Selecting a Right Interestingness Measure for Rare Association Rules

Akshat Surana R. Uday Kiran P. Krishna Reddy

Center for Data Engineering
International Institute of Information Technology-Hyderabad
Hyderabad - 500032
India.
{akshat.surana, uday_rage}@research.iiit.ac.in and pkreddy@iiit.ac.in

Abstract

In the literature, the properties of several interestingness measures have been analyzed and a framework has been proposed for selecting a right interestingness measure for extracting association rules. As *rare association rules* contain useful knowledge, researchers are making efforts to investigate efficient approaches to extract the same. In this paper, we make an effort to analyze the properties of interestingness measures for determining the interestingness of *rare association rules*. Based on the analysis, we suggest a set of properties a user should consider while selecting a measure to find the interestingness of rare associations. The experiments on real-world datasets show that the measures that satisfy the suggested properties can determine the interestingness of *rare association rules*.

1 Introduction

Association rule mining [1] finds associations between the sets of items that occur together in a transactional database. Since the traditional confidence measure may not disclose truly interesting associations [4, 22], various interestingness measures have been discussed for mining association rules [16, 19, 21]. Several interestingness measures such as lift [5] and all-confidence [15] have been proposed in the literature and are found to be useful for discovering association rules.

In [19], several key properties of a measure have been discussed and it has been shown that each measure satisfies a different set of properties making it useful for a given application domain. The authors have also proposed an approach to select a desirable measure based on the sample number of rules.

In this paper, we focus on methods to determine interestingness of *rare association rules*. We now briefly discuss about *rare association rules*. It can be observed that real-world datasets are mostly non-uniform in nature containing both frequently and relatively infrequent (or rarely) occurring entities. A *rare association rule* refers to an association rule forming between either frequent and rare items or only rare items. In literature, it has been reported that there exists useful knowledge pertaining to rare entities [14, 20]. For example, in a super-market, costly and/or durable goods such as *Bed* and *Pillow* are relatively infrequently purchased than the low cost and/or perishable goods such as *Bread* and *Jam*. However, the association between the former set of items is more interesting as it generates relatively more revenue per unit. The rare cases are more difficult to detect and generalize from because they contain fewer data. Realizing the importance of rare knowledge patterns research efforts are going on to investigate improved approaches to extract rare knowledge patterns such as *rare association rules* and rare class identification [20].

In this paper, we make an effort to analyze the properties of interestingness measures proposed in [16, 19] sensitive to *rare association rules*. Based on the analysis, we suggest a set of properties a measure should satisfy for mining *rare association rules*. The experiments on the real-world datasets show that the measures which satisfy the suggested properties are able to extract *rare association rules*.

The rest of the paper is organized as follows. In Section 2, we discuss the background and the motivation. In Section 3, we discuss the various interestingness measures and the properties examined in this paper. In Section 4 we analyze the various properties and measures with respect to mining of *rare association rules*. Experimental analysis on real-world datasets has been discussed in Section 5. Finally we summarize and discuss future work in Section 6.

2 Background and Motivation

In this section, we explain the model of association rules. Next, we discuss about *rare association rules* and explain the motivation.

15th International Conference on Management of Data
COMAD 2010, Nagpur, India, December 8–10, 2010

2.1 Association Rules

Association rules are an important class of regularities that exist in a database. Since the introduction of association rules in [1], the problem of mining association rules from transactional databases has been actively studied in the data mining community [7, 14, 20]. The basic model of association rules is as follows [1]:

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of items and T be a set of transactions (dataset). Each transaction t is a set of items such that $t \subseteq I$. An itemset (or a **pattern**) X is a set of items $\{\mathrm{i}_1, \mathrm{i}_2, ..., \mathrm{i}_k\}$, $1 \leq \mathrm{k} \leq \mathrm{n}$, such that $X \subseteq I$. The itemset containing k number of items is called a ***k*-itemset**. An implication of the form $A \Rightarrow B$, where $A \subset I$, $B \subset I$ and $A \cap B = \emptyset$ is called an **association rule** iff,

(i) The **support** of $A \Rightarrow B$, denoted as $S(A \cup B) = \frac{f(A \cup B)}{|T|}$, is not less than the user specified minimum support threshold, *minsup*.

(ii) The **confidence** of $A \Rightarrow B$, denoted as $C(A \Rightarrow B) = \frac{S(A \cup B)}{S(A)}$, is not less than the user specified minimum confidence threshold, $minconf$.

where, $f(X)$ refers to the frequency of a pattern X and $|T|$ is the transactional database size.

> **Example 1:** Consider a supermarket containing items $bread$ and $pillow$. An association between these two items is as follows: $pillow \Rightarrow bread$ $[support = 1\%,\ confidence = 75\%]$. This rule says that 1% of all the customers have bought $bread$ and $pillow$ together, and 75% of those who bought $pillow$ have also bought $bread$.

Generally, association rule mining algorithms work in two steps. In the first step, all frequent patterns that satisfy *minsup* constraint are extracted. In the second step, all association rules that satisfy $minconf$ constraint are generated from frequent patterns [1].

