
Selecting a Right Interestingness Measure for Rare Association Rules

**Akshat Surana, R. Uday Kiran and
P. Krishna Reddy**

**International Institute of Information Technology-Hyderabad,
Hyderabad, India
Email : {akshat.surana, uday_rage}@research.iiit.ac.in and
pkreddy@iiit.ac.in**

Outline

- Introduction
- Related Work
- Interestingness Measures for Mining Rare Association Rules
- Experimental Analysis
- Conclusions and Future Work

Outline

- **Introduction**
- Related Work
- Interestingness Measures for Mining Rare Association Rules
- Experimental Analysis
- Conclusions and Future Work

Association Rules

- Association rules is an important class of data mining technique like clustering and classification.
- Model of association rules under support/confidence framework
 - T be a transactional database.
 - $I = \{i_1, i_2, \dots, i_n\}$ be a set of items.
 - An association rule is of form $A \rightarrow B$, where A and B are subsets of I such that $A \cap B \neq \Phi$.
 - The terms **Support** and **Confidence** are defined as follows.
 - $\text{support}(A \cup B) = f(A \cup B)/N$.
 - $\text{confidence}(A \rightarrow B) = \text{Support}(A \cup B) / \text{Support}(A)$.
 - An association rule is interesting if it satisfies user-specified **minimum support (*minsup*)** and **minimum confidence (*minconf*)** thresholds.

Association Rule Mining: An Example

TID	Items
1	a, b
2	a, b, h
3	c, d
4	a, d
5	c, d
6	a, b
7	c, d
8	a, b
9	c, d, g
10	a, c, d

TID	Items
11	a, b, d
12	c, d
13	a, b
14	a, c, e
15	a, b, e
16	c, d
17	a, c
18	b, e, f
19	a, e, f
20	b, e, f, g

Table 1: Transactional database

- $I = \{a, b, c, d, e, f, g, h\}$. Let $\{a, b\}$ be a pattern.
 - $Support(\{a, b\}) = 7/20 = 0.35$ (or 35%)
 - $Confidence(\{a\} \rightarrow \{b\})$
 $= Support(\{a, b\}) / Support(\{a\})$
 $= (7/20) / (12/20)$
 $= 0.58$ (or 58%)
 - $Confidence(\{b\} \rightarrow \{a\}) =$
 $= Support(\{a, b\}) / Support(\{b\})$
 $= (7/20) / (9/20)$
 $= 0.78$ (or 78%)
- If user-specified $minsup = 30\%$ and $minconf = 75\%$, then
 - $\{a, b\}$ is a frequent pattern
 - $(\{b\} \rightarrow \{a\})$ is an interesting association rule
 - $(\{a\} \rightarrow \{b\})$ is NOT an interesting association rule

Contingency Table

- An association between a pair of patterns A and B can be represented using a 2 x 2 contingency table.

	B	\bar{B}	
A	f_{11} <i>Both A, B occur</i>	f_{10} <i>A occurs, B does not</i>	f_{1+} <i>A occurs</i>
\bar{A}	f_{01} <i>B occurs, A does not</i>	f_{00} <i>Neither A occurs, nor B</i>	f_{0+} <i>A does not occur</i>
	f_{+1} <i>B occurs</i>	f_{+0} <i>B does not occur</i>	N <i>Total transactions</i>

Table 2 : A 2x2 contingency table is shown for variables A and B.

- E.g. $Confidence(A \rightarrow B) = Support(A \cup B) / Support(A) = f_{11} / f_{1+} .$

Contingency Table: An Example

TID	Items	TID	Items
1	a, b	11	a, b, d
2	a, b, h	12	c, d
3	c, d	13	a, b
4	a, d	14	a, c, e
5	c, d	15	a, b, e
6	a, b	16	c, d
7	c, d	17	a, c
8	a, b	18	b, e, f
9	c, d, g	19	a, e, f
10	a, c, d	20	b, e, f, g

	<i>a</i>	\bar{a}	
<i>b</i>	7	2	9
\bar{b}	5	6	11
	12	8	20

Table 3 :Contingency table for $(\{b\} \rightarrow \{a\})$.

