Contemporary Mathematics

Objective Measures for Association Pattern Analysis

Michael Steinbach, Pang-Ning Tan, Hui Xiong, and Vipin Kumar

Abstract. Data mining is an area of data analysis that has arisen in response to new data analysis challenges, such as those posed by massive data sets or non-traditional types of data. Association analysis, which seeks to find patterns that describe the relationships of attributes (variables) in a binary data set, is an area of data mining that has created a unique set of data analysis tools and concepts that have been widely employed in business and science. The objective measures used to evaluate the interestingness of association patterns are a key aspect of association analysis. Indeed, different objective measures define different association patterns with different properties and applications. This paper first provides a general discussion of objective measures for assessing the interestingness of association patterns. It then focuses on one of these measures, h-confidence, which is appropriate for binary data sets with skewed distributions. The usefulness of h-confidence and the association pattern that it defines—a hyperclique—is illustrated by an application that involves finding functional modules from protein complex data.

1. Introduction to Association Analysis

Many different types of data analysis techniques have been developed in a wide variety of fields, including mathematics, statistics, machine learning, pattern recognition, and signal processing. Data mining is an area of data analysis that has arisen in response to new data analysis challenges, such as those posed by massive data sets or non-traditional types of data. In some cases, data mining solves current data analysis problems by combining existing data analysis techniques with innovative algorithms. In other cases, new data analysis techniques have been developed. For example, *association analysis*, which seeks to find patterns that describe the relationships of attributes (variables) in a binary data set, is an area of data mining that has created a unique set of data analysis tools and concepts that have been widely employed in both business and science.

2000 *Mathematics Subject Classification.* Primary 62-07, 68P99; Secondary 62P10.

Key words and phrases. data mining, association analysis, hypercliques, bioinformatics.

This work was partially supported by NSF grant #ACI-0325949, NSF grant IIS-0308264, and by the Army High Performance Computing Research Center under the auspices of the Department of the Army, ARL cooperative agreement 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 the AHPCRC and the Minnesota Supercomputing Institute.

©2007 American Mathematical Society

Association analysis [**AIS93, AS94**] analyzes transaction data, such as the data generated when customers purchase items in a store. (The items purchased by a customer are a transaction.) A key task of this analysis is finding *frequent itemsets*, which are sets of items that frequently occur together in a transaction. For example, baby formula and diapers are items that may often be purchased together. The strength of a frequent itemset is measured by its *support*, which is the number (or fraction) of transactions in which all items of the itemset appear together. Another important task of association analysis is the generation of *association rules* [**AS94**], where an association rule is of the form $A \rightarrow B$ (A and B itemsets) and represents the statement that the items of B occur in a transaction that contains the items of A. For instance, the purchase of a toy that does not include batteries often implies the purchase of batteries. The strength of an association rule is measured by the *confidence* of the rule, $\text{conf}(A \rightarrow B)$, which is the fraction of transactions containing all the items of A that also contain all the items of B.

Although the framework for association analysis just described has proved quite useful, support and confidence are only two of several possible objective measures for evaluating association patterns and have well known limitations. Hence, researchers have investigated the utility of a number of other measures for analyzing association patterns [**AY01, BMS97, DP01, HH99, KS96, PS91, RJBA99, TKS04, GH06**]. This paper discusses the importance of objective measures of the interestingness of association patterns and provides an extended discussion of one measure—h-confidence—that has proven particularly useful for finding association patterns in data sets with skewed distributions.

Outline of the Paper Section 2 provides the necessary background by providing a more detailed introduction to the basic concepts of traditional association analysis. Section 3 then introduces the general topic of objective measures, considers the limitations of support and confidence, and discusses alternative objective measures and their properties. One of these measures, h-confidence, and its associated pattern, hypercliques, are then considered in detail in Section 4. This section also presents an application of h-confidence that uses hypercliques to find functional modules from protein interaction data. Section 5 concludes with a summary and a discussion of future work.

2. Background

2.1. Basics. As mentioned earlier, association analysis [**TSK05**] focuses on binary transaction data, such as the data that results when customers purchase items in, for example, a grocery store. Such market basket data can be represented as a collection of transactions, where each transaction corresponds to the items purchased by a specific customer. Table 1 shows an example of a transaction data set.

Alternatively, as is more convenient for the discussion later in this paper, this data can be represented as a binary matrix, where there is one row for each transaction, one column for each item, and the ij^{th} entry is 1 if the i^{th} customer purchased the j^{th} item, and 0 otherwise. Table 2 shows how this data can be represented as a binary matrix. Nonetheless, transaction data is, strictly speaking, a special type of binary data (see [**TSK05**, Chapter 2]). More specifically, traditional association analysis is interested only in the presence, not the absence of an item (the 1's not the 0's). The patterns sought and the data analysis performed also reflects that

TABLE 1. Market basket data.

Transaction ID	Items
1	{Bread, Butter}
2	{Bread, Butter, Diapers, Milk}
3	{Coffee}
4	{Bread, Butter, Coffee, Diapers, Milk}
5	{Bread, Butter}
6	{Diapers, Milk}
7	{Bread, Tea}
8	{Coffee}
9	{Bread, Diapers, Milk}
10	{Tea, Diapers, Milk}

TABLE 2. Binary data matrix for market basket data.

ID	Bread	Butter	Coffee	Diapers	Milk	Tea
1	1	1	0	0	0	0
2	1	1	0	1	1	0
3	0	0	1	0	0	0
4	1	1	1	1	1	0
5	1	1	0	0	0	0
6	0	0	0	1	1	0
7	1	0	0	0	0	1
8	0	0	1	0	0	0
9	1	0	0	1	1	0
10	0	0	0	1	1	1

fact. For other kinds of binary data, such as the results of a true-false test taken by a number of students, 1's and 0's are equally important.

A key task of association analysis is finding *frequent itemsets*, which are sets of items that frequently occur together in a transaction. For example, milk and diapers are items that may often be purchased together. The strength of a frequent itemset is measured by its *support* [**ZO98**], which is the number (or fraction) of transactions in which all items of the itemset appear together. Thus, using either Table 1 or Table 2, the support of the set, $\{milk, diapers\}$ can be found to be 5 (or 0.5 as a fraction). Typically, the most interesting itemsets are those that have relatively high support, although the support threshold that is interesting varies with the data set and application.

Although frequent itemsets are interesting in their own right, the end goal of association analysis is often the efficient generation of *association rules* [**AIS93, AS94**], where an association rule is of the form $A \rightarrow B$ (A and B itemsets) and represents the statement that the items of B occur in a transaction that contains the items of A. The strength of an association rule is measured by the *confidence* of the rule, $\text{conf}(A \rightarrow B)$, which is the fraction of transactions containing all the items of A that also contain all the items of B. This definition of confidence is an estimate of the conditional probability of A given B. Using either Table 1 or Table 2, $\text{conf}(butter \rightarrow bread)$ = (number of times butter and bread occur together)/(number of times butter occurs) = $4/4 = 1$, since bread occurs in every

transaction in which butter occurs. However, conf($bread \rightarrow butter$) = 4/6 = 0.67, because sometimes bread is not purchased with butter.

The rules that have high confidence are typically most interesting because they have high predictive power. Another quantity of interest is the number of transactions for which the rule holds, which is known as the support of the association rule. A rule that holds for many transactions is more likely to be useful than one that holds for just a few transactions, even if the confidence of the rule is 1. As with the support threshold for frequent itemsets, the appropriate values for the support and confidence thresholds of association rules depend on the application and the data.