2.2 Rare Association Rules

Rare items are the items having low support values. The frequent patterns consisting of only rare items or both frequent and rare items are called rare frequent patterns.

An association rule forming between either frequent and rare items or only rare items is a *rare association rule*. Otherwise, it is a *frequent association rule*. *Rare association rules* can provide useful information to the users.

Rare association rules require the extraction of rare frequent patterns. Mining of rare frequent patterns under single *minsup* and single $minconf$ constraint encounters a problem known as the *rare item problem* which is as follows. At high *minsup*, we miss the frequent patterns containing rare items because rare items cannot satisfy high *minsup* constraint. To mine the frequent patterns containing rare items, we should specify a low *minsup* value. However, low *minsup* can cause combinatorial explosion, producing too many frequent patterns in which some of them can be uninteresting to the user.

To confront *rare item problem*, efforts are being made in the literature to extract rare frequent patterns under "multiple *minsup* framework" [14]. In this framework, each item is associated with a *minimum item support* (*MIS*) value. Each pattern can satisfy a *minsup* depending upon the *MIS* values of the items within it. For this framework, an Apriori-like [2] approach known as Multiple Support Apriori (MSApriori) has been discussed to mine frequent patterns. The MSApriori algorithm suffers from performance problems of the Apriori algorithm. Hence, an FP-growth-like algorithm known as Conditional Frequent Pattern-growth (CFP-growth) has been discussed in [8]. In [11], a methodology was discussed to specify *minimum item supports* depending upon the respective support values. In [10], Improved CFP-growth algorithm was discussed to efficiently mine frequent patterns. In [12], the authors made an effort to efficiently mine frequent patterns in a dataset where item frequencies vary widely.

2.3 Motivation

After discovering frequent patterns, approaches based on "multiple *minsup*s framework" use $minconf$-based rule discovery technique proposed in [1] for mining association rules containing both frequent and rare items. However, $minconf$ constraint may not disclose truly interesting association rules [4, 22].

> **Example 2:** Consider the following market-basket data T from the grocery store, focusing on the purchase of tea and $coffee$. Let $f(tea, coffee) = 20$, $f(tea) = 25$, $f(coffee) = 90$ and $|T| = 100$. Using this data, we evaluate the association rule $\{tea\} \Rightarrow \{coffee\}$ to have support=20% and confidence=80%. In other words, it can be said that out of all the people who drink tea, 80% of them drink coffee. However, 90% of all the people drink coffee regardless of the fact that they drink tea or not. Thus, the knowledge that one drinks tea decreases the chances of a customer drinking coffee from 90% to 80%. Thus the rule, $\{tea\} \Rightarrow \{coffee\}$ is slightly misleading.

As a result various interestingness measures, such as *lift, correlation* and *all-confidence* have been proposed for discovering useful association rules. Each measure has its own selection bias that justifies the rationale for preferring a set of association rules over another. As a result, selecting a right interestingness measure for mining association rules is a tricky problem. To confront this problem, a framework has been suggested in [19] for selecting a right measure. In this framework, authors have discussed various properties of a measure and suggested to choose a measure depending on the properties interesting to the user. In this paper, we make an effort to identify a set of properties that a user should consider for mining *rare association rules*.

3 About Interestingness Measures

In this section, we explain the various interestingness measures and discuss the properties of an interestingness measure.

3.1 Interestingness Measures

Since the traditional confidence measure may not disclose truly interesting association patterns [4, 22], various interestingness measures have been discussed for mining association rules [16, 19, 21]. These interestingness measures can be classified into two types: *subjective* measures and *objective* measures.

Subjective measures take into account both the data and the user. A pattern is said to be subjectively interesting if it reveals unexpected information about the data or such knowledge which could lead to profitable results. To define a subjective measure, access to the user's domain or background knowledge about the data is required. *Subjective* measures recognize that a pattern of interest to one user may or may not be of interest to another user [9, 17]. In [17], the authors have proposed *unexpectedness* and *actionability* as the two measures of *subjective* interestingness. Negative encoding length and temporal description length have been used as subjective measures in [18] and [6], respectively.

Table 1: A 2x2 Contingency Table for variables A and B.

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

Objective measures are mostly based on the theories in probability, statistics, or information theory. The *objective* measures do not require any prior knowledge about the user or domain and they measure the interestingness of an association rule in terms of the structure and the underlying data used in the discovery process. An objective measure is usually computed based on the frequency counts tabulated in a **contingency table**. A typical contingency table for a pair of binary variables, A and B, is shown in Table 1. In this table, N represents the total number of transactions in a database, f_{10} represents the number of transactions containing A but not B, f_{01} represents the number of transactions containing B but not A, f_{11} represents the number of transactions containing both A and B, f_{00} represents the number of transactions that contain neither A nor B, f_{1+} represents the number of transactions containing A and f_{+1} represents the number of transactions containing B. In context of association rule mining, the variables A and B represent patterns. Henceforth, depending upon the context we use the terms "variable" and "pattern" interchangeably.

