

Generalizing the Notion of Confidence

Michael Steinbach¹ and Vipin Kumar¹

¹Department of Computer Science and Engineering,
University of Minnesota, Minneapolis, MN, USA

Abstract. In this paper, we explore extending association analysis to non-traditional types of patterns and non-binary data by generalizing the notion of confidence. We begin by describing a general framework that measures the strength of the connection between two association patterns by the extent to which the strength of one association pattern provides information about the strength of another. Although this framework can serve as the basis for designing or analyzing measures of association, the focus in this paper is to use the framework as the basis for extending the traditional concept of confidence to Error-Tolerant Itemsets (ETIs) and continuous data. To that end, we provide two examples. First, we (1) describe an approach to defining confidence for ETIs that preserves the interpretation of confidence as an estimate of a conditional probability, and (2) show how association rules based on ETIs can have better coverage (at an equivalent confidence level) than rules based on traditional itemsets. Next, we derive a confidence measure for continuous data that agrees with the standard confidence measure when applied to binary transaction data. Further analysis of this result exposes some of the important issues involved in constructing a confidence measure for continuous data.

Keywords: confidence; support; association rules; error-tolerant itemsets; data mining

1. Introduction

Traditional association analysis (Agrawal et al. 1993) focuses on transaction data, such as the data that results when customers purchase items in a store. This market basket 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, but is 0 otherwise.¹ More recently,

Received November 30, 2005

Revised February 4, 2006

Accepted April 1, 2006

¹ This representation does not capture multiple purchases of a single item, but is a simplifying assumption commonly used in association analysis.

the concept of association analysis has also been applied to continuous data and non-traditional patterns (Steinbach et al. 2004).

A key task of association 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* (Agrawal et al. 1993), which is the number (or fraction) of transactions in which all items of the itemset appear together. Although frequent itemsets are interesting in their own right, the end goal of association analysis is typically the efficient generation of *association rules* (Agrawal et al. 1993, Agrawal & Srikant 1994), 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 .

As just described, the confidence of the association rule $A \rightarrow B$ is the ratio of the support of $A \cup B$ to the support of A . Thus, traditional confidence depends on two key elements. First, there must be a definition of support. Second, the ratio of the supports of $A \cup B$ and A must have a meaningful interpretation. However, the assumptions on which traditional confidence is based may fail to hold. We provide two examples to illustrate this point.

First, consider the non-traditional association pattern, *Error-Tolerant Itemsets* (ETIs) (Yang et al. 2001), which are itemsets in which a specified fraction of the items can be missing from a transaction. (ETIs are useful when, for example, real association patterns are distorted by noise.) To illustrate, if the specified fraction is 0.2, then for a set of 5 items, a transaction supports this itemset if it contains at least 4 out of the 5 items. As described later in this paper, the traditional definition of confidence is not appropriate for ETIs because the ratio of confidence is not meaningful in this case.

Second, consider data with continuous attributes.² Association analysis cannot be directly applied to such data. Nonetheless, if discretization techniques (Srikant & Agrawal 1996, Tan et al. 2005) are employed, then association analysis can be used for such data sets. However, discretization complicates the analysis procedure and the interpretation of its results, as well as potentially causing a loss of information. An example of an approach that directly deals with continuous data is Min-Apriori (Han et al. 1997), which uses the anti-monotone property of the \min function to produce a new version of support that also has the anti-monotone property. For binary transaction data, the support computed by Min-Apriori matches that of the standard approach if the data is pre-normalized. However, as described later, when Min-Apriori is applied to binary transaction data, the confidence that Min-Apriori computes for an association rule may not match that of traditional confidence.

To handle issues, such as those illustrated by the previous two examples, this paper presents an approach for generalizing the notion of confidence³ based on a general framework for defining the strength of the connection between two itemsets. Specifically, this framework views the strength of such a connection

² We include count attributes in this category.

³ More generally, our goal is to provide a framework for association analysis that allows it to be applied directly to both non-binary data and non-traditional types of association patterns.

(association) as the composition of two functions: (1) a function that evaluates the strength or presence of an association pattern for each transaction, producing a pattern evaluation vector, and (2) a function that measures the strength of the relationship between a pair of pattern evaluation vectors. The strength of the relationship may, for example, be measured by the extent to which one pattern evaluation vector predicts the other, or by the proximity (similarity or dissimilarity) of the two pattern evaluation vectors. Note that the traditional definition of confidence as an estimate of the conditional probability of one set of items given another is an evaluation of how well one set of items predicts another.

The following are specific contributions of this paper:⁴

1. We describe a general framework that measures the strength of the connection between two association patterns by the extent to which the strength of one association pattern provides information about the strength of another. The traditional approach to confidence is based on support, which can be regarded as a summarization of the strength of a pattern over all transactions, and represents only one way of measuring the strength of an association between sets of items. In contrast, the proposed framework is based on directly evaluating the relationship between the pattern evaluation vectors of two association patterns. The main focus of this paper is to use this framework to generalize the notion of confidence to non-traditional patterns and continuous data, but this framework has a usefulness beyond that application. For example, the framework allows any measure of association for two items to be used as a measure of association between a pair of itemsets or ETIs that contain multiple items.
2. For traditional binary transaction data, we describe how to modify the standard definition of confidence for Boolean association patterns,⁵ including ETIs, so that confidence can be viewed as an estimate of conditional probability. We provide an example of how applying the standard definition of confidence to ETI's yields a nonsensical result, while the modified definition gives the intuitively desired result. Based on these results, we provide an example involving a real-world data set that indicates of how association rules for ETIs can be more powerful than traditional association rules.
3. Using an example based on Min-Apriori, we show (1) a limitation of trying to apply the traditional definition of confidence to continuous data, and (2) demonstrate how we can use our framework to derive an alternative approach to confidence that overcomes this limitation. In doing so, we derive an interesting relationship between cosine similarity and traditional confidence. The confidence measure that we derived is further analyzed to expose some of the important issues involved in constructing a confidence measure for continuous data.

Overview Section 2 introduces the notation that will be used in this paper and quickly reviews the traditional notions of support and confidence. The limitations of confidence for non-traditional association patterns are then considered in Section 3 through an example based on ETIs. In Section 4, we describe our

⁴ A preliminary version of this paper (Steinbach & Kumar 2005) appeared in the proceedings of ICDM 2005.

⁵ A Boolean association pattern is a pattern that is either present in a transaction or not. A more precise definition will be presented shortly.

