpretation especially for pairs of items. For instance, the IS measure can be written in various ways:

$$\begin{aligned} IS(X,Y) &= \sqrt{I(X,Y) \times f_{11}} \\ &= \sqrt{\frac{P(X,Y)}{P(X)} \times \frac{P(X,Y)}{P(Y)}} \\ &= \sqrt{Conf(X \Longrightarrow Y) \times Conf(Y \Longrightarrow X)} \\ &= \frac{P(X,Y)}{\sqrt{P(X)P(Y)}} \end{aligned} \quad (1.9)$$

The third line of equation 1.9 can be interpreted as the geometric mean of confidence of the rules generated from the pair of items. The fourth line can be interpreted as the cosine angle between the vectors X and Y (cosine angles are similarity measures that often used in the text mining domain).

The main disadvantage of this measure is that it does not satisfy the first principle of objective measure since absolute independence occurs when $IS = \sqrt{f_{11}}$. Another problem is how to extend the measure beyond two items. There are various possibilities we can pursue:

$$IS_1(X_1, \cdots, X_k) = \left[\frac{P(X_1, \cdots, X_k)}{P(X_1)} \times \cdots \times \frac{P(X_1, \cdots, X_k)}{P(X_k)}\right]^{\frac{1}{k}} \quad \text{or} \quad (1.10)$$

$$IS_2(X_1, \cdots, X_k) = \sqrt{I(X_1, \cdots, X_k) \times P(X_1, \cdots, X_k)} \quad (1.11)$$

The first approach resembles the geometric mean interpretation and ensures that the range of IS is between 0 and 1. The second approach retains the interest-support product definition but allows the value to be greater than 1.

4. **conviction**

[9] proposed this measure as an alternative to interest factor because they needed an asymmetric

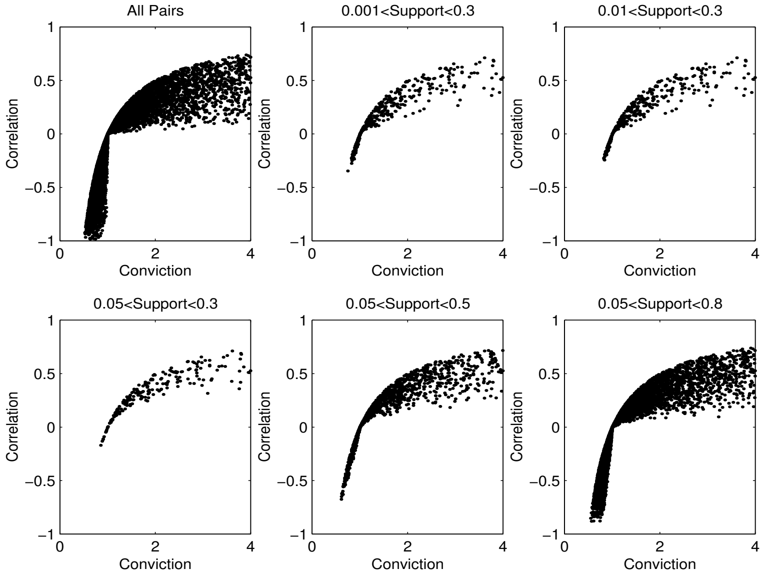

Figure 1.12: ϕ-coefficient versus conviction.

objective measure for implication rules. Conviction is defined to be :

$$\begin{aligned} \text{conviction} &= \frac{P(X)P(\neg Y)}{P(X,\neg Y)} \\ &= \frac{|T| - \sigma(Y)}{|T|(1 - \text{Conf}(X \Longrightarrow Y))} \end{aligned} \quad (1.12)$$

This measure is derived from interest factor in the following way. A rule $X \Longrightarrow Y$ is logically equivalent to $\neg(X \wedge \neg Y)$. Thus, equation 1.12 is an asymmetric way for testing independence between X and $\neg Y$. [4] The ratio between $P(X, \neg Y)$ and $P(X) \times P(\neg Y)$ is inverted due to the negation symbol in the logical expression $\neg(X \wedge \neg Y)$.

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 $P(X,Y)$ is). Figure 1.12 shows the relationship between maximum conviction and the ϕ-coefficient. This is because conviction is an asymmetric measure, thus our plot compares the conviction between the rules $X \Longrightarrow Y$ and $Y \Longrightarrow X$.