An objective measure can be either *symmetric* or *asymmetric*. For a typical 2×2 contingency table containing a pair of binary variables, A and B (see, Table 1), a measure M is said to be *symmetric* if $M(A,B) = M(B,A)$. Otherwise, M is *asymmetric*. In other words, a *symmetric* measure does not differentiate between the two association rules, $(A \Rightarrow B)$ and $(B \Rightarrow A)$; whereas, an *asymmetric* measure does differentiate between the two association rules, $(A \Rightarrow B)$ and $(B \Rightarrow A)$.

Table 2: Symmetric Interestingness Measures for the association rule $(A \Rightarrow B)$

Measure	Formula
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)	$\frac{Nf_{11}+Nf_{00}-f_{1+}f_{+1}-f_{0+}f_{+0}}{N^2-f_{1+}f_{+1}-f_{0+}f_{+0}}$
Lift (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)	$(\frac{f_{11}+f_{00}}{f_{1+}f_{+1}+f_{0+}f_{+0}}) \times$ $(\frac{N^2-f_{1+}f_{+1}-f_{0+}f_{+0}}{N-f_{11}-f_{00}})$
All-confidence (h)	min $[\ \frac{f_{11}}{f_{1+}}\ ,\ \frac{f_{11}}{f_{+1}}\ ]$
Imbalance Ratio (IR)	$\frac{\lvert f_{10}-f_{01}\rvert}{f_{11}+f_{10}+f_{01}}$
Jaccard (ζ)	$\frac{f_{11}}{f_{1+}+f_{+1}-f_{11}}$

Table 3: Asymmetric Interestingness Measures for the association rule $(A \Rightarrow B)$

Measure	Formula
Confidence ($conf$)	$\frac{f_{11}}{f_{1+}}$
Goodman-Kruskal (λ)	$\frac{(\sum_j max_k f_{jk} - max_k f_{+k})}{N - max_k f_{+k}}$
Mutual Information (M)	$\frac{(\sum_i \sum_j \frac{f_{ij}}{N} log \frac{Nf_{ij}}{f_{i+}f_{+j}})}{(-\sum_i \frac{f_{i+}}{N} log \frac{f_{i+}}{N})}$
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)	$\frac{f_{1+}}{N} \times \left[\frac{(f_{11})^2+(f_{10})^2}{(f_{1+})^2}\right] - (\frac{f_{+1}}{N})^2$ $+ \frac{f_{0+}}{N} \times \left[\frac{(f_{01})^2+(f_{00})^2}{(f_{0+})^2}\right] - (\frac{f_{+0}}{N})^2$
Laplace (L)	$(f_{11}+1)/(f_{1+}+2)$
Conviction (V)	$(f_{1+}f_{+0})/(Nf_{10})$
Certainty factor (F)	$(\frac{f_{11}}{f_{1+}} - \frac{f_{+1}}{N})/(1 - \frac{f_{+1}}{N})$
Added Value (AV)	$\frac{f_{11}}{f_{1+}} - \frac{f_{+1}}{N}$

List of *symmetric* and *asymmetric* measures that are examined in this paper are given in Tables 2 and 3, respectively.

3.2 Properties of a Measure

For a measure M, Piatetsky-Shapiro [16] has proposed the following three properties.

Property 1. *(P1) $M = 0$ if A and B are statistically independent.*

Property 2. *(P2) M monotonically increases with $P(A,B)$ when $P(A)$ and $P(B)$ remain the same.*

Table 4: **Properties of interestingness measures**

Symbol	Measure	Range	$P1$	$P2$	$P3$	$O1$	$O2$	$O3$	$O3'$	$O4$
ϕ	Correlation	$[-1,1]$	Yes	Yes	Yes	Yes	No	Yes	Yes	No
α	Odds ratio	$[0,\infty]$	Yes*	Yes	Yes	Yes	Yes	Yes*	Yes	No
κ	Kappa	$[-1,1]$	Yes	Yes	Yes	Yes	No	No	Yes	No
I	Lift	$[0,\infty)$	Yes*	Yes	Yes	Yes	No	No	No	No
IS	Cosine	$[0,1]$	No	Yes	Yes	Yes	No	No	No	Yes
PS	Piatetsky-Shapiro	$[-0.25,0.25]$	Yes	Yes	Yes	Yes	No	Yes	Yes	No
S	Collective strength	$[0,\infty)$	No	Yes	Yes	Yes	No	Yes*	Yes	No
h	All-confidence	$[0,1]$	No	Yes	Yes	Yes	No	No	No	Yes
IR	Imbalance Ratio	$[0,1]$	No	Yes	No	Yes	No	No	No	Yes
ζ	Jaccard	$[0,1]$	No	Yes	Yes	Yes	No	No	No	Yes
$conf$	Confidence	$[0,1]$	No	Yes	No	No	No	No	No	Yes
λ	Goodman-Kruskal's	$[0,1]$	Yes	No	No	No	No	No*	Yes	No
M	Mutual Information	$[0,1]$	Yes	Yes	Yes	No	No	No*	Yes	No
J	J-Measure	$[0,1]$	Yes	No	No	No	No	No	No	No
G	Gini index	$[0,1]$	Yes	No	No	No	No	No*	Yes	No
L	Laplace	$[0,1]$	No	Yes	No	No	No	No	No	No
V	Conviction	$[0.5,\infty)$	No	Yes	No	No	No	No	Yes	No
F	Certainty Factor	$[-1,1]$	Yes	Yes	Yes	No	No	No	Yes	No
AV	Added Value	$[-0.5,1]$	Yes	Yes	Yes	No	No	No	No	No