Table 1. Summary of Notation

Notation	Description
\mathcal{D}	Data matrix of M rows and N columns
$\mathcal{T} = \{t_1, \dots, t_M\}$	Set of transactions (objects, rows) of \mathcal{D}
$\mathcal{I} = \{i_1, \dots, i_N\}$	Set of items (attributes, columns) of \mathcal{D}
t	An object (transaction, row) or its index
i, j, k	An item (attribute, column) or its index
X, Y	A set of items (attributes)

general framework for defining association measures between two itemsets (two sets of attributes). A brief discussion is provided to outline the more general implications of this framework. Section 5 then uses this framework to derive a general definition of confidence for Boolean association patterns that can be interpreted in terms of conditional probability. This section also presents an example (involving ETIs and a real-life data set) that shows that our new approach to confidence can be useful for deriving association rules that are more powerful than traditional association rules. In Section 6, we discuss Min-Apriori, an approach for defining confidence in a traditional manner for continuous data, and demonstrate how this definition is not consistent with traditional confidence when Min-Apriori is applied to binary data. We also show how our framework can be used to derive a new definition for confidence that overcomes this limitation. However, further analysis reveals that yet another approach may be preferable. A discussion of related work is given in Section 7, and Section 8 provides a conclusion and directions for future work.

2. Background

After introducing some notation, we state the formal definitions of concepts, such as support and confidence, that are important to association analysis.

2.1. Notation

Table 1 provides an overview of the notation used in this and later sections. We typically use the traditional terms ‘item’ and ‘transaction’ when dealing with binary data, but commonly use the terms ‘attribute’ and ‘object’ when dealing with continuous data. In some cases, these terms are used to refer to both binary and continuous data.

2.2. Traditional Support and Confidence

Using the notation of Table 1, we briefly summarize the traditional concepts of (1) the support of an itemset, (2) a frequent itemset, (3) the support of an association rule, (4) the confidence of an association rule, and (5) the anti-monotone property of support.

Definition 1. Support

For a binary transaction data matrix \mathcal{D} with transactions \mathcal{T} and items \mathcal{I} , the

support⁶ of an itemset, i.e., a set of binary attributes, $X \subseteq \mathcal{I}$, is given by

$$\sigma(X) = |\{t \in \mathcal{T} : \mathcal{D}(t, i) = 1, \forall i \in X\}|,$$

which is simply the count of transactions containing all the items of X .

Definition 2. Frequent Itemset

Given an itemset $X \subseteq \mathcal{I}$ and a specified minimum support threshold, $minsup$, X is a frequent itemset if $\sigma(X) > minsup$.

Definition 3. Support of an Association Rule

The support of the association rule, $X \rightarrow Y$, where $X \subseteq \mathcal{I}$, $Y \subseteq \mathcal{I}$, and $X \cap Y = \emptyset$, is given by

$$\sigma(X \rightarrow Y) = \sigma(X \cup Y).$$

Definition 4. Confidence

The confidence of the association rule $X \rightarrow Y$, where $X \subseteq \mathcal{I}$, $Y \subseteq \mathcal{I}$, and $X \cap Y = \emptyset$, is given by

$$\text{conf}(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} = \frac{\sigma(X \rightarrow Y)}{\sigma(X)}.$$

Definition 5. Anti-monotone Property of Support

If X and Y , $X \subseteq Y$, are two itemsets, then $\sigma(Y) \leq \sigma(X)$.

The downward closure or anti-monotone property (Zaki & Ogihara 1998) of standard support provides an efficient way to find frequent itemsets and is the foundation of the well-known Apriori algorithm (Agrawal & Srikant 1994). The anti-monotone property for the support of itemsets allows us to find its corresponding patterns efficiently. However, as we show below, for many non-traditional patterns, such as ETIs, the lack of an anti-monotone property of support renders the notion of confidence problematic.

3. Limitations of Traditional Confidence

Here we provide an example involving ETIs to illustrate the problem with using the standard definition of confidence. Since ETIs are a type of Boolean association pattern, we begin by defining a that concept.

3.1. Boolean Association Patterns

An itemset is supported by a transaction (i.e., the pattern defined by the co-occurrence of all items in the itemset is present in a transaction) if all the items in the itemset are contained by the transaction. If we treat the items as Boolean attributes, then the process of evaluating whether a transaction supports an itemset $X = \{i_1, \dots, i_k\}$, corresponds to evaluating the truth of the Boolean formula, $i_1 \wedge i_2 \wedge \dots \wedge i_k$. This approach can be generalized—see for example

⁶ A distinction is sometimes made between ‘support count’ and ‘support,’ with the latter being defined by the ratio $\sigma(X)/|T|$. For simplicity, however, we will use ‘support’ to mean ‘support count’ in this paper unless otherwise explicitly indicated.

Table 2. Sample data set to illustrate confidence for ETIs.

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
t_1	0	1	1	1	0	0	0	0
t_2	1	0	1	1	0	0	0	0
t_3	1	1	0	1	0	0	0	0
t_4	1	1	1	0	0	0	0	0
t_5	0	0	0	0	0	1	1	1
t_6	0	0	0	0	1	0	1	1
t_7	0	0	0	0	1	1	0	1
t_8	0	0	0	0	1	1	1	1

Bollmann-Sdorra et al. 2001 or Srikant et al. 1997—in order to define more general types of association patterns by using general Boolean formulas consisting of the logical connectives \wedge (*and*), \vee (*or*), and \neg (*not*). For example, we could use the Boolean formula, $i_1 \vee i_2 \vee \dots \vee i_k$ as a pattern evaluation function. The key feature of Boolean association patterns is that, given a set of items (attributes) and a transaction, there is a formula/function/procedure that indicates whether the specified pattern is present in the transaction. As we will see, ETIs are Boolean association patterns, even though they are not defined in terms of a Boolean formula.

The computation of support for Boolean association patterns is analogous to the standard computation of support. In particular, support is the number of transactions in which the pattern occurs. Nothing has changed except the means of evaluating whether a pattern is present in a transaction. However, unlike traditional support, support for many Boolean association patterns does not have the anti-monotone property. As we illustrate with ETI’s, the traditional definition of confidence cannot be interpreted in terms of conditional probability in such cases.

3.2. Applying Traditional Confidence to ETIs

Here we attempt to use traditional confidence for a binary association pattern known as a strong ETI. Definition 6 provides a formal definition of a strong ETI (Yang et al. 2001). For brevity, we will use the term ETI instead of strong ETI.

Definition 6. Strong Error Tolerant Itemset

A strong ETI consists of a set of items $X \subseteq \mathcal{I}$, such that there exists a subset of transactions $R \subseteq \mathcal{T}$ consisting of at least $\kappa * M$ transactions, and for each $t \in R$, the fraction of items in X which are present in t is at least $1 - \epsilon$. κ is the minimum support expressed as fraction, M is the number of transactions, and ϵ is the fraction of missing items.

Our first example uses the data shown in Table 2. Each transaction must contain at least 3/8 of the specified items ($\epsilon = 5/8$) and half the transactions must support the pattern ($\kappa = 0.5$). If $X = \{i_1, i_2, i_3, i_4\}$ and $Y = \{i_5, i_6, i_7, i_8\}$, then both X and Y are ETIs with a support of 4 transactions.