Example 22 : Consider the example in Table 1.5. The conviction for the rule $X \Longrightarrow Y$ is $(240 \times 140)/(40 \times 420) = 2$.

If two items are independent, their conviction value will be equal to one. Therefore, the value (conviction - 1) would satisfy the first principle of objective measure. One can verify that conviction also obeys the remaining two principles by looking at the denominator term of equation 1.12.

5. **Piatetsky-Shapiro's rule-interest**

   The rule-interest measure was first introduced by Piatetsky-Shapiro in [23] and shown to satisfy the three basic principles of objective measures[5] :

$$RI = P(XY) - P(X)P(Y) \quad (1.13)$$

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

[5] In this paper, we use a slight variation of Piatetsky-Shapiro's rule-interest function. In the original formulation, $RI = \sigma(X \cup Y) - \sigma(X)\sigma(Y)/|T|$.

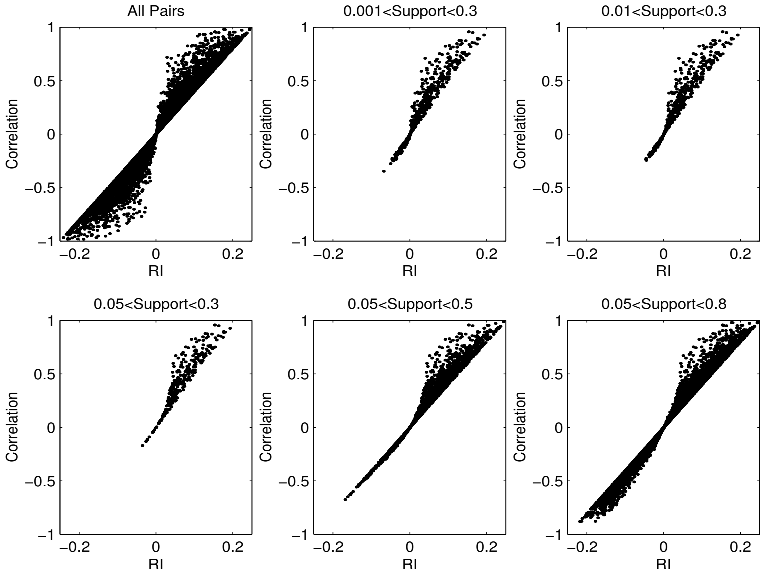

Figure 1.13: ϕ-coefficient versus rule-interest measure.

Example 23 : For the dataset shown in Table 1.5, $RI = 100/420 - 180 * 140/420/420 = 0.095$.

The range of this measure is between -0.25 and 0.25. If X and Y are independent, $RI = 0$. RI is maximum when $P(X,Y) = P(X) = P(Y) = 0.5$. This function resembles the IS measure if we rewrite the above equation into the following form :

$$\begin{aligned} RI &= \left(\frac{P(XY)}{P(X)P(Y)} - 1\right)P(X)P(Y) \\ &= (I-1)P(X)P(Y) \end{aligned} \tag{1.14}$$

Figure 1.13 illustrates the relationship between this measure with the ϕ-coefficient. Unlike the other measures, false positives (i.e. large RI but small ϕ) are not a problem with this measure. By restricting the range of support values, one can reduce the number of negatively correlated patterns.

6. **Gini index**

Gini index [41] is often used as the attribute selection criteria for building inductive decision trees. However, it can also be used to measure the interestingness of discovered rules. For an association rule of the form $X \Longrightarrow Y$:

$$\begin{aligned} Gini = \quad & P(X)(P(Y|X)^2 + P(\neg Y|X)^2)) + P(\neg X)(P(Y|\neg X)^2 + P(\neg Y|\neg X)^2)) \\ & -P(Y)^2 - P(\neg Y)^2 \end{aligned} \tag{1.15}$$

Example 24 : The Gini index for Table 1.5 is $(140/420)((100/140)^2+(40/140)^2) + (280/420)((80/280)^2 + (200/280)^2) - (180/420)^2 - (240/420)^2 = 0.0816$.

The values for the Gini index range between 0 (when the two items are independent) and 0.5 (when the two items are perfectly correlated). Figure 1.14 shows the relationship between the ϕ-coefficient and Gini index for our synthetic dataset. Since the All Pairs graph is symmetric about the horizontal axis, we can conclude that the Gini Index treats both positively and negatively correlated itemsets in the same way. However, as we restrict the range of support values, the negatively correlated pairs are eliminated. This observation suggests that the second principle of objective measure can be violated particularly when one goes from negatively correlated patterns to positively correlated patterns by perturbing the probability entries of the contingency table.