* P1, P2, P3, O1, O2, O3, O3' and O4 are discussed in Section 3
Yes* : Yes if measure is normalized
No* : Symmetry under row or column permutation

Property 3. *(P3) M monotonically decreases with $P(A)$ (or $P(B)$) when the rest of the parameters, $P(A,B)$ and $P(B)$ or $(P(A))$ remain unchanged.*

In [19], the authors have mapped a 2×2 contingency table as a 2×2 matrix formulation (**M**), i.e. $\mathbf{M} = \begin{bmatrix} f_{11} & f_{10} \\ f_{01} & f_{00} \end{bmatrix}$. The following properties (Properties 4–8) have been proposed by considering a measure as a matrix operator, O that maps the matrix **M** to a scalar value, k, i.e. $O(\mathbf{M}) = k$.

Property 4. ***Symmetry Under Variable Permutation (O1):*** *A measure O is symmetric under variable permutation, $A \leftrightarrow B$, if $O(M^T) = O(M)$ for all contingency matrices M. Otherwise it is called an asymmetric measure.*

Property 5. ***Row/Column Scaling Invariance (O2):*** *Consider two* 2×2 *matrices,* R *and* C *such that,* R = C = [k_1 0; 0 k_2]. *Now a measure O is said to be invariant under row scaling if $O(RM) = O(M)$, and is said to be invariant under column scaling if $O(MC) = O(M)$.*

Property 6. ***Antisymmetry Under Row/Column Permutation (O3):*** *For a* 2×2 *matrix* S = [0 1; 1 0], *a normalized measure O (i.e. for all contingency tables, M, $-1 \leq O(M) \leq 1$) is said to be antisymmetric under row permutation if $O(SM) = -O(M)$. Similarly, O is said to be antisymmetric under column permutation if $O(MS) = -O(M)$.*

Property 7. ***Inversion Invariance (O3'):*** *For a* 2×2 *matrix* S = [0 1; 1 0], *a measure O is said to be invariant under inversion operation if $O(SMS) = O(M)$.*

Property 8. ***Null Invariance (O4):*** *For a matrix* C = [0 0; 0 k], *a measure O is said to be null invariant if $O(M+C) = O(M)$.*

The above properties help the user to choose the interestingness measure depending on his/her requirements, i.e. they are subjective to user-interest. For example, if a user is interested in finding rules such that $(A \Rightarrow B) = (B \Rightarrow A)$, then a measure that has Property 4 ($O1$) should be selected for discovering association rules.

The list of properties either satisfied or not satisfied by a particular measure is shown in Table 4. This table also describes properties of the measures, *all-confidence* [15] and *imbalance ratio* [21] that are not discussed in [19].

4 Interestingness Measures for Mining Rare Association Rules

In this section, we first report our observation by performing an analysis on an example set of contingency tables. It is followed by discussion on the properties of an interestingness measure sensitive to *rare association rules*. Next, we discuss the framework to select an interestingness measures for mining *rare association rules*.

4.1 Analysis

Consider an example set of twelve contingency tables, E1 to E12, shown in Table 5. (The values described in these tables are based on the 2×2 contingency table shown in Table 1.) The contingency tables E1-E10 are taken from the work of Tan *et al.* [19]. It can be observed that the participating variables in Tables E1-E9 are frequent.

Table 5: Example of Contingency Tables.

Example	f_{11}	f_{10}	f_{01}	f_{00}
E1	8123	83	424	1370
E2	8330	2	622	1046
E3	3954	3080	5	2961
E4	2886	1363	1320	4431
E5	1500	2000	500	6000
E6	4000	2000	1000	3000
E7	9481	298	127	94
E8	4000	2000	2000	2000
E9	7450	2483	4	63
E10	61	2483	4	7452
E10′	61	4	2483	7452
E11	30	1	5	9964
E12	61	20	39	9880

To examine rare associations, we have added three new contingency tables, E10′, E11 and E12. Table E10 represents an association between a frequent variable and a rare variable. E10′ shows the transpose of the table E10. It is because for an asymmetric measure M, $M(A,B) \neq M(B,A)$. E11 and E12 contain less frequently occurring variables. However, relatively the variables in E12 are more frequent than those in E11.

We compute the association in each example by using measures mentioned in Tables 2 and 3. Each example is then ranked according to its measure in decreasing order of magnitude, as shown in Table 6. The following observations can be drawn from this table.

(i) Different measures can lead to substantially different orderings of contingency tables. For example, E11 is ranked highest by ϕ, α, κ, *lift* and *AV* measures, while it is ranked lowest by *PS* measure.

(ii) Some measures such as *IR* have given high ranking to the contingency tables that have high frequency variables (e.g. E7), while some measures such as *correlation (ϕ)* and *odds ratio (α)* have given high ranking to the contingency tables containing less frequent variables (e.g. E11).