Table 1: Transactional database

Contingency Table: An Example

TID	Items	TID	Items
1	a, b	11	a, b, d
2	a, b, h	12	c, d
3	c, d	13	a, b
4	a, d	14	a, c, e
5	c, d	15	a, b, e
6	a, b	16	c, d
7	c, d	17	a, c
8	a, b	18	b, e, f
9	c, d, g	19	a, e, f
10	a, c, d	20	b, e, f, g

	<i>a</i>	\bar{a}	
<i>b</i>	7	2	9
\bar{b}	5	6	11
	12	8	20

Table 3 :Contingency table for $(\{b\} \rightarrow \{a\})$.

Table 1: Transactional database

Contingency Table: An Example

TID	Items	TID	Items
1	a, b	11	a, b, d
2	a, b, h	12	c, d
3	c, d	13	a, b
4	a, d	14	a, c, e
5	c, d	15	a, b, e
6	a, b	16	c, d
7	c, d	17	a, c
8	a, b	18	b, e, f
9	c, d, g	19	a, e, f
10	a, c, d	20	b, e, f, g

	<i>a</i>	\bar{a}	
<i>b</i>	7	2	9
\bar{b}	5	6	11
	12	8	20

Table 3 :Contingency table for $(\{b\} \rightarrow \{a\})$.

Table 1: Transactional database

Contingency Table: An Example

TID	Items	TID	Items
1	a, b	11	a, b, d
2	a, b, h	12	c, d
3	c, d	13	a, b
4	a, d	14	a, c, e
5	c, d	15	a, b, e
6	a, b	16	c, d
7	c, d	17	a, c
8	a, b	18	b, e, f
9	c, d, g	19	a, e, f
10	a, c, d	20	b, e, f, g

Table 1: Transactional database

	<i>a</i>	\bar{a}	
<i>b</i>	7	2	9
\bar{b}	5	6	11
	12	8	20

Table 3 :Contingency table for $(\{b\} \rightarrow \{a\})$.

- $$\text{Confidence}(\{b\} \rightarrow \{a\}) = f_{11} / f_{1+}$$

$$= 7/9 = 0.78 \text{ (or 78\%)}$$

Issue with Confidence

	<i>Coffee</i>	<i>Coffee</i>	
<i>Tea</i>	20	5	25
<i>Tea</i>	70	5	75
	90	10	100

- Confidence measure may not disclose truly interesting associations (SIGMOD '97).
 - **Example:** $Confidence(tea \rightarrow coffee) = 20/25 = 80\%$.
However, 90% of all people drink coffee regardless of the fact they drink tea or not.
- Hence, various alternative interestingness measures have been proposed in the literature.

Alternative Interestingness Measure : Lift

Confidence

	<i>B</i>	<i>\bar{B}</i>	
<i>A</i>	<i>f₁₁</i>	<i>f₁₀</i>	<i>f₁₊</i> +
<i>\bar{A}</i>	<i>f₀₁</i>	<i>f₀₀</i>	<i>f₀₊</i> +
	<i>f₊₁</i>	<i>f₊₀</i>	<i>N</i>

Lift

	<i>B</i>	<i>\bar{B}</i>	
<i>A</i>	<i>f₁₁</i>	<i>f₁₀</i>	<i>f₁₊</i> +
<i>\bar{A}</i>	<i>f₀₁</i>	<i>f₀₀</i>	<i>f₀₊</i> +
	<i>f₊₁</i>	<i>f₊₀</i>	<i>N</i>

- $$\text{Confidence}(A \rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)} = \frac{f_{11}}{f_{1+}}$$

- $$\text{Lift}(A \rightarrow B) = \frac{\text{Support}(A \cup B)}{(\text{Support}(A) * \text{Support}(B))} = \frac{N * f_{11}}{(f_{1+} * f_{+1})}$$

Alternative Interestingness Measure : All-Confidence and Cosine Measures

All-Confidence

	<i>B</i>	\overline{B}	
<i>A</i>	f_{11}	f_{10}	f_{1+}
\overline{A}	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	<i>N</i>