Another important factor in choosing the thresholds for confidence and support is computational efficiency. Specifically, if n is the number of binary attributes in a transaction data set, there are potentially $2^n - 1$ possible non-empty itemsets. Because transaction data is typically sparse, i.e., mostly 0's, the number of frequent itemsets is far less than $2^n - 1$. However, the actual number depends greatly on the support threshold that is chosen. Likewise, the potential number of association rules is large and is quite sensitive to the thresholds that are chosen for support and confidence. Nonetheless, with judicious choices for support and confidence thresholds, the number of patterns in a data set can be made manageable, and a variety of efficient algorithms have been developed to find frequent itemsets and association rules [**GZ03**].

As a result, association analysis has been very successful. For retail sales data, association analysis has been used for planning product layout, designing marketing campaigns, or managing inventory. Association analysis has also been applied to areas of science, e.g., to analyze Earth science and genomics data [**TSK$^+$01, XHD$^+$05b**]. Furthermore, association analysis has been extended to handle sequential data [**JKK00, TSK05**] and graph data [**KK04, TSK05**]. Algorithms for association analysis are readily available, either in commercial data mining tools [**Int05, Ent05, Cle05, Ins05**] or public domain software packages [**GZ03, R05**]. Thus, association analysis has become a standard technique for data analysis both inside and outside the data mining field.

2.2. A Broader View of Association Patterns and Their Measures. More generally, an itemset pattern or association rule is defined by the measure that is selected to evaluate the strength of the association. Traditionally, support is used to measure the strength of an itemset, while support and confidence are used to measure the strength of an association rule. However, by defining different association measures, it is possible to find different types of association patterns or rules that are appropriate for different types of data and applications. This situation is analogous to that of using different objective functions for measuring the goodness of a set of clusters in order to obtain different types of clusterings for different types of data and applications (see [**TSK05**, Chapter 8]).

Thus, association analysis is fundamentally concerned with defining new association measures. These measures, together with a threshold, select itemsets or rules that are of interest. What might motivate the creation of a new association measure? Most often, the development of new measures is motivated by the limitations of support and/or confidence or the desirable properties of some new measure.

However, besides providing new capabilities, these new association measures must be cognizant of the practical realties addressed by the current association

measures of support and confidence. In particular, two important goals are computational efficiency and distinguishing interesting patterns from spurious ones. As the size and dimensionality of real world databases can be very large, one could easily end up with thousands or even millions of patterns, many of which might not be interesting. It is therefore important to establish the appropriate criteria for evaluating the quality of the derived patterns. There are two criteria often used to prune uninteresting patterns. First, patterns that involve a set of mutually independent items or cover very few transactions are often considered uninteresting. Second, redundant patterns are considered uninteresting because they correspond to sub-patterns of other interesting patterns. In both cases, various *objective interestingness measures* [**TKS04, GH06**] have been proposed to help evaluate the patterns.

The next section presents an overview of such measures. While the focus of that section is on interestingness measures, there are several aspects of pattern evaluation that are not considered. First, patterns can be evaluated through subjective arguments. A pattern is considered subjectively uninteresting unless it reveals unexpected information about the data or provides useful knowledge that can lead to profitable actions. Incorporating subjective knowledge into pattern evaluation requires extensive amount of prior information from domain experts [**GH06**] and thus goes beyond the scope of this paper. Second, pattern evaluation can be complicated by the presence of partial associations among items within the pattern. For example, some relationships may appear or disappear when conditioned upon the value of certain items. This problem is known as *Simpson's paradox* [**FF99**] and goes beyond the scope of this paper. Third, the problem of multiple comparison due to the exploratory nature of the task is not considered in this paper. Interested readers may refer to the references such as [**Web06**] and [**BHA02**].

3. Objective Measures of Interestingness

For simplicity, the discussion in this section focuses primarily on objective measures of interestingness that are for rules or, more generally, pairs of itemsets. Furthermore, only pairs of binary variables are considered. Nonetheless, most of this discussion is relevant to the more general situation.

3.1. Definition of an Objective Measures of Interestingness. An objective measure is a data-driven approach for evaluating the quality of association patterns. It is domain-independent and requires minimal input from the users, other than to specify a threshold for filtering low-quality patterns. An objective measure is usually computed based on the frequency counts tabulated in a **contingency table**. Table 3 shows an example of a contingency table for a pair of binary variables, A and B. We use the notation $\overline{A}$ ($\overline{B}$) to indicate that A (B) is absent from a transaction. Each entry f_{ij} in this 2×2 table denotes a frequency count. For example, f_{11} is the number of times A and B appear together in the same transaction, while f_{01} is the number of transactions that contain B but not A. The row sum f_{1+} represents the support count for A, while the column sum f_{+1} represents the support count for B.

3.2. Limitations of the Support-Confidence Framework. Existing association rule mining formulation relies on the support and confidence measures to

TABLE 3. A 2-way 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

eliminate uninteresting patterns. The drawback of support was that many potentially interesting patterns involving low support items might be eliminated by the support threshold. The drawback of confidence is more subtle and is best demonstrated with the following example from Brin, Motwani, and Silverstein [**BMS97**].

EXAMPLE 3.1. Suppose we are interested in analyzing the relationship between people who drink tea and coffee. We may gather information about the beverage preferences among a group of people and summarize their responses into a table such as the one shown in Table 4.

TABLE 4. Beverage preferences among a group of 1000 people.

	$Coffee$	$\overline{Coffee}$	
Tea	150	50	200
$\overline{Tea}$	650	150	800
	800	200	1000

The information given in this table can be used to evaluate the association rule $\{Tea\} \longrightarrow \{Coffee\}$. At first glance, it may appear that people who drink tea also tend to drink coffee because the rule's support (15%) and confidence (75%) values are reasonably high. This argument would have been acceptable except that the fraction of people who drink coffee, regardless of whether they drink tea, is 80%, while the fraction of tea drinkers who drink coffee is only 75%. Thus knowing that a person is a tea drinker actually decreases her probability of being a coffee drinker from 80% to 75%! The rule $\{Tea\} \longrightarrow \{Coffee\}$ is therefore misleading despite its high confidence value.

The pitfall of confidence can be traced to the fact that the measure ignores the support of the itemset in the rule consequent. Indeed, if the support of coffee drinkers is taken into account, we would not be surprised to find that many of the people who drink tea also drink coffee. What is more surprising is that the fraction of tea drinkers who drink coffee is actually less than the overall fraction of people who drink coffee, which points to an inverse relationship between tea drinkers and coffee drinkers.

3.3. Alternative Objective Interestingness Measures. Because of the limitations in the support-confidence framework, various alternative measures have been used to evaluate the quality of association patterns. Table 5 provides the definitions for some of these measures in terms of the frequency counts of a 2×2 contingency table.

Given the wide variety of measures available, it is reasonable to question whether the measures can produce similar ordering results when applied to a set of association patterns. If the measures are consistent, then we can choose any one of them as our evaluation metric. Otherwise, it is important to understand what their differences are in order to determine which measure is more suitable for analyzing certain types of patterns.

Suppose we apply the measures to rank the ten contingency tables shown in Table 6. These contingency tables are chosen to illustrate the differences among the existing measures. The orderings produced by these measures are shown in Table 7 (with 1 as the most interesting and 10 as the least interesting table). Although

TABLE 5. Objective measures for association patterns.