(iii) Most important, variance of the rankings given by the measures is high in contingency tables containing either only frequent variables (E7) or only rare variables (E11). This shows that some measures have favored contingency tables containing high frequency variables (i.e., *frequent association rules*), while some others have favored contingency tables containing rare variables (i.e., *rare association rules*).

4.2 Properties Sensitive to Rare Association Rules

Let M be a given measure that is being considered for mining association rules in a transactional database. Here we discuss which properties are to be satisfied by M for discovering *rare association rules*.

Property 1 (i.e., $P1$) is not mandatory for M to discover *rare association rules*. It is because some measures may take a value other than 0 to represent the case when A, B are statistically independent (A and B are statistically independent means $P(A,B) = P(A) \times P(B)$). For example, *lift* takes the value 1 when A, B are statistically independent.

The measure M which satisfies Property 2 (i.e., $P2$) can be used to mine *rare association rules*. It is understandable that the association between A and B becomes more interesting when $P(A,B)$ increases while keeping both $P(A)$ and $P(B)$ constant.

The measure M which satisfies Property 3 (i.e., $P3$) is significant for mining *rare association rules*. It can be illustrated as follows. Consider two rare variables A and B. Now if $P(A)$ is increased keeping $P(B)$ and $P(A,B)$ constant, A no more remains a *rare variable*. Thus, the association between A and B, becomes less interesting.

Properties 4-7 help the user to choose the interestingness measure depending on his/her requirements, i.e. they are subjective to user-interest. For example, if a user is interested in finding rules such that $(A \Rightarrow B) = (B \Rightarrow A)$, then a measure that has Property 4 ($O1$) should be selected for discovering association rules. Thus, the properties *O1, O2, O3* and *O3'* are subjective to user interest.

For mining *rare association rules*, the *null invariance (O4)* property needs to be considered. A measure which satisfies the *null invariance* property is not influenced by the co-absence of the participating variables. A transaction is said to be a *null transaction* with respect to an association rule $A \Rightarrow B$ if neither A nor B is contained in that transaction. In case of a *rare association rule*, there will be a large number of *null transactions*. So to prevent the pruning of *rare association rules*, such a measure needs to be selected which does not take into account the *null transactions* while calculating the interestingness value.

Based on the above analysis, a measure that satisfies the properties, *P2, P3* and *O4* should be considered for mining *rare association rules*,

It can be observed from Table 4 that only *cosine (IS), all-confidence (h)* and *jaccard* (ζ) satisfy all the three properties among the *symmetric* measures listed in Table 2, and none of the *asymmetric* measures listed in Table 3 satisfy all the three properties. However, among the *asymmetric* measures, *mutual information (M), certainty factor (F)* and *added value (AV)* satisfy two ($P2$ and $P3$) out of the three properties. Therefore, any one of these measures can be considered for mining *rare association rules*.

4.3 Selecting a measure

In [19], the authors have shown that by selecting a small subset of "well-separated" contingency tables (or association rules), an appropriate measure to mine association rules can be found by comparing how well each measure agrees with the expectation of the users (or domain experts).

Given a dataset, the steps for selecting an appropriate interestingness measure to extract association rules are as follows.

Table 6: Rankings of contingency tables using different measures (Rank 1 being most interesting, 12 being least). The values under the column “Var” represents the variance of the ranks given by different measures for that particular contingency table. For symmetric measures, E10 $=$ E10$'$. Therefore, both tables receive same rank.

Example	Symmetric Measures											Asymmetric Measures									
	ϕ	α	κ	I	IS	PS	S	h	IR	ζ	Var	$conf$	λ	M	J	G	L	V	F	AV	Var
E1	2	5	2	8	2	2	2	2	4	2	4.100	2	2	2	1	1	4	3	3	7	3.444
E2	3	2	3	9	3	5	3	3	5	3	4.100	1	4	4	2	3	1	1	1	8	5.444
E3	5	4	6	6	6	1	5	10	11	7	8.100	10	7	5	5	2	3	8	8	6	6.500
E4	6	10	5	5	8	3	6	6	2	8	5.656	7	6	8	3	4	11	5	5	3	6.694
E5	7	9	8	4	11	6	8	11	10	11	5.611	11	9	9	4	6	9	7	7	4	5.750
E6	8	11	7	7	7	4	7	7	7	6	2.989	8	5	10	6	5	8	6	6	5	3.028
E7	9	8	9	11	1	8	9	1	3	1	16.000	3	9	7	11	9	5	9	9	11	7.111
E8	10	12	10	10	10	7	10	7	1	10	9.567	8	9	11	9	7	12	10	10	9	2.278
E9	11	6	11	12	5	10	11	5	9	5	8.500	6	8	6	12	10	2	11	11	12	11.750
E10	12	7	12	3	12	11	12	12	12	12	9.389	12	9	12	10	12	7	12	12	10	3.250
E11	1	1	1	1	4	12	1	4	6	4	12.278	4	1	1	8	11	6	2	2	1	13.000
E12	4	3	4	2	9	9	4	9	8	9	8.544	5	3	3	7	8	10	4	4	2	7.111
E10$'$	-	-	-	-	-	-	-	-	-	-	-	5	9	7	9	11	7	4	4	3	7.528

(i) A random sample of n contingency tables is chosen from the set of discovered frequent patterns.