Computing $\text{conf}(X \rightarrow Y)$ using the traditional definition of support yields

$$\text{conf}(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} = \frac{8}{4} = 2.$$

Table 3. Another set of sample data set to illustrate confidence for ETIs.

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}
t_1	1	1	1	0	1	1	1	1	1	1
t_2	1	1	1	1	0	1	1	1	1	1
t_3	1	1	1	1	0	1	1	1	1	1
t_4	1	1	1	0	1	1	1	1	1	1
t_5	0	0	1	1	1	1	1	1	1	1
t_6	0	1	0	1	1	1	1	1	1	1
t_7	0	0	1	1	1	1	1	1	1	1
t_8	0	0	1	1	1	1	1	1	1	1

This seems quite odd because (1) the confidence is larger than 1 and (2) the ETI pattern in X never co-occurs with the ETI pattern in Y . Thus, for ETIs, the traditional notion of confidence does not seem appropriate. Later, we will indicate how to define confidence for ETIs and other Boolean association patterns.

Our second examples show that similar problems with confidence exist for ETIs with lower values of ϵ . Consider the data shown in Table 3. Each transaction must contain at least $4/5$ of the specified items ($\epsilon = 1/5$) and half the transactions must support the pattern ($\kappa = 0.5$). If $X = \{i_1, i_2, i_3, i_4, i_5\}$ and $Y = \{i_6, i_7, i_8, i_9, i_{10}\}$, then X (Y) is an ETI with a support of 4 (8), and $X \cup Y$ has a support of 8. Thus, $\text{conf}(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} = \frac{8}{4} = 2$.

Again, the problem with confidence that is illustrated by these two examples results from the fact that ETIs do not have the anti-monotone property of support defined on page 5. Many other non-traditional Boolean association patterns have similar problems.

4. A General Framework for Association Measures Between Sets of Attributes

4.1. Pattern Evaluation Functions

To generalize the notion of confidence, we created a general framework for association measures between sets of attributes. The key to this framework is the notion of evaluating, in each transaction, the strength of a pattern involving a specified set of items. The evaluation of the strength of a pattern can take various forms. Most commonly, and this is the case for traditional association analysis, the pattern is either present, i.e., the pattern strength is 1, or it is absent, i.e., the pattern strength is 0. (As noted, we call such patterns *Boolean association patterns*.) More generally, the evaluation of strength can be a real number. This will be particularly relevant when we discuss confidence for continuous data.

Formally, a pattern is defined by a pattern evaluation function. In particular, an evaluation function, *eval*, is a function that takes a set of items $X \subseteq \mathcal{I}$ as an argument, and returns a *pattern evaluation vector*, \mathbf{v} , whose i^{th} component is the strength of the pattern in the i^{th} transaction. Thus, we can write

$$\mathbf{v}(t) = \text{eval}(t, X), \forall t \in \mathcal{T} \quad \text{or} \quad \mathbf{v} = \text{eval}(X) \quad (1)$$

If there are several sets of items under consideration, e.g., X and Y , then we will distinguish between their pattern evaluation vectors by using subscripts,

Table 4. Pattern evaluation functions.

	<i>eval</i> function	Definition
1	\wedge (and)	$eval_{\wedge}(t, X) = \mathcal{D}(t, i_1) \wedge \dots \wedge \mathcal{D}(t, i_k)$
2	\prod (product)	$eval_{\prod}(t, X) = \mathcal{D}(t, i_1) * \dots * \mathcal{D}(t, i_k)$
3	min	$eval_{\min}(t, X) = \min_{1 \leq j \leq k} \{\mathcal{D}(t, i_j)\}$
4	max	$eval_{\max}(t, X) = \max_{1 \leq j \leq k} \{\mathcal{D}(t, i_j)\}$
5	range	$eval_{\text{range}}(t, X) = eval_{\max}(t, X) - eval_{\min}(t, X)$
6	strong ETI	$eval_{eti, \epsilon}(t, X) = \frac{\sum_{i \in X} \mathcal{D}(t, i)}{ X } \geq 1 - \epsilon$

Table 5. Computation of eval vectors using $eval_{\wedge}$ and $eval_{\min}$.

(a) $eval_{\wedge}$.				(b) $eval_{\min}$.			
	item 1	item 2	v		attr 1	attr 2	v
1	1	0	0	1	0.37	0.25	0.25
2	0	1	0	2	0.22	1.00	0.22
3	1	1	1	3	0.81	0.11	0.11
4	1	1	1	4	0.19	0.05	0.05
5	1	0	0	5	0.33	0.55	0.33

e.g., \mathbf{v}_X and \mathbf{v}_Y . Various *eval* functions are shown in Table 4. Note that $X = \{i_1, i_2, \dots, i_k\} \subseteq \mathcal{I}$ and that some of the *eval* functions are valid for both binary and continuous data. An illustration of the operation of evaluation functions for binary and continuous data is provided by Tables 5(a) and 5(b), respectively.

Conceptually an itemset (set of attributes in the continuous case) is replaced by a vector, whose length is the number of transactions. Each component of this pattern vector measures the strength of the pattern in a particular transaction. This mapping of each itemset to a pattern vector creates a collection of pattern vectors and further analysis involves only these vectors. This is analogous to computing a similarity matrix from a data matrix and then performing the clustering using only this similarity matrix. For clustering, pairs of objects (transactions) are transformed into values in a similarity space, while for our framework, itemsets (sets of attributes) are transformed into vectors in a pattern space. This approach can also be used to create a framework for generalizing support by equating support with measures of the magnitude of a pattern vector (Steinbach et al. 2004).

4.2. Association Measures Between Sets of Attributes

Most generally, the strength of a connection (association) between two association patterns can be viewed as a measure of the information that one association pattern provides about another, where by association pattern, we mean any pattern involving an itemset (set of attributes) that is defined by a pattern evaluation vector. Specifically, the strength of an association is a function that quantifies the relationship between the evaluation vectors of a pair of itemsets (pair of sets of attributes). This is captured by the following definition.

Definition 7. A General Definition of the Strength of an Association

If $X \subseteq \mathcal{I}$ and $Y \subseteq \mathcal{I}$, $X \cap Y = \emptyset$, are itemsets (sets of attributes), then the strength of the association between them can be defined in terms of a function π that maps the evaluation vectors of the two association patterns to a real value. More formally, $\text{assoc} : (\mathfrak{R}^M, \mathfrak{R}^M) \rightarrow \mathfrak{R}$, where \mathfrak{R} is the set of real numbers and M is the number of transactions (objects).

$$\text{assoc}(X, Y) = \pi(\text{eval}(X), \text{eval}(Y)) = \pi(\mathbf{v}_X, \mathbf{v}_Y) \quad (2)$$

We are not interested in just any function, π , but rather those functions that meaningfully capture the strength of a relationship between two association patterns. For example, π may (1) capture the extent to which the strength of one association pattern can be used to predict another, or (2) capture the proximity (similarity or dissimilarity) between the two association patterns. Although our focus in this paper will be on using this framework to extend the notion of confidence, we briefly describe some of the possible alternatives to confidence and then discuss some of the more general implications of this framework. After, this brief discussion, the remainder of this paper will focus on using the framework to generalize the notion of confidence.