Cosine

	<i>B</i>	\overline{B}	
<i>A</i>	f_{11}	f_{10}	f_{1+}
\overline{A}	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	<i>N</i>

- $All-Confidence(A \rightarrow B) = \min(f_{11} / f_{1+}, f_{11} / f_{+1})$

- $Cosine(A \rightarrow B) = f_{11} / \sqrt{f_{1+} f_{+1}}$

Rare Association Rules

- Real world datasets contain both frequent and rare items.
- A rare association rule is an association rule containing rare items.
- Knowledge pertaining to rare items can provide useful information.
 - E.g. {Bed} → {Pillow} is more interesting than {Jam} → {Bread}.

Issues While Mining Rare Association Rules

- Issue 1: Mining frequent patterns containing both frequent and rare items leads to *rare item problem*.

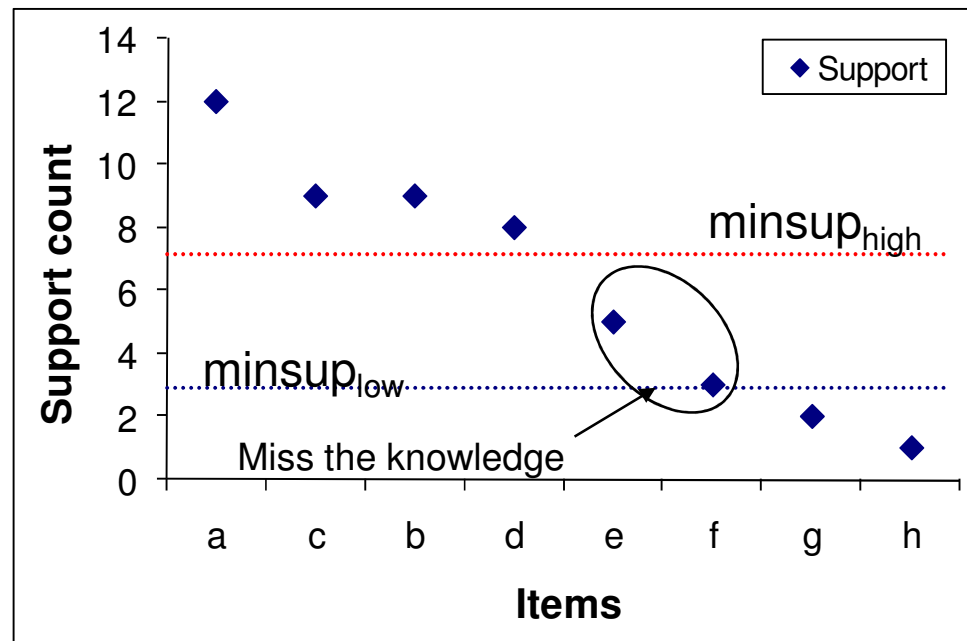


Figure 1. Items vs. Support count

Issues While Mining Rare Association Rules

- Issue 2: Selecting a measure to mine rare association rules.
 - No measure consistently performs better over others.
 - Each measure has its own selection bias.
 - Which measures can mine rare association rules is unknown.

Contribution of this Paper

- We address the problem of selecting an interestingness measure for mining rare association rules.
- Analyze various properties of a measure and suggest a set of properties to be considered for selecting a measure.
- Through experimental results, we show that the measures satisfying the suggested properties can efficiently discover rare association rules.

Outline

- Introduction
- **Related Work**
- Interestingness Measures for Mining Rare Association Rules
- Experimental Analysis
- Conclusions and Future Work

Related Work

- To address rare item problem, efforts have been made in the literature to mine frequent patterns using multiple minimum support framework (KDD '99, KDIR 2009, CIDM 2009, DASFAA 2010).
- Each item is specified with a *minimum item support* (MIS).
- Minsup of a pattern is specified with respect to MIS values of the items with it.
 - $\text{Minsup}(i_1, i_2, \dots, i_k) = \min(\text{MIS}(i_1), \text{MIS}(i_2), \dots, \text{MIS}(i_k))$
where, $\text{MIS}(i_j)$ is MIS value of the i_j , $1 \leq j \leq k$.