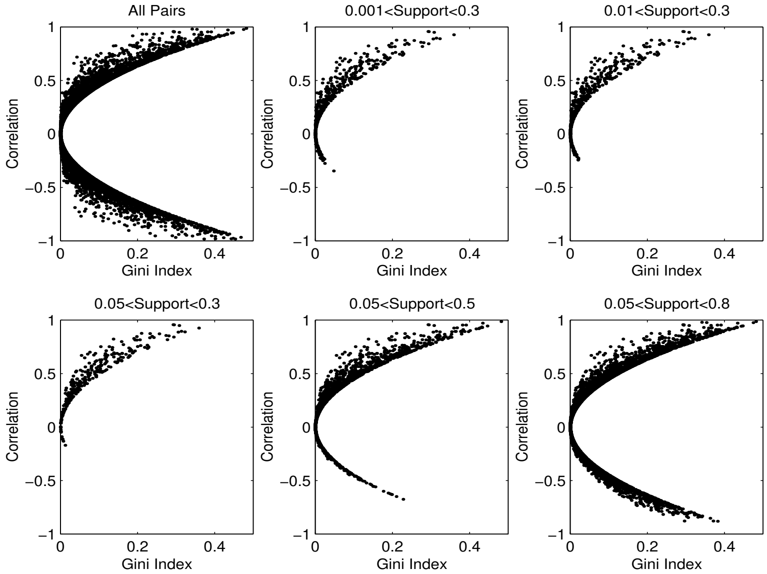

Figure 1.14: ϕ-coefficient versus Gini Index.

7. **Information-based measures**

Another important class of interestingness measures are developed from information theory. In this section, we will describe two such measures: information gain ratio (entropy) and J-measure.

The information gain ratio for a rule $X \Longrightarrow Y$ is given by :

$$GainRatio = \frac{H_X + H_Y - H_{XY}}{H_X} \tag{1.16}$$

where $H(X) = -P(X)\log P(X) - P(\neg X)\log P(\neg X)$, $H(Y) = -P(Y)\log P(Y) - P(\neg Y)\log P(\neg Y)$ and $H(X,Y) = \Sigma_i \Sigma_j P(X=i, Y=j)\log P(X=i, Y=j)$.

The range of this measure is between 0 (for absolute independence) and 1 (for perfect correlation). Figure 1.15 illustrates the relationship between this measure and the ϕ-coefficient.

The J-measure [22] uses information theory to describe the information content of a probabilistic rule. It is defined as :

$$J(X,Y) = P(X)\Big(P(Y|X)\log\big(\frac{P(Y|X)}{P(Y)}\big) + P(\neg Y|X)\log\big(\frac{P(\neg Y|X)}{1-P(Y)}\big)\Big) \tag{1.17}$$

This measure can be interpreted in the following way : the first term $P(X)$ on the right hand side of equation 1.17 requires that the antecedent of a rule must occur relatively often in order for the rule to be interesting. The rest of the right hand side term is a measure of the distance (called cross-entropy) between the posterior and prior beliefs about X. Figure 1.16 illustrates the relationship between J-Measure and the ϕ-coefficient.

Example 25 : For the dataset provided in Table 1.5, the rule $X \Longrightarrow Y$ has an gain ratio of $GainRatio = 0.133$ and J-measure equals to $J(X,Y) = 0.0802$.

Note the symmetry of the two graphs with respect to their horizontal axes (when no support thresholds are specified). This shows that both measures treat positive and negatively correlated items in the same way. As a result, they can violate the second principle of objective measure unless we consider only rules that are positively correlated.