(ii) Sample tables are ranked by users based on the perceived interestingness. We call this the user-given ranking vector U_r.

(iii) For each interestingness measure M, its value is computed for all the contingency tables. We calculate the ranks of these measure values as follows. The highest value receives the first rank; the second highest value receives the second rank and so on. The resulting ranking values are called a ranking vector for a measure M denoted as R_M.

(iv) The similarity between user-given ranking vector (U_r) and the ranking vectors of each measure (R_M) is computed. The measure with the highest similarity value is selected as the interestingness measure for mining association rules for that dataset.

The similarity between two measures is equivalent to the similarity value between the corresponding ranking vectors. A similarity measure such as *cosine similarity* or *Pearson's correlation* can be used to find the similarity between two ranking vectors.

The sample set of contingency tables is chosen randomly from the set of discovered frequent patterns. For mining *rare association rules*, the contingency tables related to rare frequent patterns should be included as a part of sample set. In other words, considering a 2×2 contingency table for variables A and B, the sample set is chosen such that it contains the following combinations: (i) both A and B are frequent variables, (ii) both A and B are rare variables, (iii) A is a frequent variable while B is a rare variable, and (iv) A is a rare variable while B is a frequent variable.

5 Experimental Analysis

In this section, we performed experimental analysis on various real-world datasets available at Frequent Itemset MIning (FIMI) Repository (http://fimi.cs.helsinki.fi/data/). We confine our analysis to the following real-world datasets: Retail and BMS-WebView-1. Retail dataset [3] is a large sparse dataset containing 16,470 distinct items in 88,162 transactions. BMS-WebView-1 [13] is also a large sparse dataset with 497 distinct items in 59,602 transactions.

To mine *rare association rules*, we used "multiple *minsup*s framework" [14] to discover the set of frequent patterns F from a dataset D. In this approach, each item is associated with a "minimum item support" (MIS) value. Let $S(i)$ denote the support/frequency of an item i and $MIS(i)$ denote the minimum item support value for item i. To specify items' MIS values, we use the approach given in [14], which is as follows.

$$MIS(i) = min(\beta \times S(i), LS) \quad (1)$$

where β ($0 \leq \beta \leq 1$) is a parameter that controls how the MIS values for items should be related to their frequencies (or supports) and LS is the user-specified lowest minimum item support. For both the datasets, Retail and BMS-WebView-1, we have set $\beta = 0.05$ and $LS = 0.1\%$.

5.1 Experiment 1: Similarity between various measures

In this experiment, we analyzed the performance of the measures which possess the properties sensitive to *rare association rules* against the other measures.

For a given dataset, we extracted the set of frequent patterns. Next, a sample of frequent patterns was selected and all corresponding contingency tables were generated. We find the similarity between all pairs of measures by finding the *Pearson's correlation* measure between the corresponding *ranking vectors*.

By considering only symmetric measures, the similarity matrices computed for Retail and BMS-WebView-1 datasets are shown in Tables 7 and 8, respectively. Since the similarity matrices are symmetric in nature, only the lower triangular part of the matrices is presented in the tables. It can be observed that in both the datasets, the measures, *jaccard* (ζ), *all-confidence (h)* and *cosine (IS)* are highly similar to each other (the corresponding similarity values are underlined in Tables 7 and 8) because all three measures share the three properties, $P2$, $P3$ and $O4$ (refer Section 3) sensitive to *rare association rules*.

By considering only asymmetric measures, the similarity matrices computed for Retail and BMS-WebView-1 datasets are shown in Tables 9 and 10, respectively. It can be observed that in both the datasets, the measures, *certainty factor (F), added value (AV)* and *mutual information (M)* are highly similar (the corresponding similarity values are underlined in Tables 9 and 10) to each other because all three measures share two ($P2$ and $P3$) properties sensitive to *rare association rules*.

Overall the experiments show that the measures satisfying the properties sensitive to *rare association rules* exhibit high similarity. Any one of the measures which are possessing properties sensitive to *rare association rules* can be chosen for extracting *rare association rules*. The measure should be selected based on the suitability of other properties of the measure for the data mining task.

5.2 Experiment 2: Measure Selection for Mining Rare Association Rules

We have carried out an experiment to select a measure for extracting *rare association rules* for a given dataset. After extracting frequent patterns, we have selected a sample of 15 contingency tables from each of the two datasets, Retail and BMS-WebView-1. A group of users examined the contingency tables and gave the appropriate ranks. To give importance to *rare association rules*, the users gave high ranks to *rare association rules*.