Minimum constraint model: Illustration

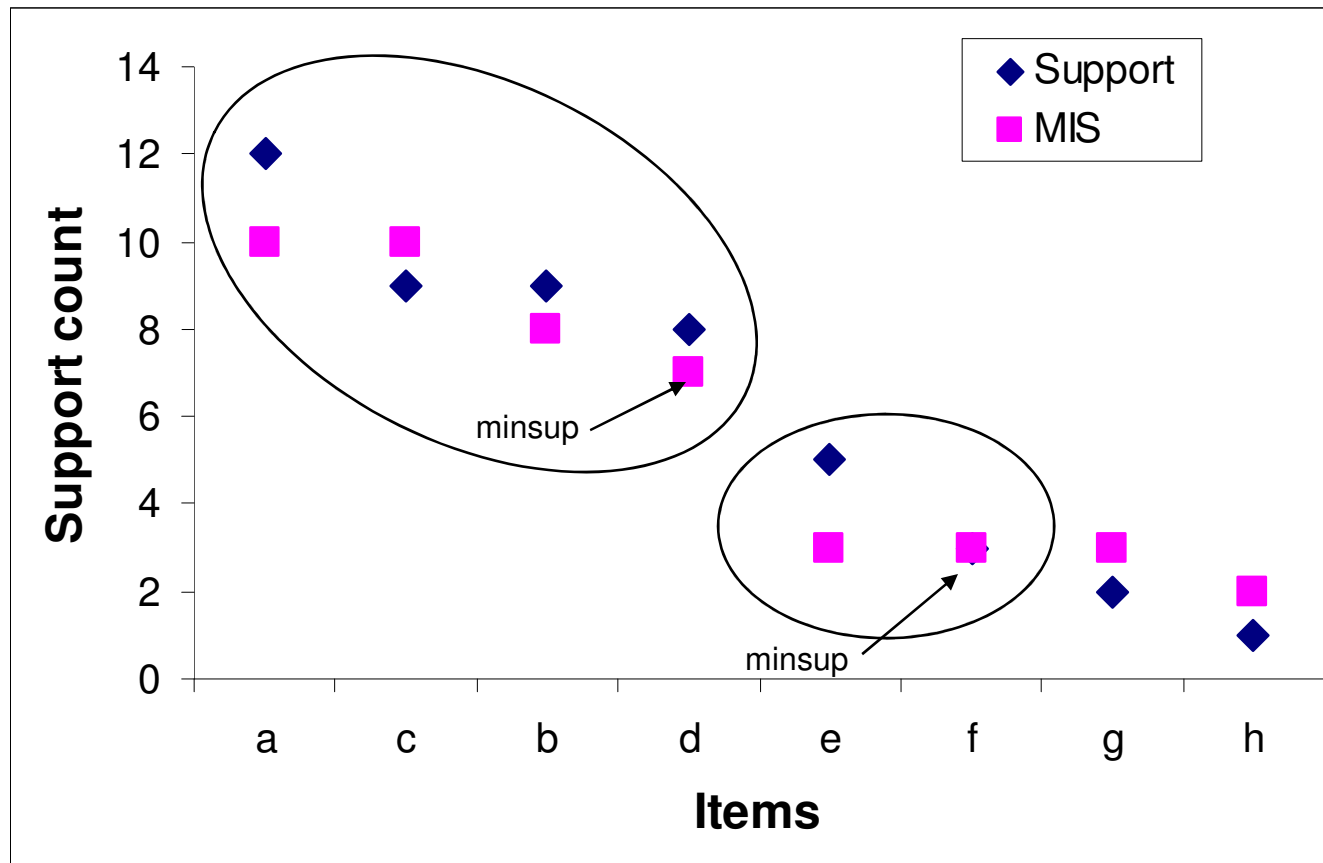


Figure 2. Minsup's specified for the patterns in minimum constraint model.