Measure (Symbol)	Definition
Correlation (ϕ)	$\frac{Nf_{11}-f_{1+}f_{+1}}{\sqrt{f_{1+}f_{+1}f_{0+}f_{+0}}}$
Odds ratio (α)	$\big(f_{11}f_{00}\big)/\big(f_{10}f_{01}\big)$
Kappa (κ)	$\frac{Nf_{11}+Nf_{00}-f_{1+}f_{+1}-f_{0+}f_{+0}}{N^2-f_{1+}f_{+1}-f_{0+}f_{+0}}$
Interest (I)	$\big(Nf_{11}\big)/\big(f_{1+}f_{+1}\big)$
Cosine (IS)	$\big(f_{11}\big)/\big(\sqrt{f_{1+}f_{+1}}\big)$
Piatetsky-Shapiro (PS)	$\frac{f_{11}}{N}-\frac{f_{1+}f_{+1}}{N^2}$
Collective strength (S)	$\frac{f_{11}+f_{00}}{f_{1+}f_{+1}+f_{0+}f_{+0}}\times\frac{N-f_{1+}f_{+1}-f_{0+}f_{+0}}{N-f_{11}-f_{00}}$
Jaccard (ζ)	$f_{11}/\big(f_{1+}+f_{+1}-f_{11}\big)$
All-confidence (h)	$\min\big[\frac{f_{11}}{f_{1+}},\frac{f_{11}}{f_{+1}}\big]$
Goodman-Kruskal (λ)	$\left[\frac{\sum_j \max_k f_{jk}+\sum_k \max_j f_{jk}-\max_j f_{j+}-max_k f_{+k}}{2N-\max_j f_{j+}-\max_k f_{+k}}\right]$
Mutual Information (M)	$\frac{\sum_i\sum_j \frac{f_{ij}}{N}\log\frac{Nf_{ij}}{f_{i+}f_{+j}}}{\min\left[-\sum_i \frac{f_{i+}}{N}\log\frac{f_{i+}}{N},-\sum_j \frac{f_{+j}}{N}\log\frac{f_{+j}}{N}\right]}$
J-Measure (J)	$\frac{f_{11}}{N}\log\frac{Nf_{11}}{f_{1+}f_{+1}}+\max\left[\frac{f_{10}}{N}\log\frac{Nf_{10}}{f_{1+}f_{+0}},\frac{f_{01}}{N}\log\frac{Nf_{01}}{f_{0+}f_{+1}}\right]$
Gini index (G)	$\max\left[\frac{f_{1+}}{N}\times[(\frac{f_{11}}{f_{1+}})^2+(\frac{f_{10}}{f_{1+}})^2]+\frac{f_{0+}}{N}\times[(\frac{f_{01}}{f_{0+}})^2+(\frac{f_{00}}{f_{0+}})^2]-(\frac{f_{+1}}{N})^2-(\frac{f_{+0}}{N})^2,\ \frac{f_{+1}}{N}\times[(\frac{f_{11}}{f_{+1}})^2+(\frac{f_{01}}{f_{+1}})^2]+\frac{f_{+0}}{N}\times[(\frac{f_{10}}{f_{+0}})^2+(\frac{f_{00}}{f_{+0}})^2]-(\frac{f_{1+}}{N})^2-(\frac{f_{0+}}{N})^2\right]$
Laplace (L)	$\max\left[\frac{f_{11}+1}{f_{1+}+2},\frac{f_{11}+1}{f_{+1}+2}\right]$
Conviction (V)	$\max\left[\frac{f_{1+}f_{+0}}{Nf_{10}},\frac{f_{0+}f_{+1}}{Nf_{01}}\right]$
Certainty factor (F)	$\max\left[\frac{\frac{f_{11}}{f_{1+}}-\frac{f_{+1}}{N}}{1-\frac{f_{+1}}{N}},\frac{\frac{f_{11}}{f_{+1}}-\frac{f_{1+}}{N}}{1-\frac{f_{1+}}{N}}\right]$
Added Value (AV)	$\max\left[\frac{f_{11}}{f_{1+}}-\frac{f_{+1}}{N},\frac{f_{11}}{f_{+1}}-\frac{f_{1+}}{N}\right]$

TABLE 6. Example of contingency tables.

Example	f_{11}	f_{10}	f_{01}	f_{00}
E_1	8123	83	424	1370
E_2	8330	2	622	1046
E_3	3954	3080	5	2961
E_4	2886	1363	1320	4431
E_5	1500	2000	500	6000
E_6	4000	2000	1000	3000
E_7	9481	298	127	94
E_8	4000	2000	2000	2000
E_9	7450	2483	4	63
E_{10}	61	2483	4	7452

TABLE 7. Rankings of contingency tables using the measures given in Table 5.

	ϕ	α	κ	I	IS	PS	S	ζ	h	λ	M	J	G	L	V	F	AV
E_1	1	3	1	6	2	2	1	2	2	1	1	1	1	4	2	2	5
E_2	2	1	2	7	3	5	2	3	3	2	2	2	3	5	1	1	6
E_3	3	2	4	4	5	1	3	6	8	5	3	5	2	2	6	6	4
E_4	4	8	3	3	7	3	4	7	5	4	6	3	4	9	3	3	1
E_5	5	7	6	2	9	6	6	9	9	9	7	4	6	8	5	5	2
E_6	6	9	5	5	6	4	5	5	7	3	8	6	5	7	4	4	3
E_7	7	6	7	9	1	8	7	1	1	7	5	9	8	3	7	7	9
E_8	8	10	8	8	8	7	8	8	7	8	9	7	7	10	8	8	7
E_9	9	4	9	10	4	9	9	4	4	6	4	10	9	1	9	9	10
E_{10}	10	5	10	1	10	10	10	10	10	10	10	8	10	6	10	10	8

some of the measures appear to be consistent with each other, there are certain measures that produce quite different ordering results. For example, the rankings given by the ϕ-coefficient tend to agree with those provided by κ and collective strength, but are quite different than the rankings produced by interest factor and odds ratio. Furthermore, a contingency table such as E_{10} is ranked lowest according to the ϕ-coefficient, but highest according to interest factor.

In general, measures that were developed for different applications and types of data often give quite different results (ranking) and the choice of the proper association measure depends on the nature of the application and the data type.

3.4. Properties of Objective Measures. The results shown in Table 7 suggest that a significant number of the measures provide conflicting information about the quality of a pattern. To understand their differences, we need to examine the properties of these measures. The following is a summary three important properties. For a more comprehensive list of properties, readers should refer to [**TKS02, TKS04, HH99, GH06**].

3.4.1. *Inversion Property.* Consider the bit vectors shown in Figure 1. The 0/1 bit in each column vector indicates whether a transaction (row) contains a particular item (column). For example, the vector **A** indicates that item a belongs to the first and last transactions, whereas the vector **B** indicates that item b is

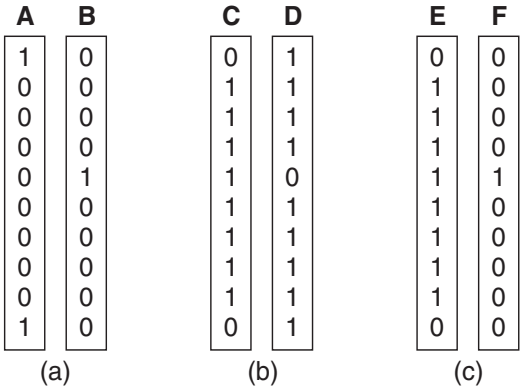

FIGURE 1. Effect of the inversion operation. The vectors C and E are inversions of vector A, while the vector D is an inversion of vectors B and F.

contained only in the fifth transaction. The vectors **C** and **E** are in fact related to the vector **A**—their bits have been inverted from 0's (absence) to 1's (presence), and vice versa. Similarly, **D** is related to vectors **B** and **F** by inverting their bits. The process of flipping a bit vector is called **inversion**. If a measure is invariant under the inversion operation, then its value for the vector pair $(\mathbf{C}, \mathbf{D})$ should be identical to its value for $(\mathbf{A}, \mathbf{B})$. The inversion property of a measure can be stated as follows.