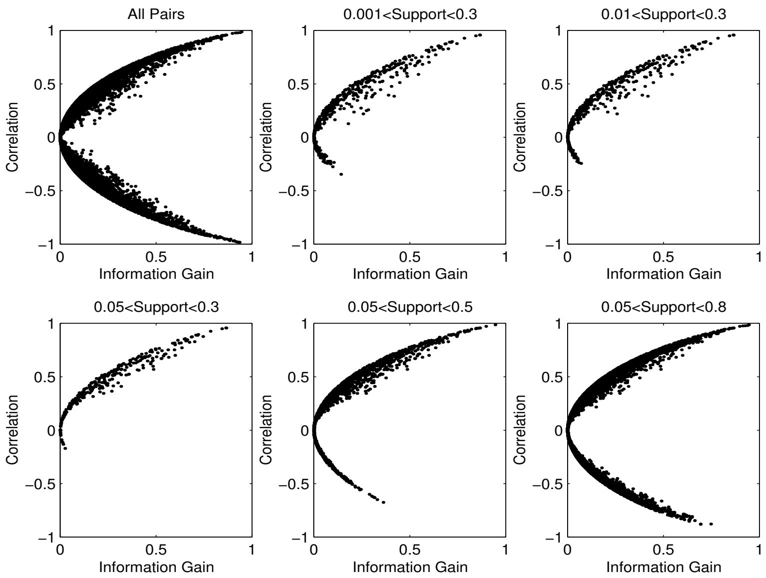

Figure 1.15: ϕ-coefficient versus gain ratio.

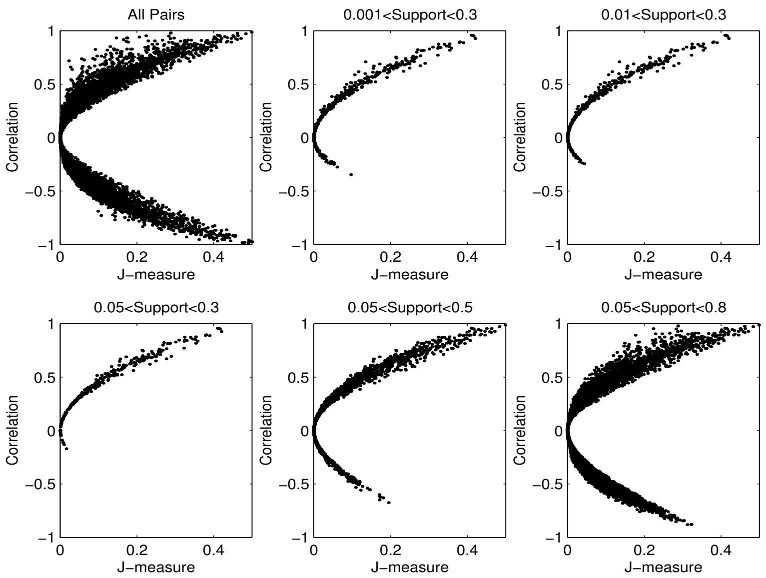

Figure 1.16: ϕ-coefficient versus J-Measure.

Comparison between Objective Measures

So far, we have studied various objective measure of interestingness that have been proposed in statistics, machine learning and data mining literature. All the measures have their own strengths and weaknesses. In summary, we have looked at two important criteria for comparing the various measures.

1. **Compliance with the basic principles of objective measures**

   The three basic principles of objective measures provide a good guideline for evaluating how well an objective measure can be used to represent the interestingness of a pattern. Sometimes, one may want to relax or tighten some of these conditions under certain scenario. For example, critics may argue that the first principle is too rigid. A more flexible principle would be that the interest measure for independent items should be less than the interest measure for positively correlated items. Similarly, the second principle can be violated by objective measures that treat positive and negative correlation equally. One may want to tighten the principle by comparing only patterns containing items that either positively correlated or negatively correlated, but not both.

2. **Agreement with statistical correlation**

   Since the goal of objective measure of interestingness is to determine the degree of association between items in a pattern, it should correspond significantly to statistical measures of association (such as the ϕ-coefficient). In other words, a good objective measure should exhibit high linearity with respect to statistical correlation, especially within the range of support values commonly encountered in practical applications.

In terms of the first criteria, the RI measure, interest factor and conviction are some of the reasonable measures of interestingness. However, upon comparison with statistical measure of correlation, IS is perhaps one of the better choices.

1.5.2 Subjective Measures

Proponents of subjective measures have argued that objective measures alone may not capture all the intricacies of the knowledge discovery process. For example, [42] suggested that the interestingness of a pattern should somehow be related to the estimated utility from taking a specific response action based on the discovered pattern. [43] gave three scenarios in which a rule with high objective measures may not be interesting to data analysts :

- A rule that correspond to prior knowledge, or expectations.
- A rule that refer to uninteresting attributes.
- A redundant rule; ie. a rule that does not add any new predictive power. For example, the rule $X, Y, Z \Longrightarrow C$ may be redundant with respect to $X, Y \Longrightarrow C$.