Table 11 and Table 12 show the sample contingency tables chosen from the Retail and BMS-WebView-1 datasets, respectively. The last column in these tables shows the rankings provided by the users. It can be observed that the

Table 7: Similarity between different *symmetric* measures for Retail dataset

	ϕ	α	κ	I	IS	PS	S	h	IR	ζ
ϕ	1									
α	0.95	1								
κ	0.718	0.796	1							
I	0.957	0.979	0.814	1						
IS	0.943	0.875	0.668	0.875	1					
PS	0.646	0.521	0.043	0.496	0.639	1				
S	-0.593	-0.657	-0.811	-0.693	-0.507	0.189	1			
h	0.907	0.845	0.685	0.871	0.951	0.532	-0.594	1		
IR	0.768	0.671	0.614	0.743	0.779	0.389	-0.557	0.882	1	
ζ	0.889	0.827	0.671	0.845	0.958	0.525	-0.572	0.996	0.86	1

Table 8: Similarity between different *symmetric* measures for BMS-WebView-1 dataset

	ϕ	α	κ	I	IS	PS	S	h	IR	ζ
ϕ	1									
α	0.754	1								
κ	0.939	0.596	1							
I	0.707	0.964	0.6	1						
IS	0.993	0.707	0.95	0.65	1					
PS	0.696	0.196	0.675	0.064	0.732	1				
S	-0.771	-0.8	-0.796	-0.871	-0.732	-0.2	1			
h	0.821	0.375	0.921	0.346	0.854	0.693	-0.643	1		
IR	-0.104	-0.464	0.168	-0.425	-0.057	0.175	-0.014	0.411	1	
ζ	0.879	0.479	0.971	0.482	0.9	0.657	-0.739	0.975	0.304	1

Table 9: Similarity between different *asymmetric* measures for Retail dataset

	$conf$	λ	M	J	G	L	V	F	AV
$conf$	1								
λ	0.617	1							
M	0.657	0.746	1						
J	-0.033	0.402	0.318	1					
G	0.343	0.198	0.493	-0.208	1				
L	0.686	0.449	0.471	0.293	-0.232	1			
V	0.75	0.777	0.932	0.384	0.411	0.532	1		
F	0.768	0.777	0.946	0.249	0.429	0.511	0.975	1	
AV	0.768	0.777	0.946	0.249	0.429	0.511	0.975	1	1

Table 10: Similarity between different *asymmetric* measures for BMS-WebView-1 dataset

	$conf$	λ	M	J	G	L	V	F	AV
$conf$	1								
λ	0.541	1							
M	0.961	0.57	1						
J	0.668	0.085	0.682	1					
G	-0.254	-0.399	-0.418	0.096	1				
L	0.461	0.541	0.568	0.336	-0.371	1			
V	0.986	0.541	0.986	0.696	-0.311	0.554	1		
F	0.986	0.541	0.986	0.696	-0.311	0.554	1	1	
AV	0.986	0.541	0.986	0.696	-0.311	0.554	1	1	1

Table 11: Sample set of contingency tables taken from Retail dataset

	f_{11}	f_{10}	f_{01}	f_{00}	U_r
T1	130	74	71	87887	1
T2	124	127	56	87855	2
T3	106	41	120	87895	3
T4	99	175	201	87687	4
T5	106	90	138	87828	5
T6	5402	3053	36733	42974	6
T7	224	3673	3033	81232	7
T8	1740	19	13856	72547	8
T9	1206	853	40929	45174	9
T10	1416	40719	1520	44507	10
T11	98	50577	88	37399	11
T12	1116	41019	921	45106	12
T13	93	39	50582	37448	13
T14	93	48	50582	37439	14
T15	92	50583	49	37438	15

Table 12: Sample set of contingency tables taken from BMS-WebView-1 dataset

	f_{11}	f_{10}	f_{01}	f_{00}	U_r
T1	118	33	128	59323	1
T2	359	549	821	57873	2
T3	102	114	218	59168	3
T4	1204	2408	2454	53536	4
T5	615	1333	807	56847	5
T6	307	806	601	57888	6
T7	771	2678	1600	54553	7
T8	496	2301	2953	53852	8
T9	60	622	664	58256	9
T10	62	434	528	58578	10
T11	63	8	1425	58106	11
T12	72	2196	50	57284	12
T13	228	3430	2569	53375	13
T14	83	3529	52	55938	14
T15	68	3544	138	55852	15

contingency table $T1$ has been given high rank because it represents an association between two rare variables, and $T15$ has been given a low rank because it is an association between a highly frequent and a rare variable.

The *Pearson's correlation measure* has been used to find the similarity between the ranking vectors.

For symmetric measures, the similarity values between ranking vector of each measure (R_M) and user given ranking vector (U_r) is given in Tables 13 and 14 for Retail and BMS-WebView-1 datasets, respectively. It can be observed that the rankings given by the measures, *jaccard (ζ), all-confidence (h)* and *cosine (IS)* are highly similar to the user-specified ranking vectors U_r.

Similarly, for asymmetric measure, the similarity values between ranking vector of each measure (R_M) and user given ranking vector (U_r) is given in Tables 15 and 16 for Retail and BMS-WebView-1 datasets, respectively. It can be observed that the ranks given by *mutual information* (M) measure are the most similar to the actual rankings expected by the users. *Added value* (AV) and *certainty factor* (F) also give ranks similar to the ranks expected by the users.