DEFINITION 3.2 (Inversion Property). An objective measure M is invariant under the inversion operation if its value remains the same when exchanging the frequency counts f_{11} with f_{00} and f_{10} with f_{01}.

Among the measures that remain invariant under this operation include the ϕ-coefficient, odds ratio, κ, and collective strength. These measures may not be suitable for analyzing asymmetric binary data. For example, the ϕ-coefficient between **C** and **D** is identical to the ϕ-coefficient between **A** and **B**, even though items c and d appear together more frequently than a and b. Furthermore, the ϕ-coefficient between **C** and **D** is less than that between **E** and **F** even though items e and f appear together only once! For asymmetric binary data, measures that do not remain invariant under the inversion operation are preferred [**HSM01**]. Some of the non-invariant measures include interest factor, IS, PS, and the Jaccard coefficient.

3.4.2. *Null Addition Property.* Suppose we are interested in analyzing the relationship between a pair of words, such as `data` and `mining`, in a set of documents. If a collection of articles about ice fishing is added to the data set, then one would expect that the association between `data` and `mining` to remain unchanged. This process of adding unrelated data (in this case, documents) to a given data set is known as the **null addition** operation.

DEFINITION 3.3 (Null Addition Property). An objective measure M is invariant under the null addition operation if its value does not change when f_{00} is increased, while all other frequencies in the contingency table stay the same.

For applications such as document analysis or market basket analysis, the measure is expected to remain invariant under the null addition operation. Otherwise, the relationship between words may disappear simply by adding enough documents that do not contain both words! Examples of measures that satisfy this property include cosine (IS) and Jaccard (ξ) measures, while those that violate this property include interest factor, PS, odds ratio, and the ϕ-coefficient.

3.4.3. *Scaling Property.* Table 8 shows the contingency tables for gender and the grades achieved by students enrolled in a particular course in 1993 and 2004. This example is inspired by Mosteller [**Mos68**]. The data in these tables showed that the number of male students has doubled since 1993, while the number of female students has increased by a factor of 3. The correlation between grade and gender in both tables are different. However, the male students in 2004 are not performing any better than those in 1993 because the ratio of male students who achieve a high grade to those who achieve a low grade is still the same, i.e., 3:4. Similarly, the female students in 2004 are performing no better than those in 1993. According to Mosteller's analysis method, 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.

TABLE 8. The grade-gender example.

	Male	Female	
High	30	20	50
Low	40	10	50
	70	30	100

(a) Sample data from 1993.

	Male	Female	
High	60	60	120
Low	80	30	110
	140	90	230

(b) Sample data from 2004.

DEFINITION 3.4 (Scaling Invariance Property). An objective measure M is invariant under the row/column scaling operation if $M(T) = M(T')$, where T is a contingency table with frequency counts $[f_{11};\ f_{10};\ f_{01};\ f_{00}]$, T' is a contingency table with scaled frequency counts $[k_1k_3f_{11};\ k_2k_3f_{10};\ k_1k_4f_{01};\ k_2k_4f_{00}]$, and k_1, k_2, k_3, k_4 are positive constants.

From Table 9, notice that only the odds ratio (α) is invariant under the row and column scaling operations. All other measures such as the ϕ-coefficient, κ, IS, interest factor, and collective strength (S) change their values when the rows and columns of the contingency table are rescaled.

4. An Objective Measure for Skewed Support Distributions

The performances of many association analysis algorithms are influenced by properties of their input data. For example, the computational complexity of the *Apriori* algorithm depends on properties such as the number of items in the data and average transaction width. This section examines another important property that has significant influence on the performance of association analysis algorithms as well as the quality of extracted patterns. More specifically, we focus on data sets with skewed support distributions, where most of the items have relatively low to moderate frequencies, but a small number of them have very high frequencies.

TABLE 9. Properties of association measures.

Symbol	Measure	Inversion	Null Addition	Scaling
ϕ	ϕ-coefficient	Yes	No	No
α	odds ratio	Yes	No	Yes
κ	Cohen's	Yes	No	No
I	Interest	No	No	No
IS	Cosine	No	Yes	No
PS	Piatetsky-Shapiro's	Yes	No	No
S	Collective strength	Yes	No	No
ζ	Jaccard	No	Yes	No
h	All-confidence	No	Yes	No
s	Support	No	No	No
λ	Goodman-Kruskal	Yes	No	No
M	Mutual Information	Yes	No	No
J	J-Measure	No	No	No
G	Gini index	Yes	No	No
L	Laplace	No	No	No
V	Conviction	Yes	No	No
F	Certainty factor	Yes	No	No
AV	Added value	No	No	No

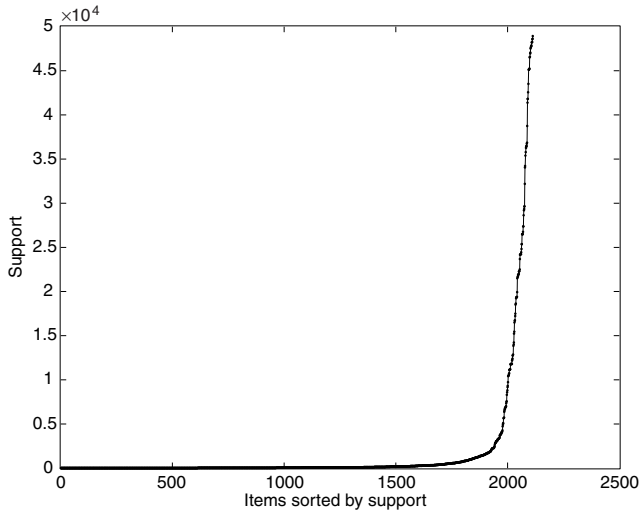

FIGURE 2. Support distribution of items in the census data set.

A measure, h-confidence, which performs well in the presence of skewed support is introduced and its properties are described. An example of its usefulness on real biological data is also presented.

4.1. The Effect of a Skewed Support Distribution. An example of a real data set that exhibits a skewed support distribution is shown in Figure 2. The data, taken from the PUMS (Public Use Microdata Sample) census data, contains 49,046 records and 2113 asymmetric binary variables. While more than 80% of the items have support less than 1%, a handful of them have support greater than 90%.

TABLE 10. Grouping the items in the census data set based on their support values.

Group	G_1	G_2	G_3
Support	$< 1\%$	$1\% - 90\%$	$> 90\%$
Number of Items	1735	358	20

To illustrate the effect of skewed support distribution on frequent itemset mining, we divide the items into three groups, G_1, G_2, and G_3, according to their support levels. The number of items that belong to each group is shown in Table 10.

Choosing the right support threshold for mining this data set can be quite tricky. If we set the threshold too high (e.g., 20%), then we may miss many interesting patterns involving the low support items from G_1. In market basket analysis, such low support items may correspond to expensive products (such as jewelry) that are seldom bought by customers, but whose patterns are still interesting to retailers. Conversely, when the threshold is set too low, it becomes difficult to find the association patterns due to the following reasons. First, the computational and memory requirements of existing association analysis algorithms increase considerably with low support thresholds. Second, the number of extracted patterns also increases substantially, many of which relate a high-frequency item such as milk to a low-frequency item such as caviar. Such patterns, which are called **cross-support** patterns, are likely to be spurious. For example, at a support threshold equal to 0.05%, there are 18,847 frequent pairs involving items from G_1 or G_3, or both. Out of these, 93% of them are cross-support patterns; i.e., the patterns contain items from both G_1 and G_3. The maximum correlation obtained from the cross-support patterns is 0.029, which is much lower than the maximum correlation obtained from frequent patterns involving items from the same group (which is as high as 1.0). Similar statement can be made about many other interestingness measures discussed in the previous section. This example shows that a large number of weakly correlated cross-support patterns can be generated when the support threshold is sufficiently low. Before defining an association measure that can eliminate such patterns, we formally define the concept of cross-support patterns.