Minimum Constraint Model: Illustration

Pattern	Support count	MIS	Single minsup	multiple minsup
{a}	12	10	✓	✓
{c}	9	10	✓	✓
{b}	9	8	✓	✓
{d}	8	7	✓	✓
{e}	5	3	✓	✓
{f}	3	3	✓	✓
{a, b}	8	-	✓	✓
{a, c}	3	-	✓	✗
{a, e}	3	-	✓	✓
{c, d}	7	-	✓	✓
{e, f}	3	-	✓	✓

Table 2. Frequent patterns generated in different models.

Related Work

- Tan *et. al* (KDD 2002) have showed that all measures have different selection criteria.
- Introduced several properties and suggested to select a measure depending on the properties interesting to the user.
- However, they have not discussed which properties a user should consider for mining rare association rules.

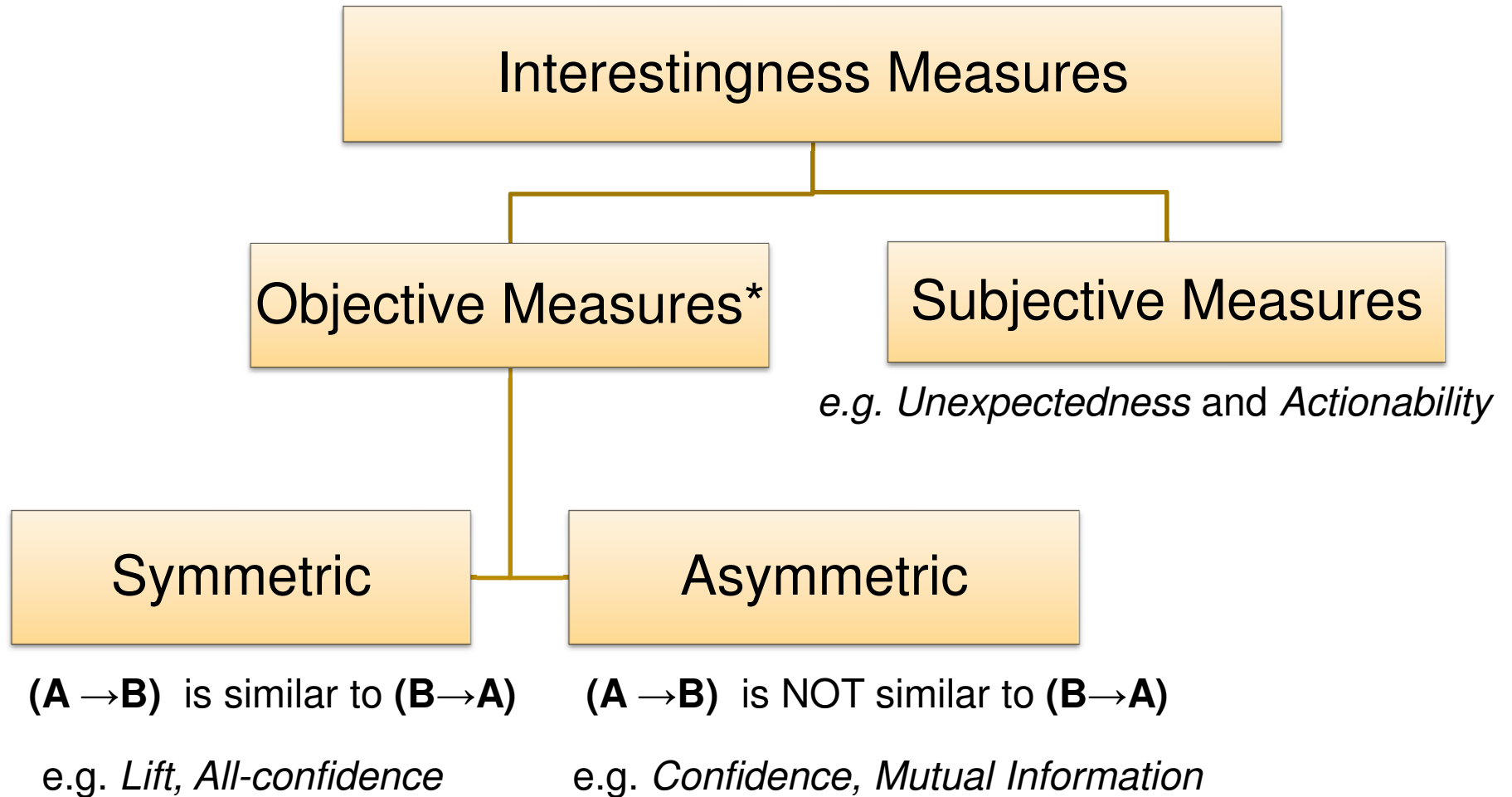
Outline

- Introduction
- Related Work
- **Interestingness Measures for Mining Rare Association Rules**
- Experimental Analysis
- Conclusions and Future Work

What is An Interestingness Measure?

- A measure used to assess the interestingness of a relationship between a pair of patterns.
- They are mostly based on the theories in probability, statistics and information theory.

Types of Interestingness Measure



List of Measures Studied

Correlation (ϕ)
Odds ratio (α)
Kappa (κ)
Lift (I)
Cosine (IS)
Piatetsky-Shapiro (PS)
Collective strength (S)
All-confidence (h)
Imbalance Ratio (IR)
Jaccard (ζ)

Table 4: **Symmetric Measures**

Confidence ($conf$)
Goodman-Kruskal (λ)
Mutual Information (M)
J-Measure (J)
Gini index (G)
Laplace (L)
Conviction (V)
Certainty factor (F)
Added Value (AV)

Table 5: **Asymmetric measures**

Analysis of Various Interestingness Measures

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

Both variables are frequent