Principles of Subjective Measures

Silberschatz et al.[24] presented two principles in which a pattern (e.g. rule) can be interesting from a subjective point of view.

- A pattern is interesting if it contradicts the expectation of a user.
- A pattern is interesting if it is actionable.

With this formalism, patterns that are subjectively interesting can appear in three forms: (1) patterns that are both unexpected and actionable, (2) patterns that are unexpected but not actionable, and (3) patterns that are expected but actionable. Even though actionability is a important requirement, in practice, it is difficult to capture this concept formally. However, Silberschatz et al [24] suggested that unexpectedness is a good approximation to actionability because the majority of actionable patterns are unexpected.

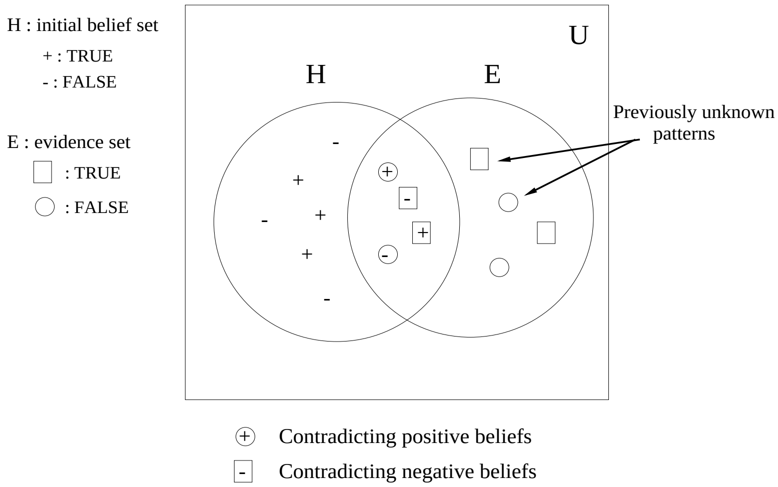

Figure 1.17: Unexpected subjective measure

Subjective measures based on unexpectedness

Unexpectedness of a rule can be modeled by defining the belief set of the problem domain. In [24], beliefs are classified into two categories :

1. **Hard beliefs** are beliefs that can never be changed despite contradictory evidence.
2. **Soft beliefs** are beliefs that a user is willing to compromise if compelling new evidence are found.

Interesting patterns for hard beliefs are patterns that contradict with these beliefs. For soft beliefs, one can define the interestingness measure in terms of how much the existing belief changed as a result of the observed pattern. In [24], the authors suggested several ways to model the soft beliefs using techniques from Bayesian statistics, Dempster-Shafer theory, frequentist approach, Cyc's approach and hypothesis testing methods. However, they assume that the initial belief set of the domain expert encompasses the entire universe of discourse. Liu et al.[25] relaxed this assumption and defined the concept of unexpectedness to include patterns that are previously unknown to the user.

The concept of unexpectedness can be illustrated using the Venn diagram in Figure 1.17. Each element in the Venn diagram represents a logical proposition, which can be true or false depending on the hypothesis used. A hypothesis is a set of propositions, describing either the initial beliefs of a domain expert or experimental evidence collected for a particular domain. H represents the set of initial beliefs (prior knowledge) about the domain. The symbols + (TRUE) and $-$ (FALSE) are used to represent the truth values of each proposition contained in H. Meanwhile, E represents the set of evidence deduced from the discovered patterns. The truth values of each proposition in E are represented by rectangles (TRUE) and circles (FALSE).

Silberschatz et al.[24] defines unexpectedness in terms of evidence that are contradictory to the prior beliefs. This is represented as the contradicting positive beliefs and contradicting negative beliefs in Figure 1.17. On the other hand, Liu et al.[25] extends the notion of unexpectedness to include previously unknown patterns (i.e. propositions that fall in the region E - (E $\cap$ H)).

Cooley et al.[44] proposed a general framework for combining beliefs from the initial belief set (H) and the evidence set (E) using Dempster-Shafer theory of evidence combination. Their framework has been successfully applied to the Web Usage Mining domain.

Similarly, Dong et al.[26] proposed an approach for capturing the unexpectedness of a rule based on its neighborhood information. First, they defined the notion of distance between rules. As a result, rules that are close to one another forms a neighborhood of rules. A rule is interesting if there are unexpected differences between the confidence of the rule with all the other rules in the neighborhood. In [26], unexpected rules are also defined in terms of the density of the neighborhood (i.e. isolated rules are interesting).