From Experiment 2, we can conclude that the measures satisfying the suggested properties can be used for mining *rare association rules*, giving rankings similar to the user-perceived rankings.

6 Conclusion

In this paper, we have analyzed how various interestingness measures perform in extracting *rare association rules*. Through analysis, we found out that the measures which possess certain properties are appropriate for extracting *rare association rules*. By carrying out experiments on real world datasets, it has been shown that the measures satisfying the prescribed properties are able to mine *rare association rules*.

It can be observed that a single measure may not be appropriate to mine interesting association rules from all kinds of frequent patterns (both rare and frequent) for a given dataset. So as a part of future work, we will make an effort to investigate the approaches to extract association rules by dividing the frequent patterns extracted from the dataset into multiple groups and applying an appropriate interestingness measure for each group.

Acknowledgement

We thank the anonymous referees for the useful comments. We are also thankful to all authors for providing their datasets in FIMI repository for conducting experiments.

References

[1] R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In *SIGMOD Conference*, pages 207–216, 1993.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *VLDB*, pages 487–499, 1994.

[3] T. Brijs, G. Swinnen, K. Vanhoof, and G. Wets. Using association rules for product assortment decisions: A case study. In *KDD*, pages 254–260, 1999.

[4] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *SIGMOD Conference*, pages 265–276, 1997.

[5] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *SIGMOD Conference*, pages 255–264, 1997.

Table 13: Ranking of *symmetric* measures for sample Retail data

Measure M	ϕ	α	κ	I	IS	PS	S	h	IR	ζ
$Sim(R_M, U_r)$	0.893	0.825	0.661	0.861	0.914	0.496	-0.614	0.984	0.896	0.973

Table 14: Ranking of *symmetric* measures for sample BMS-WebView-1 data

Measure M	ϕ	α	κ	I	IS	PS	S	h	IR	ζ
$Sim(R_M, U_r)$	0.875	0.5	0.954	0.489	0.896	0.675	-0.743	0.961	0.321	0.975

Table 15: Ranking of *asymmetric* measures for sample Retail data

Measure M	$conf$	λ	M	J	G	L	V	F	AV
$Sim(R_M, U_r)$	0.279	0.586	0.807	0.282	0.593	-0.018	0.711	0.729	0.729

Table 16: Ranking of *asymmetric* measures for sample BMS-WebView-1 data

Measure M	$conf$	λ	M	J	G	L	V	F	AV
$Sim(R_M, U_r)$	0.793	0.313	0.804	0.8	-0.286	0.214	0.789	0.789	0.789

[6] S. Chakrabarti, S. Sarawagi, and B. Dom. Mining surprising patterns using temporal description length. In *VLDB*, pages 606–617, 1998.

[7] B. Goethals. Survey on frequent pattern mining. Technical report, 2003.

[8] Y.-H. Hu and Y.-L. Chen. Mining association rules with multiple minimum supports: a new mining algorithm and a support tuning mechanism. *Decision Support Systems*, 42(1):1–24, 2006.

[9] M. Kamber and R. Shinghal. Evaluating the interestingness of characteristic rules. In *KDD*, pages 263–266, 1996.

[10] R. U. Kiran and P. K. Reddy. An improved frequent pattern-growth approach to discover rare association rules. In *KDIR*, pages 43–52, 2009.

[11] R. U. Kiran and P. K. Reddy. An improved multiple minimum support based approach to mine rare association rules. In *CIDM*, pages 340–347, 2009.

[12] R. U. Kiran and P. K. Reddy. Mining rare association rules in the datasets with widely varying items' frequencies. In *DASFAA (1)*, pages 49–62, 2010.

[13] R. Kohavi, C. Brodley, B. Frasca, L. Mason, and Z. Zheng. KDD-Cup 2000 organizers' report: Peeling the onion. *SIGKDD Explorations*, 2(2):86–98, 2000.

[14] B. Liu, W. Hsu, and Y. Ma. Mining association rules with multiple minimum supports. In *KDD*, pages 337–341, 1999.

[15] E. Omiecinski. Alternative interest measures for mining associations in databases. *IEEE Trans. Knowl. Data Eng.*, 15(1):57–69, 2003.

[16] G. Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. In *Knowledge Discovery in Databases*, pages 229–248. AAAI/MIT Press, 1991.

[17] A. Silberschatz and A. Tuzhilin. On subjective measures of interestingness in knowledge discovery. In *KDD*, pages 275–281, 1995.

[18] E. Suzuki. Negative encoding length as a subjective interestingness measure for groups of rules. In *PAKDD*, pages 220–231, 2009.

[19] P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *KDD*, pages 32–41, 2002.

[20] G. M. Weiss. Mining with rarity: a unifying framework. *SIGKDD Explorations*, 6(1):7–19, 2004.

[21] T. Wu, Y. Chen, and J. Han. Re-examination of interestingness measures in pattern mining: a unified framework. *Data Mining and Knowledge Discovery*, 2010.

[22] X. Wu, C. Zhang, and S. Zhang. Efficient mining of both positive and negative association rules. *ACM Trans. Inf. Syst.*, 22(3):381–405, 2004.