One variable is frequent and another variable is rare

Both variables are rare

Table 6: Example of Contingency Tables

Analysis for Symmetric Measures

Table 7: Ranking of Contingency Tables using different Symmetric measures

Example	Symmetric Measures										Var
	ϕ	α	κ	I	IS	PS	S	h	IR	ζ	
E1	2	5	2	8	2	2	2	2	4	2	4.100
E2	3	2	3	9	3	5	3	3	5	3	4.100
E3	5	4	6	6	6	1	5	10	11	7	8.100
E4	6	10	5	5	8	3	6	6	2	8	5.656
E5	7	9	8	4	11	6	8	11	10	11	5.611
E6	8	11	7	7	7	4	7	7	7	6	2.989
E7	9	8	9	11	1	8	9	1	3	1	16.000
E8	10	12	10	10	10	7	10	7	1	10	9.567
E9	11	6	11	12	5	10	11	5	9	5	8.500
E10	12	7	12	3	12	11	12	12	12	12	9.389
E11	1	1	1	1	4	12	1	4	6	4	12.278
E12	4	3	4	2	9	9	4	9	8	9	8.544
E10'	-	-	-	-	-	-	-	-	-	-	-

Analysis for Asymmetric Measures

Table 8: Ranking of Contingency Tables using different Asymmetric measures

Example	Asymmetric Measures									Var
	<i>conf</i>	λ	<i>M</i>	<i>J</i>	<i>G</i>	<i>L</i>	<i>V</i>	<i>F</i>	<i>AV</i>	
E1	2	2	2	1	1	4	3	3	7	3.444
E2	1	4	4	2	3	1	1	1	8	5.444
E3	10	7	5	5	2	3	8	8	6	6.500
E4	7	6	8	3	4	11	5	5	3	6.694
E5	11	9	9	4	6	9	7	7	4	5.750
E6	8	5	10	6	5	8	6	6	5	3.028
E7	3	9	7	11	9	5	9	9	11	7.111
E8	8	9	11	9	7	12	10	10	9	2.278
E9	6	8	6	12	10	2	11	11	12	11.750
E10	12	9	12	10	12	7	12	12	10	3.250
E11	4	1	1	8	11	6	2	2	1	13.000
E12	5	3	3	7	8	10	4	4	2	7.111
E10'	5	9	7	9	11	7	4	4	3	7.528


Properties of a Measure

- Piatetsky-Shapiro (KDD '91) introduced the following properties for selecting a right interestingness measure M .
 - **P1:** $M = 0$ if A and B are statistically independent.
 - **P2:** M monotonically increases with $P(A,B)$ when $P(A)$ and $P(B)$ remain the same.
 - **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.