DEFINITION 4.1 (Cross-Support Pattern). A cross-support pattern is an itemset $X = \{i_1, i_2, \ldots, i_k\}$ whose support ratio

$$r(X) = \frac{\min\left[s(i_1), s(i_2), \ldots, s(i_k)\right]}{\max\left[s(i_1), s(i_2), \ldots, s(i_k)\right]}, \tag{1}$$

is less than a user-specified threshold h_c.

EXAMPLE 4.2. Suppose the support for milk is 70%, while the support for sugar is 10% and caviar is 0.04%. Given $h_c = 0.01$, the frequent itemset {milk, sugar, caviar} is a cross-support pattern because its support ratio is

$$r = \frac{\min\left[0.7, 0.1, 0.0004\right]}{\max\left[0.7, 0.1, 0.0004\right]} = \frac{0.0004}{0.7} = 0.00058 < 0.01.$$

Existing measures such as support and confidence are not sufficient to eliminate cross-support patterns, as illustrated by the data set shown in Figure 3. Assuming that $h_c = 0.3$, the itemsets $\{p, q\}$, $\{p, r\}$, and $\{p, q, r\}$ are cross-support patterns because their support ratios, which are equal to 0.2, are less than the threshold

p	q	r
0	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
0	0	0
0	0	0
0	0	0
0	0	0

FIGURE 3. A transaction data set containing three items, p, q, and r, where p is a high support item and q and r are low support items.

h_c. Although we can apply a high support threshold, say, 20%, to eliminate the cross-support patterns, this may come at the expense of discarding other interesting patterns such as the strongly correlated itemset, $\{q, r\}$ that has support equal to 16.7%.

Confidence pruning also does not help because the confidence of the rules extracted from cross-support patterns can be very high. For example, the confidence for $\{q\} \longrightarrow \{p\}$ is 80% even though $\{p, q\}$ is a cross-support pattern. The fact that the cross-support pattern can produce a high-confidence rule should not come as a surprise because one of its items (p) appears very frequently in the data. Therefore, p is expected to appear in many of the transactions that contain q. Meanwhile, the rule $\{q\} \longrightarrow \{r\}$ also has high confidence even though $\{q, r\}$ is not a cross-support pattern. This example demonstrates the difficulty of using the confidence measure to distinguish between rules extracted from cross-support and non-cross-support patterns.

4.2. Definition of H-confidence and Hypercliques. Returning to the previous example, notice that the rule $\{p\} \longrightarrow \{q\}$ has very low confidence because most of the transactions that contain p do not contain q. In contrast, the rule $\{r\} \longrightarrow \{q\}$, which is derived from the pattern $\{q, r\}$, has very high confidence. This observation suggests that cross-support patterns can be detected by examining the lowest confidence rule that can be extracted from a given itemset. The proof of this statement can be understood as follows.

(1) Note the following anti-monotone property of confidence:

$$conf(\{i_1 i_2\} \longrightarrow \{i_3, i_4, \ldots, i_k\}) \leq conf(\{i_1 i_2 i_3\} \longrightarrow \{i_4, i_5, \ldots, i_k\}).$$

This property suggests that confidence never increases as we shift more items from the left- to the right-hand side of an association rule. Because of this property, the lowest confidence rule extracted from a frequent itemset contains only one item on its left-hand side. We denote the set of all rules with only one item on its left-hand side as R_1.

(2) Given a frequent itemset $\{i_1, i_2, \ldots, i_k\}$, the rule

$$\{i_j\} \longrightarrow \{i_1, i_2, \ldots, i_{j-1}, i_{j+1}, \ldots, i_k\}$$

has the lowest confidence in R_1 if $s(i_j) = \max\left[s(i_1), s(i_2), \ldots, s(i_k)\right]$. This follows directly from the definition of confidence as the ratio between the rule's support and the support of the rule antecedent.

Thus, for a set of items $\{i_1, i_2, \ldots, i_k\}$, we can define a new measure known as the **h-confidence** [**XTK06**], which is the lowest confidence attainable from that itemset.

DEFINITION 4.3. The **h-confidence** of an itemset $X = \{i_1, i_2, \cdots, i_m\}$, denoted as $hconf(X)$, is a measure that reflects the overall affinity among items within the itemset. This measure is defined as

$$min\{conf\{i_1 \rightarrow i_2, \ldots, i_m\}, conf\{i_2 \rightarrow i_1, i_3, \ldots, i_m\}, \ldots, conf\{i_m \rightarrow i_1, \ldots, i_{m-1}\}\},$$

where $conf$ is the conventional definition of association rule confidence.

Note that h-confidence can also be expressed as

$$\frac{s(\{i_1, i_2, \ldots, i_k\})}{\max\left[s(i_1), s(i_2), \ldots, s(i_k)\right]}.$$

Furthermore, h-confidence is equivalent to the **all-confidence** measure defined by Omiecinski [**Omi03**].

$$\min\{\text{conf}(X_1 \rightarrow X_2)\} \text{ for every } X_1, X_2 \text{ such that } X_1 \cup X_2 = X, X_1 \cap X_2 = \phi.$$

4.3. Illustration of H-Confidence. Consider an itemset $X = \{i_1, i_2, i_3\}$. Assume that $supp(\{i_1\}) = 0.1$, $supp(\{i_2\}) = 0.1$, $supp(\{i_3\}) = 0.06$, and $supp(\{i_1, i_2, i_3\}) = 0.06$, where $supp$ is the support of an itemset. Then

$$\begin{aligned} conf\{i_1 \rightarrow i_2, i_3\} &= supp(\{i_1, i_2, i_3\})/supp(\{i_1\}) = 0.6 \\ conf\{i_2 \rightarrow i_1, i_3\} &= supp(\{i_1, i_2, i_3\})/supp(\{i_2\}) = 0.6 \\ conf\{i_3 \rightarrow i_1, i_2\} &= supp(\{i_1, i_2, i_3\})/supp(\{i_3\}) = 1 \end{aligned}$$

Hence, $hconf(X) = min\{conf\{i_2 \rightarrow i_1, i_3\}, conf\{i_1 \rightarrow i_2, i_3\}, conf\ \{i_3 \rightarrow i_1, i_2\}\}$ $= 0.6$.

4.4. Properties of the H-confidence measure. The h-confidence measure has three important properties, namely the anti-monotone property, the cross-support property, and the strong affinity property. Detailed descriptions of these three properties were provided in our earlier paper [**XTK06**]. Here, we provide only the following brief summaries.

The anti-monotone property. The h-confidence measure is anti-monotone, i.e.,

$$\text{h-confidence}(\{i_1, i_2, \ldots, i_k\}) \geq \text{h-confidence}(\{i_1, i_2, \ldots, i_{k+1}\}).$$

This property is analogous to the anti-monotone property of the support measure used in association-rule mining [**AIS93**] and allows us to use h-confidence-based pruning to speed the search for hyperclique patterns in the same way that support-based pruning is used to speed the search for frequent itemsets.