4.2.1. Examples of Association Measures

We begin by describing association measures, which are called interestingness or objective measures, that have been developed to address the issue that association measures with properties different from that of traditional confidence can be useful. A list of many such measures is given in Table 6. Note that these measures are given in terms of two binary variables i_1 and i_2 and the frequencies with which their 0 and 1 values co-occur, as defined in Table 7. These measures have been extensively investigated by numerous researchers. A survey of these measures, their properties, and related issues is provided in Tan et al. 2002, Tan et al. 2004, and Tan et al. 2005.

More specifically, the measures in Table 6 can be characterized in a number of different ways. For example, there are several, including confidence, that have an interpretation in terms of conditional probability or related notions involving statistical independence or information theory. *All-confidence* is the minimum of the confidence measured for all possible rules that can be created from the union of the two itemsets. *Lift* is the confidence divided by the support of the second itemset, while *interest* is, conceptually, the probability that the two items co-occur divided by the probability that would be observed if they are independent. *Mutual information* is an information theoretic measure that quantifies the degree of information that the first itemset (the antecedent) provides about second itemset (the consequent). As an example of measures of a different sort, some of the other measures in Table 6 correspond to similarity measures, namely, the *cosine measure*, *correlation*, and the *Jaccard coefficient*.

Indeed, when the pattern evaluation vectors \mathbf{v}_X and \mathbf{v}_Y are continuous, it seems especially useful to consider confidence measures based on some of the similarity or distance measures that have been developed for evaluating the strength of a connection between two continuous vectors. Some of these measures are the continuous analogues of similarity measures in Table 6, such as the cosine, correlation, and the extended Jaccard (Strehl et al. 2000) measures. Other alternatives include *Euclidean distance* and a relatively new set of measures, known as *Bregman divergences* (Banerjee et al. 2004).

Table 6. Examples of objective (interestingness) measures for the itemset $\{i_1, i_2\}$.

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

4.2.2. General Usefulness of the Framework

We mention three general uses for this framework. First, the framework allows any binary measure of association, i.e., any measure is defined only for two items (attributes), to be used when the itemsets, X and Y , contain more than two items. Specifically, we first compute the pattern evaluation vectors of X and Y and then calculate the strength of the association in terms of this pair of pattern evaluation vectors. Although some binary measures, including those of Table 6, have natural extensions to the case where X and Y contain multiple items, many do not. Second, using the framework, the measures of Table 6 can be automatically applied for non-traditional patterns, such as ETIs. As we show below, a generalized version of confidence for ETIs seems quite useful, and thus, generalized versions of other association measures for ETIs may also prove valuable. Third, our framework provides a way of defining various measures of association when the pattern evaluation functions that are applied to the data (continuous, binary, etc.) produce continuous pattern evaluation vectors. In particular, any of the many similarity and dissimilarity measures that have been defined for pairs of continuous vectors can be used as association measures between two sets of continuous attributes.

Table 7. Contingency table for two binary variables, i_1 and i_2 .

	$i_1 = 1$	$i_1 = 0$
$i_2 = 1$	f_{11}	f_{10}
$i_2 = 0$	f_{01}	f_{00}

5. Confidence for Boolean Evaluation Functions

In this section, we use the framework described above to generalize the interpretation of confidence as conditional probability to the case of general Boolean association patterns. We then show how this generalized version of confidence provides a reasonable value of confidence for the first example of Section 3. Finally, we provide an example with a real-life data set that demonstrates the practical value of association rules based on ETIs.

5.1. Derivation of Confidence from Conditional Probability

Given two attributes X and Y , and a Boolean pattern evaluation function, what we are seeking is a measure of the strength of the connection between the pattern evaluation vectors of X and Y that corresponds to traditional confidence in the sense that it can be interpreted as an estimate of conditional probability. For any Boolean pattern evaluation function, $eval_b$, the resulting pattern evaluation vectors, $\mathbf{v}_X = eval_b(X)$ and $\mathbf{v}_Y = eval_b(Y)$, are binary vectors. Furthermore, we assume that only the presence of a pattern (a value of 1) is important. This reduces the problem to one of computing, $Prob(\mathbf{v}_Y(t) = 1 | \mathbf{v}_X(t) = 1)$, which is the probability that an entry of \mathbf{v}_Y is 1 when the corresponding entry of \mathbf{v}_X is 1. An estimate of this conditional probability is the fraction of entries where \mathbf{v}_X and \mathbf{v}_Y are both 1; i.e., $Prob(\mathbf{v}_Y(t) = 1 \wedge \mathbf{v}_X(t) = 1)$, divided by the fraction of entries where $\mathbf{v}_X(t)$ is 1, i.e., $Prob(\mathbf{v}_X(t) = 1)$. This discussion leads to the following general definition of the confidence of an association rule when using a Boolean association pattern:

Definition 8. Confidence of an Association Rule for Boolean Association Patterns

Given two disjoint sets of attributes X and Y , the traditional support function, σ , and a Boolean pattern evaluation function $eval_b$ that defines the Boolean association pattern and generates pattern evaluation vectors, $eval_b(X) = \mathbf{v}_X$ and $eval_b(Y) = \mathbf{v}_Y$, the generalized confidence, $gconf(X \rightarrow Y)$, of the association rule $X \rightarrow Y$ with respect to the Boolean association pattern is given by

$$\begin{aligned}
 gconf(X \rightarrow Y) &= Prob(Y|X) \\
 &= \frac{Prob(X \wedge Y)}{Prob(X)} \\
 &= \frac{|\{t \in \mathcal{T} : \mathbf{v}_X(t) = 1 \wedge \mathbf{v}_Y(t) = 1\}|}{|\{t \in \mathcal{T} : \mathbf{v}_X(t) = 1\}|} \\
 &= \frac{\sigma(\{\mathbf{v}_X, \mathbf{v}_Y\})}{\sigma(\{\mathbf{v}_X\})} \\
 &= conf(\{\mathbf{v}_X\} \rightarrow \{\mathbf{v}_Y\})
 \end{aligned}$$

where we have used the fact that $|\{t \in \mathcal{T} : \mathbf{v}_X(t) = 1\}| = \sigma(\{\mathbf{v}_X\})$ and $|\{t \in \mathcal{T} : \mathbf{v}_X(t) = 1 \wedge \mathbf{v}_Y(t) = 1\}| = \sigma(\{\mathbf{v}_X, \mathbf{v}_Y\})$.