Properties of a Measure

- *Tan et. al* (KDD 2002) proposed the following properties by
 - Mapping the 2x2 contingency table to a 2x2 matrix \mathbf{M} .
 - Considering a measure to be a matrix operator O .
- **Symmetry Under Variable Permutation (O1):** A measure O is symmetric under variable permutation, $A \leftrightarrow B$, if $O(MT) = O(M)$ for all contingency matrices M . Otherwise it is called an asymmetric measure.

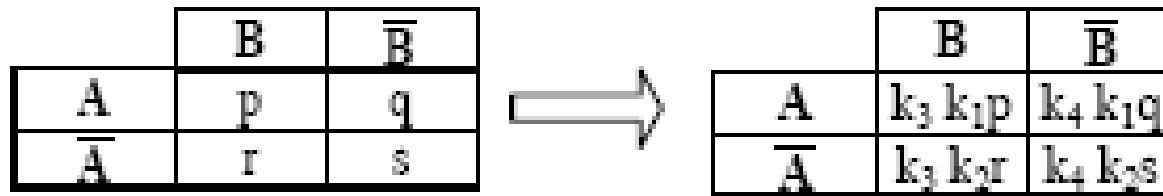
	B	\bar{B}
A	p	q
\bar{A}	r	s



	A	\bar{A}
B	p	r
\bar{B}	q	s

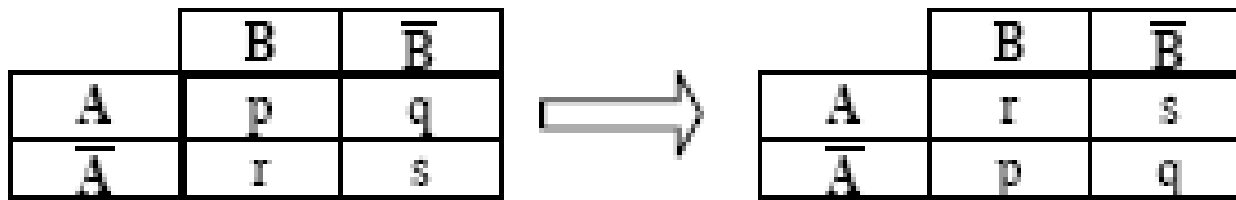
Properties of a Measure

- **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 $O(MC) = O(M)$.



Properties of a Measure

- **Antisymmetry Under Row/Column Permutation (O3):** For a 2×2 matrix $S = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, 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 $O(MS) = -O(M)$.



Properties of a Measure

- Inversion Invariance (O3')**: For a 2×2 matrix $S = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, a measure O is said to be invariant under inversion operation if $O(SMS) = O(M)$.

	B	\bar{B}
A	p	q
\bar{A}	r	s

→

	B	\bar{B}
A	s	r
\bar{A}	q	p

- Null Invariance (O4)**: For a matrix $C = \begin{bmatrix} 0 & 0 \\ 0 & k \end{bmatrix}$, a measure O is said to be null invariant if $O(M+C) = O(M)$.

	B	\bar{B}
A	p	q
\bar{A}	r	s

→

	B	\bar{B}
A	p	q
\bar{A}	r	s+k

Properties of a Measure

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

Table 9: Properties of a measure

Properties Sensitive to Rare Association Rules

- With respect to mining rare association rules, the following two questions are unclear in the work of Tan *et. al.*
 - Can any measure be used for mining rare association rules?
 - What properties a user should consider for mining rare association rules?

Properties Sensitive to Rare Association Rules

- Property **P1** is not mandatory.
 - A measure can take any value (not zero necessarily) to indicate that A, B are independent.
- Property **P2** is interesting
 - Association between rare variables increases with increase in $P(A,B)$.
- Property **P3** is interesting
 - Consider two rare variables A, B . If $P(A)$ is increased keeping $P(A,B)$ constant, A no longer remains rare.
 - An association between a rare and a frequent variable is not as interesting as that between two rare variables.