The cross-support property. Because of the anti-monotone property of support, the numerator of the h-confidence measure is bounded by the minimum support of any item that appears in the frequent itemset. In other words, the h-confidence of an itemset $X = \{i_1, i_2, \ldots, i_k\}$ must not exceed the following expression:

$$\text{h-confidence}(X) \leq \frac{\min\left[s(i_1), s(i_2), \ldots, s(i_k)\right]}{\max\left[s(i_1), s(i_2), \ldots, s(i_k)\right]}.$$

Note the equivalence between the upper bound of h-confidence and the support ratio (r) given in Equation 1. Because the support ratio for a cross-support pattern is always less than h_c, the h-confidence of the pattern is also guaranteed to be less than h_c. Therefore, cross-support patterns can be eliminated by ensuring that the h-confidence values for the patterns exceed h_c. The computation of this upper bound is much cheaper than the computation of the exact h-confidence value, since the it only relies on the support values of individual items in the itemset. Thus, using the cross-support property, we can design a partition-based approach that allows us to efficiently eliminate patterns involving items with different support levels.

The strong affinity property. H-confidence ensures that the items contained in an itemset are strongly associated with each other. For example, suppose the h-confidence of an itemset X is 80%. If one of the items in X is present in a transaction, there is at least an 80% chance that the rest of the items in X also belong to the same transaction. This property can also be stated in terms of similarity. For instance, the strong affinity property guarantees that if an itemset has an h-confidence value of h_c, then every pair of items within the hyperclique pattern must have a cosine similarity greater than or equal to h_c. A similar result can be proved for the Jaccard coefficient. The overall affinity of hyperclique patterns can be controlled by setting an h-confidence threshold.

As demonstrated in our previous paper [**XTK06**], the anti-monotone and cross-support properties form the basis of an efficient hyperclique mining algorithm that has much better performance than frequent itemset mining algorithms, particularly at low levels of support. Also, the number of hyperclique patterns is significantly less than the number of frequent itemsets.

4.5. Applications of H-ypercliques. The hyperclique pattern has been shown to be useful for various applications, including clustering [**XSTK04**], semi-supervised classification [**XSK05**], data cleaning [**XPSK06**], and finding functionally coherent sets of proteins [**XHD$^+$05a**]. We describe this last application in more detail.

The fact that hypercliques can find relatively pure patterns that have low support in the presence of noise has been used to analyze protein interaction networks

[**XHD$^+$05a**]. For this particular analysis, the data consisted of 252 protein complexes, where each protein complex was a collection of proteins. Functional modules are groups of proteins that occur in more than one protein complex and represent groups of proteins that, informally, share a similar function or belong to the same cellular process. Thus, if the protein complexes are taken as transactions and the proteins as items, this problem becomes one of finding itemsets. In this domain, we want to avoid cross support patterns that could result when a frequently occurring protein is included in an itemset (functional module) simply because of its frequency and not because it works with other proteins to perform some biologically meaningful task. Since hypercliques do not contain such cross support patterns unless the h-confidence threshold is low, hypercliques were considered as candidates for groups of proteins (itemsets) that could be functional modules.[1]

The results were evaluated using the Gene Ontology (GO) [**GO 06**], which is a set of three separate hierarchies that impose a hierarchical set of terms on biological functions, processes, and components. For example, the function of a protein is a set of terms that start at the root of the function hierarchy and proceeds down the tree to the most specific function (or functions) known for the protein. Since a protein can have multiple functions or participate in multiple processes or be a part of multiple components, the function, process, or component characteristics of a protein is often expressed as a subtree of the function, process, or component hierarchies. Indeed, the description of a set of proteins is often visualized as a subtree of one of these hierarchies. If the proteins are concentrated at only a single leaf of the tree, then the group of proteins are strongly related. We will consider only the process hierarchy in this discussion.

The analysis of the hyperclique results using GO, was quite encouraging. For most patterns, many of the proteins in a hyperclique (candidate functional module) were concentrated mostly at one leaf of the function or process tree. We have only included a couple of results here, but many of the results can be viewed online at [**Xio05**]. Additional details are available in [**XHD$^+$05b**].

Figure 4 is for the hyperclique pattern containing eight proteins—Pre2 Pre4 Pre5 Pre8 Pup3 Pre6 Pre9 Scl1. As is shown in Figure 4, all of these proteins share the same leaf node in the process tree, *ubiquitin dependent protein catabolic process*, and thus, form a coherent group from a biological perspective. Figure 5 is for the hyperclique pattern containing seven proteins—Clf1 Lea1 Rse1 YLR424W Prp46 Smd2 Snu114. Although this figure is more complicated, all the proteins share the function, *nuclear mRNA splicing, via spliceosome*. The additional complexity of the figure comes from the fact that some proteins in the group have additional functions. Note that these figures were produced using the Saccharomyces Genome Database (SGD) GO Gene Finder [**SGD07**].

[1]Actually, closed hypercliques were used. Closed hypercliques are sets of items for which there exists no larger set of hypercliques that contains the original set of items and has the same h-confidence.