Thus, we obtain a generalized version of confidence which can be interpreted as conditional probability for the case of a general Boolean association pattern. However, our results also show that the standard approach for computing confidence cannot be used in the general case. In particular, the approach employed in standard association analysis would use a numerator of $\sigma(X \cup Y) = |\{t \in \mathcal{T} : \mathbf{v}_{X \cup Y}(t) = 1\}|$ instead of $\sigma(\{\mathbf{v}_X, \mathbf{v}_Y\}) = |\{t \in \mathcal{T} : \mathbf{v}_X(t) = 1 \wedge \mathbf{v}_Y(t) = 1\}|$, as given in Definition 8. The equivalence of $\sigma(X \cup Y)$ and $\sigma(\{\mathbf{v}_X, \mathbf{v}_Y\})$ is a special case and does not hold in general.

5.2. Example: Strong Error-Tolerant Itemsets (Continued)

Computing $\text{conf}(X \rightarrow Y)$, as given in Definition 8, requires computing $\sigma(\{\mathbf{v}_X, \mathbf{v}_Y\})$. In turn, this requires us to ‘and’ the pattern evaluation vectors of X and Y and then sum the entries of the resultant vector. The ETI pattern of X occurs only in the first four transactions, while that of Y occurs only in the last four transactions. If \wedge denotes the componentwise ‘and’ function, then

$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \wedge \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

and consequently, $\sigma(\{\mathbf{v}_X, \mathbf{v}_Y\}) = 0$.

Therefore,

$$\text{gconf}(X \rightarrow Y) = \frac{\sigma(\{\mathbf{v}_X, \mathbf{v}_Y\})}{\sigma(\{\mathbf{v}_X\})} = 0/4 = 0,$$

which is much more intuitive than the previous result: $\text{conf}(X \rightarrow Y) = 2$.

5.3. An Example of the Application of Generalized Confidence to the Mushroom Data Set

Here we present an example of the power of confidence for ETIs. We use the mushroom dataset from the FIMI website (Goethals & Zaki 2003), which has binary attributes.⁷ In the binary format, the data set consists of 119 binary attributes (items) and 8124 rows (transactions). Each transaction represents a mushroom and each item an attribute, e.g., odor or color. The first item, denoted by p , indicates whether a mushroom is poisonous, while the second item, e , indicates a mushroom is edible. There were 3916 poisonous mushrooms and 4208 edible ones.

For a complete analysis, it would be desirable to seek rules involving any number of attributes, but here the focus is on illustrating the practical usefulness

⁷ The mushroom data set with categorical attributes is available from the UCI machine learning repository (Newman et al. 1998).

of confidence involving ETIs versus regular itemsets. Thus, we restrict ourselves to rules involving only three items on the left hand side and the item p on the right hand side. Specifically, we found all itemsets of three items (excluding p and e) and computed the confidence between these itemsets and the p item, both by treating the itemset as a standard itemset and as an ETI.

First, we computed confidence in the traditional manner, as described by Definition 4. Second, we treated each itemset as an ETI that had to have at least two of the three items and calculated confidence of the rule as given in Definition 8. We kept only those rules that had a confidence of 1. We give two examples.

One rule was $\{29, 48, 90\} \rightarrow p$. With traditional association analysis, this rule has a confidence of 1 and a support of 576. On the other hand, if we require only two of the three items, i.e., treat $\{29, 48, 90\}$ as an ETI, then we still have a confidence of 1, but the support is 3312. Thus, in this case, using ETIs, with the modified approach to confidence, results in a rule of equal quality, but expanded coverage.

Another rule was $\{48, 85, 95\} \rightarrow p$. With traditional association analysis, this rule has a confidence and support of 0. However, if we require only two of the three items, i.e., treat $\{48, 85, 95\}$ as an ETI, then the confidence is 1 and the support is 3024. The reason for this is that items 48 and 95 never co-occur with one another, but they do frequently co-occur individually with item 85. This type of disjunctive rule (Elble et al. 2003, Nanavati et al. 2001, Zelenko 1999) would not be produced by standard association analysis.

In summary, generalizing the notion of confidence allows us to meaningfully use association rules based on ETIs. In turn, as this example has illustrated, such rules can provide better coverage (at an equivalent level of confidence) than rules based on traditional itemsets. Furthermore, rules based on ETIs automatically incorporate disjunctive characteristics.

6. Confidence for Continuous Data

In this section, we consider, for continuous data, the traditional notion of confidence as defined as a ratio of supports. The basis for this discussion will be the Min-Apriori algorithm Han et al. (1997), which we will describe next. On the surface, this algorithm defines support and confidence for continuous data in a way that seems almost identical to the way these concepts are defined for binary data. However, when applied to binary data, Min-Apriori's definition of confidence produces results that differ from those of traditional confidence applied to the same data. Using the framework introduced in Section 4, we analyze this issue and produce a new measure of confidence for Min-Apriori that agrees with traditional confidence for binary data. This result is further analyzed to highlight some of the important issues in constructing a confidence measure for continuous data.

6.1. Min Apriori

The Min-Apriori algorithm operates as follows. First, to adjust for possible differences in the scales of attributes, Min-Apriori normalizes the data in each column (attribute) of the data matrix by dividing each column entry by the

Table 8. Computation of support for continuous data.

	i_1	i_2	$\min(\{i_1, i_2\})$
t_1	0.45	0.55	0.45
t_1	0.25	0.05	0.05
t_1	0.30	0.25	0.25
t_1	0.00	0.15	0.00
<i>Support</i>	1.00	1.00	0.75

Table 9. Data for confidence calculations of Min-Apriori.

(a) Sample data.				(b) Normalized data.				(c) Eval vectors.		
A	B	C	D	A	B	C	D	AB	CD	ABCD
1	1	1	1	1/6	1/3	1/5	1/4	1/6	1/5	1/6
1	1	1	1	1/6	1/3	1/5	1/4	1/6	1/5	1/6
1	0	1	1	1/6	0	1/5	1/4	0	1/5	0
1	0	1	1	1/6	0	1/5	1/4	0	1/5	0
1	1	0	0	1/6	1/3	0	0	1/6	0	0
1	0	1	0	1/6	0	1/5	0	0	0	0

sum of the column entries. Given this normalization, the support of a set of attributes is computed by taking the minimum value in each row and summing the resultant values. In what follows, we will indicate this support function with the notation, σ_{\min} to distinguish this support function from the standard one.⁸ The Min-Apriori process of computing the support of already normalized data is illustrated in Table 8.