Properties Sensitive to Rare Association Rules

- Properties **O1**, **O2**, **O3** and **O3'** are subjective to user interest.
- *Null Invariance Property (O4)* is interesting
 - A measure satisfying *null invariance property* is not influenced by the co-absence of the participating variables.
 - A transaction containing neither *A* nor *B* is a *null transaction* with respect to the rule $(A \rightarrow B)$.
 - Number of *null transactions* are huge for rare association rules.
 - To prevent pruning of rare association rules, *null transactions* should not be considered.

Properties Sensitive to Rare Association Rules

- It is preferable to select a measure that satisfies properties **P2**, **P3** and **O4**.
- Among symmetric measures, *cosine (IS)*, *all-confidence (h)* and *jaccard (ζ)* satisfy all the three suggested properties.
- Among asymmetric measures, *mutual information (M)*, *certainty factor (F)* and *added value (AV)* satisfy two (P2 and P3) out of the three suggested properties.

Outline

- Introduction
- Related Work
- Interestingness Measures for Mining Rare Association Rules
- **Experimental Analysis**
- Conclusions and Future Work

Experimental Analysis

- Real-world datasets used :-
 - **Retail dataset** : Sparse dataset with 88,162 transactions.
 - BMS-WebView-1 dataset : Sparse dataset with 59,602 transactions.

- The following two experiments are conducted
 - Experiment 1 : Similarity between the measures while mining rare association rules.
 - Experiment 2 : Selecting an appropriate measure for mining rare association rules.

Experiment 1: Similarity between various measures

- For a dataset, a set of all contingency tables is derived from the set of frequent patterns.
- For each measure, the corresponding ranking vectors were computed.
- Similarity between measures was computed by finding *Pearson's correlation* between the corresponding ranking vectors.

Similarity between different *Symmetric Measures* for Retail dataset

Table 10: 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	<u>0.951</u>	0.532	-0.594	1		
IR	0.768	0.671	0.614	0.743	<u>0.779</u>	0.389	-0.557	0.882	1	
ζ	0.889	0.827	0.671	0.845	<u>0.958</u>	0.525	-0.572	<u>0.996</u>	0.86	1

Jaccard (ζ), *all-confidence* (h) and *cosine* (IS) are highly similar to each other.

Similarity between different *Asymmetric Measures* for *Retail dataset*

Table 11: Similarity between different *asymmetric measures* for *Retail dataset*

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

Certainty factor (F), added value (AV) and mutual information (M) are highly similar to each other.

Experiment 2: Selecting a Measure

- Choose a random set of n contingency tables such that the following combinations for A and B are present.
 - Both A and B are frequent.
 - Both A and B are rare.
 - A is frequent and B is rare.
 - A is rare and B is frequent.

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

Table 12: Sample set of contingency tables taken from Retail dataset

Experiment 2: Selecting a Measure

- User ranks the n contingency tables based on the perceived interestingness. Call this the ranking vector U_r
- Using each measure M , rank the sample set in decreasing order of magnitude. Call this ranking vector R_M
- Find similarity between the ranking vectors U_r and R_M . Measure with highest similarity value is selected.
 - *Pearson's correlation* can be used to find similarity between ranking vectors.

	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 Retail dataset

Experiment 2: Selecting a measure

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

- *All-confidence* (h), *Jaccard* (ζ) and *cosine* (IS) gave highly similar rankings to user-given rankings.

Table 14: 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

- *Mutual information* (M) gave most similar rankings.
- *Added value* (AV) and *certainty factor* (F) also give ranks similar to the ranks expected by the users.

Outline

- Introduction
- Related Work
- Interestingness Measures for Mining Rare Association Rules
- Experimental Analysis
- **Conclusions and Future Work**

Conclusions and Future Work

- We analyzed how various interestingness measures perform in extracting *rare association rules*.
- Suggested a set of properties, one should consider when selecting a measure for mining *rare association rules*.
- Performed experimental analysis to support the fact that measures satisfying the prescribed properties are able to mine rare association rules.
- In future, we would like to investigate approaches to divide the set of frequent patterns into different groups, and applying a different interestingness measure on each group.