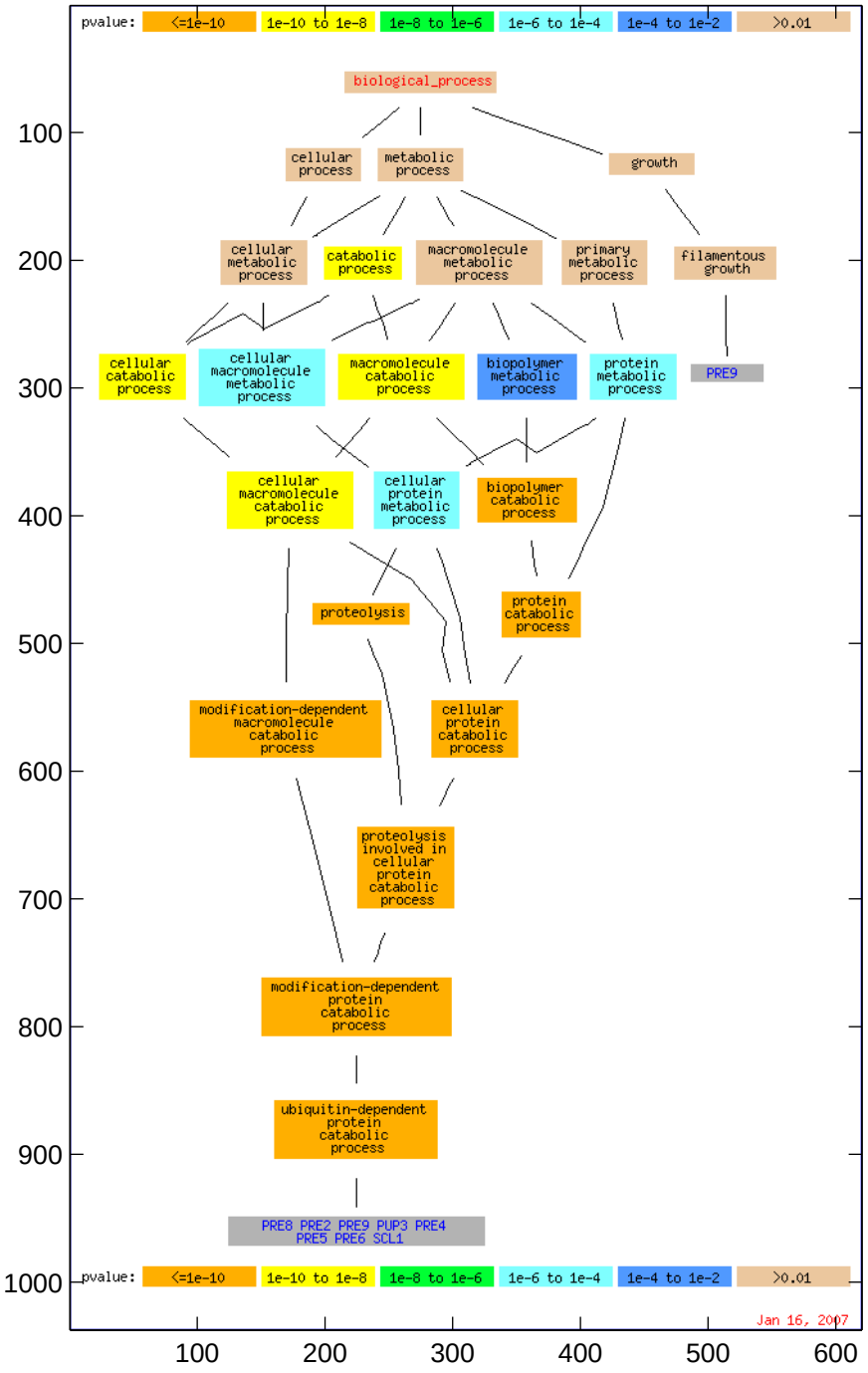

FIGURE 4. The process hierarchy tree for the hyperclique containing {Pre2 Pre4 Pre5 Pre8 Pup3 Pre6 Pre9 Scl1}

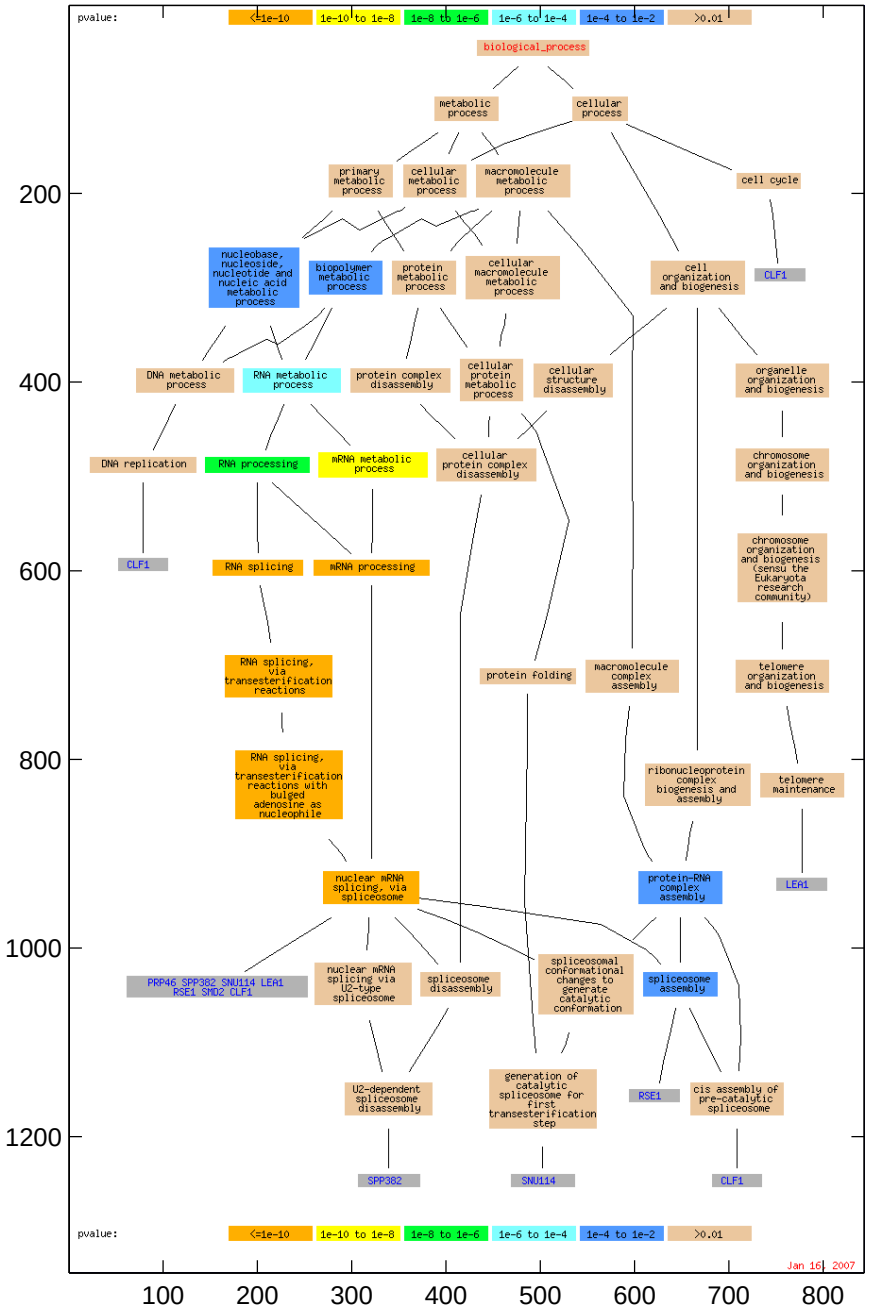

FIGURE 5. The process hierarchy tree for the hyperclique containing {Clf1 Lea1 Rse1 YLR424W Prp46 Smd2 Snu114}

5. Conclusion and Future Work

In this paper we discussed objective measures for assessing the interestingness of association patterns—itemsets and rules—for binary transaction data. Traditionally, the measures of support and confidence have been used to evaluate itemsets and association rules, but, as was described, these measures are not appropriate in all situations. More generally, it is most appropriate measure for association analysis depends on the application and the type of data. To illustrate this, we described one such measure, h-confidence. This measure and its associated pattern, hypercliques, are more appropriate when the data has a skewed distribution. A concrete example of the usefulness of hypercliques was illustrated by application to finding functional modules from protein interaction data.

There is considerable room for future work in this area, both in terms of defining new measures and in exploring which measures (and associated patterns) are most suited to various applications and types of data. Although it may seem that there are already a large number of measures, only most of these measures are defined in terms of pairs of items, and only some have been extended to sets of items. For those that have been extended, the method of extension has been specific to the given measure, and thus, a more systematic approach could be useful. There is also a need for measures that will lend themselves to the association analysis of non-binary data [**STXK04, SK05**], including data with mixed attributes. Additionally, there has not been much work in exploring the statistical aspects of many of these measures [**DP01, BHA02, Web06**]. In particular, h-confidence has shown itself to be useful, but its statistical properties and distribution in various types of data sets has not been investigated. Finally, although there are efficient algorithms for finding association patterns using support and confidence, those algorithms often cannot be applied for finding patterns defined using other measures.

References

[AIS93] Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami, *Mining association rules between sets of items in large databases*, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data (Washington, D.C.) (Peter Buneman and Sushil Jajodia, eds.), 26–28 1993, pp. 207–216.

[AS94] Rakesh Agrawal and Ramakrishnan Srikant, *Fast algorithms for mining association rules*, Proc. of the 20th VLDB Conf. (VLDB 94), 1994, pp. 487–499.

[AY01] C. C. Aggarwal and P. S. Yu, *Mining associations with the collective strength approach*, IEEE Transactions on Knowledge and Data Engineering **13** (2001), no. 6, 863–873.

[BHA02] Richard J. Bolton, David J. Hand, and Niall M. Adams, *Determining hit rate in pattern search*, Proceedings of the ESF Exploratory Workshop on Pattern Detection and Discovery (London, UK), Springer-Verlag, 2002, pp. 36–48.