Min-Apriori uses the traditional, support-based definition of confidence, although with the definition of support as defined above, i.e., $\text{conf}(X \rightarrow Y) = \sigma_{\min}(X \cup Y) / \sigma_{\min}(Y)$. However, if Min-Apriori is given binary transaction data, it does not always yield the same confidence values as the standard definition of confidence on the original data. This is illustrated by the data shown in Table 9. The original data is shown in Table 9(a), Table 9(b) shows the data after it has been normalized, and Table 9(c) shows the evaluation vectors. Using traditional confidence, we obtain

$$\text{conf}(CD \rightarrow AB) = \sigma(ABCD) / \sigma(CD) = 2/4 = 0.5.$$

However, using Min-Apriori's definition of support, we get

$$\text{conf}(CD \rightarrow AB) = \frac{\sigma_{\min}(ABCD)}{\sigma_{\min}(CD)} = (2/6) / (4/5) = 5/12 = 0.42.$$

6.2. A New Confidence Measure for Continuous Data

To illustrate the utility of our framework when applied to continuous data, we derive a confidence measure for continuous data that agrees with traditional

⁸ In Steinbach et al. 2004, we use slightly different notation and provide a more formal description of support in terms of summarizing a pattern evaluation vector. To enable this discussion to stand on its own, we simplify both the notation and discussion.

confidence when used on binary transaction data. We first prove a supporting theorem that relates the cosine measure and traditional confidence and then use this result to derive a new confidence measure that is consistent with confidence for binary data. An example illustrates that this new measure achieves the desired consistency. However, a further analysis of the new measure of confidence more clearly reveals a key issue involved in extending confidence to continuous data, namely, the need to normalize continuous data to remove dependencies of scale and the fact that the confidence measure is not invariant to such a normalization.

6.2.1. Relationship Between the Cosine Measure and Traditional Confidence

We begin by proving Theorem 6.1, which indicates that the traditional confidence of the association rule $X \rightarrow Y$ for binary data equals the cosine measure between the evaluation vectors of X and Y multiplied by a factor that depends on the relative support of the two itemsets. This theorem serves as the basis for our approach, as well as being interesting in its own right. Note that if both association patterns have the same support, then the confidence simply reduces to the cosine of their pattern evaluation vectors, i.e., $\cos(eval_{\wedge}(X), eval_{\wedge}(Y))$. In this case, confidence is also symmetric, i.e., $\text{conf}(X \rightarrow Y) = \text{conf}(Y \rightarrow X)$.

Theorem 6.1. Given traditional support and the corresponding evaluation function, $eval_{\wedge}$, we get

$$\text{conf}(X \rightarrow Y) = \cos(eval_{\wedge}(X), eval_{\wedge}(Y)) \sqrt{\frac{\sigma(Y)}{\sigma(X)}} \quad (3)$$

Proof.

$$\begin{aligned} \text{conf}(X \rightarrow Y) &= \frac{\sigma(X \cup Y)}{\sigma(X)} \\ &= \frac{eval_{\wedge}(X) \cdot eval_{\wedge}(Y)}{\|eval_{\wedge}(X)\|_2^2} \\ &= \cos(eval_{\wedge}(X), eval_{\wedge}(Y)) \frac{\|eval_{\wedge}(Y)\|_2}{\|eval_{\wedge}(X)\|_2} \\ &= \cos(eval_{\wedge}(X), eval_{\wedge}(Y)) \sqrt{\frac{\sigma(Y)}{\sigma(X)}} \end{aligned}$$

where $eval_{\wedge}(X) \cdot eval_{\wedge}(Y)$ is the dot product of evaluation vectors of X and Y , respectively. We have used the definition of the cosine measure, $\cos(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y} / \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$, as well as the following pair of facts: $\sigma(X \cup Y) = eval_{\wedge}(X) \cdot eval_{\wedge}(Y)$ and $\sigma(X) = \|eval_{\wedge}(X)\|_1 = \|eval_{\wedge}(X)\|_2^2$. Note that $\|\cdot\|_1$ and $\|\cdot\|_2$ are the L_1 and L_2 vector norms, respectively. \square

6.2.2. A Measure Consistent With Traditional Confidence

If two vectors are multiplied by (potentially different) non-zero constants, their cosine measure is unchanged. As a result, if each attribute of a binary transaction matrix is normalized to have an L_1 norm of 1, then the cosine measure between attributes does not change. Thus, if we can express $\sqrt{\sigma(Y)/\sigma(X)}$ in

terms of support based on the Min-Apriori support function, σ_{\min} , then we will have an alternative definition of confidence that matches the standard definition of confidence for binary transaction data and works for continuous data. The following theorem provides the details.

Theorem 6.2. For traditional binary transaction data whose attributes have been normalized to have an L_1 norm of 1 (the Min-Apriori normalization) and an itemset X , $\sigma_{\min}(X)/m = \sigma(X)$, where m is the mean of the non-zero entries of $eval_{\min}(X)$ and $\sigma(X)$ is the traditional support of the attributes X for the original, unnormalized data.

Proof. First, notice that when attribute i is normalized, its non-zero entries become $1/\sigma(i)$. Thus, all the non-zero entries of $eval_{\min}(X)$ have the value $\min_{i \in X}(1/\sigma(i))$. Trivially, the average of the non-zero values of $eval_{\min}(X)$ will be $m = \min_{i \in X} 1/\sigma(i)$. Furthermore, $eval_{\min}(X)$ will have $\sigma(X)$ such entries, one for each transaction where all the attributes of X are all non-zero. Thus, $\sigma_{\min}(X) = \min_{i \in X} 1/\sigma(i) \sigma(X) = m \sigma(X)$, and the theorem follows. \square

Using the previous two results, we get a new definition of confidence for continuous data that is normalized to have an L_1 norm of 1.

Definition 9. New Definition of Confidence for Continuous Data

$$\text{conf}(X \rightarrow Y) = \cos(eval_{\min}(X), eval_{\min}(Y)) \sqrt{\frac{\sigma_{\min}(Y)/m_Y}{\sigma_{\min}(X)/m_X}},$$

where m_X (m_Y) is the mean of the non-zero values of $eval_{\min}(X)$ ($eval_{\min}(Y)$).

6.2.3. Example: Confidence for Continuous Data Using Min-Apriori (Continued)

We provide a numerical illustration of the result that we have just derived, using the data in Table 9. Recall that $\text{conf}(CD \rightarrow AB) = 0.5$ when traditional confidence is used. Using Table 9(c), we apply Definition 9 with $X = \{C, D\}$, and $Y = \{A, B\}$. Summing column AB of Table 9(c), we obtain $\sigma_{\min}(AB) = 3/6$. Summing column CD of Table 9(c), we get $\sigma_{\min}(CD) = 4/5$. From Table 9(c), we see $m_{AB} = 1/5$ and $m_{CD} = 1/6$. Finally, $\cos(eval_{\min}(CD), eval_{\min}(AB)) = 1/\sqrt{3}$. Thus, the confidence given by Definition 9 is

$$\text{conf}(CD \rightarrow AB) = \frac{1}{\sqrt{3}} \sqrt{\frac{(3/6)/(1/6)}{(4/5)/(1/5)}} = \frac{1}{\sqrt{3}} \sqrt{\frac{3}{4}} = \sqrt{\frac{1}{4}} = 0.5.$$

This new value agrees with that produced by traditional confidence as illustrated in the example in Section 6.1.

6.2.4. Further Analysis

Although we have produced a measure of confidence that addresses a potential limitation of the Min-Apriori version of confidence, further analysis is warranted to better understand the meaning of this measure. In particular, at first glance, division by the mean of the non-zero entries seems mysterious. However, the role of this factor is clear if we realize that its effect is to counteract the

normalization of the data performed by Min-Apriori as an initial preprocessing step. An examination of the previous example provides an illustration of this.

However, a more formal understanding of this situation is possible. In Tan et al. 2004, various properties of association measures were considered, including invariance with respect to the scaling of the data. The normalization performed by Min-Apriori corresponds to an independent scaling of the columns of the data, and, as was indicated in Tan et al. 2004, confidence is not invariant to such a transformation. In light of this, the formula for confidence that we derived above becomes much more understandable; i.e., to match confidence for unnormalized binary data, the effects of normalization must be counteracted. It is worth noting that if the normalization step of Min-Apriori is omitted, then Min-Apriori produces confidence results for binary data that are consistent with traditional confidence.⁹

In summary, there are two approaches that extend the traditional confidence measure to continuous data in a way that yields the same value as traditional confidence for the original binary data. However, both approaches run counter to the purpose of normalization, which may be desirable in situations involving continuous data. Indeed, despite the fact that Min-Apriori is not always consistent with traditional confidence on binary data, the Min-Apriori approach to confidence may be preferable for continuous data because of the normalization issue. However, other types of association measures should also be considered. More investigation is needed into these issues.

6.3. A Final Observation

Here we make the observation, which some may find interesting, that the Min-Apriori approach to confidence is only partially inconsistent with confidence for binary data. Consider the following example that again uses Tables 9a-c, but computes $\text{conf}(AB \rightarrow CD)$ instead of $\text{conf}(CD \rightarrow AB)$. Using traditional confidence, we obtain

$$\text{conf}(AB \rightarrow CD) = \sigma(ABCD)/\sigma(AB) = 2/3 = 0.67.$$

Using Min-Apriori's definition of support, we get

$$\text{conf}(AB \rightarrow CD) = \frac{\sigma_{\min}(ABCD)}{\sigma_{\min}(AB)} = (2/6)/(3/6) = 2/3 = 0.67.$$

In this case, the traditional confidence of unnormalized binary data matches that obtained from Min-Apriori. Indeed, when using Min-Apriori, the confidence of one of the association rules $\text{conf}(X \rightarrow Y)$ or $\text{conf}(Y \rightarrow X)$ will always match that of traditional confidence when applied to binary data. We omit a formal proof, but this is readily established using the relationship, $\sigma_{\min}(X) = \min_{i \in X} 1/\sigma(i) \sigma(X)$, that was used in the previous proof.

⁹ This depends on the fact that, for binary data, the min evaluation function is equivalent to an evaluation function based on *and*. In this case, the support corresponds to a count not a fraction, but this changes nothing important.

7. Related Work

Although there has been considerable work on various association measures (Tan et al. 2002, 2004), we are unaware of any work similar to ours in constructing a framework for such measures or in generalizing confidence. The most relevant previous work concerns interestingness/objective measures. These alternative association measures and their connection to the framework were discussed in Section 4 and are surveyed in Tan et al. 2002, Tan et al. 2004, and Tan et al. 2005.

Another area that has some connection to our work is quantitative association rules (Aumann & Lindell 1999, Okoniewski et al. 2003, Srikant & Agrawal 1996, Webb 2001), which try to predict the values of one set of attributes from another. There has also been some work in integrating regression with association analysis (Ozgur et al. 2004). However, both these approaches are distinct from our view of association rules and confidence, which involves only the prediction of the strength/presence of one pattern from another.

Finally, we mention our previous work on generalizing support (Steinbach et al. 2004), which provides a framework for generalizing support. The work in this paper, although independent of the support work, is complementary to that work. Our previous work on Min-Apriori (Han et al. 1997) also played an important role in this work by providing a concrete example of confidence for continuous data.

8. Conclusions and Future Work

We described a general framework for measures of the strength of an association between two itemsets (sets of attributes) and showed how this framework could be used to generalize confidence. First, we showed how to define confidence for Boolean association patterns, including ETIs, in a way that preserves the interpretation of confidence as an estimate of conditional probability. Second, we constructed an estimate of confidence for continuous data that agrees with traditional confidence when applied to binary data. However, a further analysis of this measure showed that its approach runs counter to the normalization that may be desirable for continuous data.

With respect to future work, there is a need for efficient algorithms for finding high confidence association rules for non-traditional patterns and non-binary data sets. Also, more experiments are needed to evaluate the usefulness of such association rules. Although preliminary, the results of using ETIs to form association rules for the mushroom data set were very promising and need to be further investigated. In particular, we would like to conduct an extensive comparison of the usefulness of rules based on ETIs and traditional association rules. Also, the connection of rules based on ETIs to disjunctive rules and more general Boolean patterns needs to be further explored. In the context of Boolean association patterns, it would be worthwhile to explore using various objective (interestingness) measures (Tan et al. 2002) as association measures, particularly in the context of ETIs or other Boolean association patterns. Finally, more research is needed to understand the appropriate association measures for continuous association analysis.

Acknowledgements. We thank the anonymous reviewers for their very useful comments and suggestions. This work was partially supported by NASA grant #NCC 2

1231, NSF grant #ACI-0325949 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.

References

- Agrawal, R., Imielinski, T. & Swami, A. N. (1993), Mining Association Rules between Sets of Items in Large Databases, *in* P. Buneman & S. Jajodia, eds, 'Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data', Washington, D.C., pp. 207–216.
- Agrawal, R. & Srikant, R. (1994), Fast Algorithms for Mining Association Rules, *in* 'Proc. of the 20th VLDB Conf. (VLDB 94)', pp. 487–499.
- Aumann, Y. & Lindell, Y. (1999), A Statistical Theory for Quantitative Association Rules, *in* 'Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 99)', ACM Press, pp. 261–270.
- Banerjee, A., Merugu, S., Dhillon, I. S. & Ghosh, J. (2004), Clustering with Bregman Divergences, *in* 'Proc. of the 2004 SIAM Intl. Conf. on Data Mining', Lake Buena Vista, FL, pp. 234–245.
- Bollmann-Sdorra, P., Hafez, A. & Raghavan, V. V. (2001), A Theoretical Framework for Association Mining Based on the Boolean Retrieval Model, *in* Y. Kambayashi, W. Winiwarter & M. Arikawa, eds, 'Data Warehousing and Knowledge Discovery, Third International Conference, DaWaK 2001, Munich, Germany, September 5-7, 2001, Proceedings', Vol. 2114 of *Lecture Notes in Computer Science*, Springer, pp. 21–30.
- Elble, J., Heeren, C. & Pitt, L. (2003), Optimized Disjunctive Association Rules via Sampling, *in* 'ICDM '03: Proceedings of the Third IEEE International Conference on Data Mining', IEEE Computer Society, Washington, DC, p. 43.
- Goethals, B. & Zaki, M. J. (2003), 'Frequent Itemset Mining Implementations Repository'. This site contains a wide-variety of algorithms for mining frequent, closed, and maximal itemsets, <http://fimi.cs.helsinki.fi/>.
- Han, E.-H., Karypis, G. & Kumar, V. (1997), TR# 97-068: Min-Apriori: An Algorithm for Finding Association Rules in Data with Continuous Attributes, Technical report, Department of Computer Science, University of Minnesota, Minneapolis, MN.
- Nanavati, A. A., Chitrapura, K. P., Joshi, S. & Krishnapuram, R. (2001), Mining Generalised Disjunctive Association Rules, *in* 'CIKM '01: Proceedings of the Tenth International Conference on Information and Knowledge Management', ACM Press, New York, NY, pp. 482–489.
- Newman, D., Hettich, S., Blake, C. & Merz, C. (1998), 'UCI Repository of machine learning databases'.
URL: <http://www.ics.uci.edu/~mllearn/MLRepository.html>
- Okoniewski, M., Gancarz, L. & Gawrysiak, P. (2003), Mining Multi-Dimensional Quantitative Associations, *in* O. Bartenstein, U. Geske, M. Hannebauer & O. Yoshie, eds, '14th International Conference on Applications of Prolog, INAP