[BMS97] S. Brin, R. Motwani, and C. Silverstein, *Beyond market baskets: Generalizing association rules to correlations*, Proc. ACM SIGMOD Intl. Conf. Management of Data (Tucson, AZ), 1997, pp. 265–276.

[Cle05] *SPSS: Clementine. SPSS, Inc.*, http://www.spss.com/clementine/index.htm, 2005.

[DP01] William DuMouchel and Daryl Pregibon, *Empirical bayes screening for multi-item associations*, KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (New York, NY, USA), ACM Press, 2001, pp. 67–76.

[Ent05] *SAS: Enterprise Miner. SAS Institute Inc.*, http://www.sas.com/technologies/ analytics/datamining/miner/index.html, 2005.

[FF99] C. C. Fabris and A. A. Freitas, *Discovering surprising patterns by detecting occurrences of Simpson's paradox*, Proc. of the 19th SGES Intl. Conf. on Knowledge-Based Systems and Applied Artificial Intelligence) (Cambridge, UK), December 1999, pp. 148–160.

[GH06] Liqiang Geng and Howard J. Hamilton, *Interestingness measures for data mining: A survey*, ACM Computing Surveys **38** (2006), no. 3, 9.

[GO 06] GO Consortium, *The Gene Ontology (GO) project in 2006*, Nucleic Acids Research **34** (2006), no. Database issue, D322–D326.

[GZ03] Bart Goethals and Mohammed J. Zaki, *Frequent itemset mining implementations repository*, 2003, This site contains a wide-variety of algorithms for mining frequent, closed, and maximal itemsets, http://fimi.cs.helsinki.fi/.

[HH99] R. Hilderman and H. Hamilton, *Knowledge discovery and interestingness measures: A survey*, Tech. Report TR-99-04, Department of Computer Science, University of Regina, 1999.

[HSM01] David J. Hand, Padhraic Smyth, and Heikki Mannila, *Principles of data mining*, MIT Press, Cambridge, MA, USA, 2001.

[Ins05] *S-PLUS. Insightful Miner, Insightful Corporation*, http://www.insightful.com/data-mining.asp, 2005.

[Int05] *IBM: DB2 Intelligent Miner for Data, IBM Inc.*, http://www-306.ibm.com/software/data/iminer/, 2005.

[JKK00] Mahesh Joshi, George Karypis, , and Vipin Kumar, *A universal formulation of sequential patterns*, Tech. Report 99-021, University of Minnesota, 2000.

[KK04] Michihiro Kuramochi and George Karypis, *An efficient algorithm for discovering frequent subgraphs*, IEEE Transactions on Knowledge and Data Engineering **16** (2004), no. 9, 1038–1051.

[KS96] M. Kamber and R. Shinghal, *Evaluating the interestingness of characteristic rules*, Second International Conference on Knowledge Discovery and Data Mining (KDD-96) (Portland, Oregon), 1996, pp. 263–266.

[Mos68] F. Mosteller, *Association and estimation in contingency tables*, Journal of the American Statistical Association **63** (1968), 1–28.

[Omi03] Edward R. Omiecinski, *Alternative interest measures for mining associations in databases*, IEEE TKDE **15** (2003), no. 1, 57–69.

[PS91] G. Piatetsky-Shapiro, *Discovery, analysis and presentation of strong rules*, Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, eds.), MIT Press, Cambridge, MA, 1991, pp. 229–248.

[R05] *R: A language and environment for statistical computing and graphics. The R Project for Statistical Computing*, http://www.r-project.org/, 2005.

[RJBA99] Jr. Roberto J. Bayardo and Rakesh Agrawal, *Mining the most interesting rules*, KDD '99: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (New York, NY, USA), ACM Press, 1999, pp. 145–154.

[SGD07] *Saccharomyces genome database (sgd) go gene finder*, 2007, http://db.yeastgenome.org/cgi-bin/GO/goTermFinder.

[SK05] Michael Steinbach and Vipin Kumar, *Generalizing the notion of confidence*, Proceedings of the 2005 IEEE International Conference on Data Mining (ICDM 2005), 27-30 Novemeber 2005, Houston, TX, IEEE Computer Society, 2005.

[STXK04] Michael Steinbach, Pang-Ning Tan, Hui Xiong, and Vipin Kumar, *Generalizing the notion of support*, KDD '04: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, 2004, pp. 689–694.

[TKS02] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava, *Selecting the right interestingness measure for association patterns*, Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, 2002, pp. 32–41.

[TKS04] ———, *Selecting the right objective measure for association analysis*, Information Systems **29** (2004), no. 4, 293–313.

[TSK+01] P. N. Tan, M. Steinbach, V. Kumar, S. Klooster, C. Potter, and A. Torregrosa, *Finding spatio-temporal patterns in earth science data*, KDD 2001 Workshop on Temporal Data Mining (San Francisco, CA), 2001.

[TSK05] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, *Introduction to data mining*, Pearson Addison-Wesley, May 2005.

[Web06] Geoffrey I. Webb, *Discovering significant rules*, KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (New York, NY, USA), ACM Press, 2006, pp. 434–443.

[XHD+05a] Hui Xiong, X. He, Chris Ding, Ya Zhang, Vipin Kumar, and Stephen R. Holbrook, *Identification of functional modules in protein complexes via hyperclique pattern discovery*, Proc. of the Pacific Symposium on Biocomputing, 2005, pp. 221–232.

[XHD+05b] Hui Xiong, Xiaofeng He, Chris H. Q. Ding, Ya Zhang, Vipin Kumar, and Stephen R. Holbrook, *Identification of functional modules in protein complexes via hyperclique pattern discovery.*, Biocomputing 2005, Proceedings of the Pacific Symposium (Hawaii) (Russ B. Altman, Tiffany A. Jung, Teri E. Klein, A. Keith Dunker, and Lawrence Hunter, eds.), World Scientific, Jan. 4-8 2005.

[Xio05] Hui Xiong, *Online results of using hypercliques to find functional modules in protein complexes.*, 2005, http://cimic.rutgers.edu/ hui/pfm/pfm.html.

[XPSK06] Hui Xiong, Gaurav Pandey, Michael Steinbach, and Vipin Kumar, *Enhancing data analysis with noise removal.*, IEEE Trans. Knowl. Data Eng. **18** (2006), no. 2, 304–319.

[XSK05] Hui Xiong, Michael Steinbach, and Vipin Kumar, *Privacy leakage in multi-relational databases via pattern based semi-supervised learning*, Proceedings of the ACM Conference on Information and Knowledge Management (CIKM 2005), November 2005.

[XSTK04] Hui Xiong, Michael Steinbach, Pang-Ning Tan, and Vipin Kumar, *Hicap: Hierarchical clustering with pattern preservation.*, Proceedings of the Fourth SIAM International Conference on Data Mining (Lake Buena Vista, Florida) (Michael W. Berry, Umeshwar Dayal, Chandrika Kamath, and David B. Skillicorn, eds.), SIAM, April 22-24 2004.

[XTK06] Hui Xiong, Pang-Ning Tan, and Vipin Kumar, *Hyperclique pattern discovery*, Data Min. Knowl. Discov. **13** (2006), no. 2, 219–242.

[ZO98] Mohammed Javeed Zaki and Mitsunori Ogihara, *Theoretical foundations of association rules*, 3rd SIGMOD'98 Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD), ACM Press, June 1998, pp. 7:1–7:8.

Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55343

E-mail address: steinbach@cs.umn.edu

Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824

E-mail address: ptan@cse.msu.edu

Management Science and Information Systems Department, Rutgers, Newark, NJ 07102

E-mail address: hxiong@andromeda.rutgers.edu

Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55343

E-mail address: kumar@cs.umn.edu