- 2001, Tokyo, Japan, October 20-22, (LNCS Volume', Vol. 2543 of *Lecture Notes in Computer Science*, Springer.
- Ozgur, A., Tan, P.-N. & Kumar, V. (2004), RBA: An Integrated Framework for Regression Based on Association Rules., in 'Proc. of the 2004 SIAM Intl. Conf. on Data Mining'.
- Srikant, R. & Agrawal, R. (1996), Mining Quantitative Association Rules in Large Relational Tables, in H. V. Jagadish & I. S. Mumick, eds, 'Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data', Montreal, Quebec, Canada, pp. 1–12.
- Srikant, R., Vu, Q. & Agrawal, R. (1997), Mining Association Rules with Item Constraints, in D. Heckerman, H. Mannila, D. Pregibon & R. Uthurusamy, eds, 'Proc. 3rd Int. Conf. Knowledge Discovery and Data Mining, KDD', AAAI Press, pp. 67–73.
- Steinbach, M. & Kumar, V. (2005), Generalizing the Notion of Confidence., in 'Proceedings of the 5th IEEE International Conference on Data Mining (ICDM 2005)', IEEE Computer Society, pp. 402–409.
- Steinbach, M., Tan, P.-N., Xiong, H. & Kumar, V. (2004), Generalizing the Notion of Support, in 'KDD '04: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM Press, pp. 689–694.
- Strehl, A., Ghosh, J. & Mooney, R. (2000), Impact of Similarity Measures on Web-page Clustering, in 'Workshop of Artificial Intelligence for Web Search (AAAI 2000)', AAAI Press, pp. 58–64.
- Tan, P.-N., Kumar, V. & Srivastava, J. (2002), Selecting the Right Interestingness Measure for Association Patterns, in 'Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM Press, pp. 32–41.
- Tan, P.-N., Kumar, V. & Srivastava, J. (2004), 'Selecting the Right Objective Measure for Association Analysis', *Information Systems* **29**(4), 293–313.
- Tan, P.-N., Steinbach, M. & Kumar, V. (2005), *Introduction to Data Mining*, Pearson Addison-Wesley.
- Webb, G. I. (2001), Discovering Associations With Numeric Variables, in 'Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2001)', ACM Press, pp. 383–388.
- Yang, C., Fayyad, U. M. & Bradley, P. S. (2001), Efficient Discovery of Error-Tolerant Frequent Itemsets in High Dimensions, in 'Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2001)', ACM Press, pp. 194–203.
- Zaki, M. J. & Ogihara, M. (1998), Theoretical Foundations of Association Rules, in '3rd SIGMOD'98 Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD)', ACM Press, pp. 7:1–7:8.
- Zelenko, D. (1999), Optimizing Disjunctive Association Rules, in 'PKDD '99: Proceedings of the Third European Conference on Principles of Data Mining and Knowledge Discovery', Springer-Verlag, London, UK, pp. 204–213.

Author Biographies



Michael Steinbach earned a B.S. degree in Mathematics, an M.S. degree in Statistics, and the M.S. and Ph.D. degrees in Computer Science from the University of Minnesota. He also has held a variety of software engineering, analysis, and design positions in industry at Silicon Biology, Racotek, and NCR. Steinbach is currently a research associate in the Department of Computer Science and Engineering at the University of Minnesota, Twin Cities. He is a co-author of the textbook, *Introduction to Data Mining*, and has published numerous technical papers in peer-reviewed journals and conference proceedings. His research interests include data mining, statistics, and bioinformatics. He is a member of the IEEE and the ACM.



Vipin Kumar is currently William Norris Professor and Head of the Computer Science and Engineering Department at the University of Minnesota. He received the B.E. degree in electronics & communication engineering from University of Roorkee, India, in 1977, the M.E. degree in electronics engineering from Philips International Institute, Eindhoven, Netherlands, in 1979, and the Ph.D. degree in computer science from University of Maryland, College Park, in 1982. Kumar's current research interests include high-performance computing and data mining. His research has resulted in the development of the concept of isoefficiency metric for evaluating the scalability of parallel algorithms, as well as highly efficient parallel algorithms and software for sparse matrix factorization (PSPASES), graph partitioning (METIS, ParMetis, hMetis), and dense hierarchical solvers. He has authored over 200 research articles, and has coedited or coauthored 9 books including the widely used text books *Introduction to Parallel Computing* and *Introduction to Data Mining*, both published by Addison Wesley. Kumar has served as chair/co-chair for many conferences/workshops in the area of data mining and parallel computing, including the IEEE International Conference on Data Mining (2002) and the 15th International Parallel and Distributed Processing Symposium (2001). Kumar serves as the chair of the steering committee of the SIAM International Conference on Data Mining, and is a member of the steering committee of the IEEE International Conference on Data Mining. Kumar serves or has served on the editorial boards of *Data Mining and Knowledge Discovery*, *Knowledge and Information Systems*, *IEEE Computational Intelligence Bulletin*, *Annual Review of Intelligent Informatics*, *Parallel Computing*, *Journal of Parallel and Distributed Computing*, *IEEE Transactions of Data and Knowledge Engineering* (93-97), *IEEE Concurrency* (1997-2000), and *IEEE Parallel and Distributed Technology* (1995-1997). He is a Fellow of the ACM and IEEE, and a member of SIAM.

Correspondence and offprint requests to: Michael Steinbach, Department of Computer Science and Engineering, University of Minnesota, 4-192 EE/CS Building, 200 Union Street SE, Minneapolis, MN 55455, USA. Email: steinbac@cs.umn.edu