SELECT THE RIGHT INTERESTINGNESS MEASURE FOR ASSOCIATION PATTERNS

Pang-Ning Tan Vipin Kumar
Jaideep Srivastava

presentation : Zhipeng Cai

ABSTRACT

- Many techniques for association rule mining and feature selection require a suitable metric to capture the dependencies among variables in a data set.
- However, many such measures provide conflicting information about the interestingness of a pattern and best metric to use for a given application domain is rarely known.

Specific contributions

- 1: Present an overview of various measures proposed in the statistics,machine learning and data mining literature.
- 2: Describe several key properties one should examine in order to select the right measure for a given application domain.A comparative study of these properties is made using twenty one of the existing measures.

Specific contributions

- 3:we present two scenario in which most of the existing measures agree with each other.

  namely, support-based pruning and table standardization

4: present an algorithm to select a small set of tables such that an expert can select a desirable measure by looking at just a small set of table.

INTRODUCTION

- The central task of association rule mining is to find sets of binary variables that co-occur together frequently in a transaction database.
- Analysis often requires a suitable metric to capture the dependencies among variables.
- These metrics are defined in terms of the frequency counts tabulated in a 2*2 contingency table.

*Table1:A 2*2 contingency table for variables A and B*

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

Table 2:Example of contingency tables

Example	f_{11}	f_{10}	f_{01}	f_{00}
E1	8123	83	424	1370
E2	8330	2	622	1046
E3	9481	94	127	298
E4	3954	3080	5	2961
E5	2886	1363	1320	4431
E6	1500	2000	500	6000
E7	4000	2000	1000	3000
E8	4000	2000	2000	2000
E9	1720	7121	5	1154
E10	61	2483	4	7452

Table 3:Ranking of contingency table using various interestingness measures

Table 3: Rankings of contingency tables using various interestingness measures.

Example	ϕ	λ	α	Q	Y	κ	M	J	G	s	c	L	V	I	IS	PS	F	AV	S	ζ	K
E1	1	1	3	3	3	1	2	2	1	3	5	5	4	6	2	2	4	6	1	2	5
E2	2	2	1	1	1	2	1	3	2	2	1	1	1	8	3	5	1	8	2	3	6
E3	3	3	4	4	4	3	3	8	7	1	4	4	6	10	1	8	6	10	3	1	10
E4	4	7	2	2	2	5	4	1	3	6	2	2	2	4	4	1	2	3	4	5	1
E5	5	4	8	8	8	4	7	5	4	7	9	9	9	3	6	3	9	4	5	6	3
E6	6	6	7	7	7	7	6	4	6	9	8	8	7	2	8	6	7	2	7	8	2
E7	7	5	9	9	9	6	8	6	5	4	7	7	8	5	5	4	8	5	6	4	4
E8	8	9	10	10	10	8	10	10	8	4	10	10	10	9	7	7	10	9	8	7	9
E9	9	9	5	5	5	9	9	7	9	8	3	3	3	7	9	9	3	7	9	9	8
E10	10	8	6	6	6	10	5	9	10	10	6	6	5	1	10	10	5	1	10	10	7

Interestingness Measures for Association Patterns

#	Measure	Formula
1	ϕ-coefficient	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
2	Goodman-Kruskal's (λ)	$\frac{\sum_j \max_k P(A_j,B_k)+\sum_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-\max_k P(B_k)}$
3	Odds ratio (α)	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(A,\overline{B})P(\overline{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(\overline{A}\overline{B})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{A}\overline{B})+P(A,\overline{B})P(\overline{A},B)} = \frac{\alpha-1}{\alpha+1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\overline{A}\overline{B})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{A}\overline{B})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$
6	Kappa (κ)	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
7	Mutual Information (M)	$\frac{\sum_i \sum_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i)\log P(A_i),-\sum_j P(B_j)\log P(B_j))}$
8	J-Measure (J)	$\max\Big(P(A,B)\log(\frac{P(B\|A)}{P(B)})+P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),$ $P(A,B)\log(\frac{P(A\|B)}{P(A)})+P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})\Big)$
9	Gini index (G)	$\max\Big(P(A)[P(B\|A)^2+P(\overline{B}\|A)^2]+P(\overline{A})[P(B\|\overline{A})^2+P(\overline{B}\|\overline{A})^2]$ $-P(B)^2-P(\overline{B})^2,$ $P(B)[P(A\|B)^2+P(\overline{A}\|B)^2]+P(\overline{B})[P(A\|\overline{B})^2+P(\overline{A}\|\overline{B})^2]$ $-P(A)^2-P(\overline{A})^2\Big)$

Interestingness Measures for Association Patterns

#	Measure	Formula
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B\|A),P(A\|B))$
12	Laplace (L)	$\max\left(\frac{NP(A,B)+1}{NP(A)+2},\frac{NP(A,B)+1}{NP(B)+2}\right)$
13	Conviction (V)	$\max\left(\frac{P(A)P(\overline{B})}{P(A\overline{B})},\frac{P(B)P(\overline{A})}{P(B\overline{A})}\right)$
14	Interest (I)	$\frac{P(A,B)}{P(A)P(B)}$
15	cosine (IS)	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
16	Piatetsky-Shapiro's (PS)	$P(A,B)-P(A)P(B)$
17	Certainty factor (F)	$\max\left(\frac{P(B\|A)-P(B)}{1-P(B)},\frac{P(A\|B)-P(A)}{1-P(A)}\right)$
18	Added Value (AV)	$\max(P(B\|A)-P(B),P(A\|B)-P(A))$
19	Collective strength (S)	$\frac{P(A,B)+P(\overline{A}\overline{B})}{P(A)P(B)+P(\overline{A})P(\overline{B})}\times\frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{A}\overline{B})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
21	Klosgen (K)	$\sqrt{P(A,B)}\max(P(B\|A)-P(B),P(A\|B)-P(A))$

Two situation

- 1: the measures may become highly correlated when support-based pruning is used.
- 2: after standardizing the contingency tables to have uniform margins, many of the well-known measures become equivalent each other.

Preliminaries

- T(D)={t1,t2,t3….t n} denote the set of patterns .
- P is the set of measures available to an analyst.
- $M \in P$
- M(T)={m1,m2,m3….m n},which corresponds to the values of M for each contingency table that belongs to T(D).
- M(T) can also be transformed into a ranking vector Om(T)={O1,O2,….On}.

Definition 1:

- [Similarity between measures]
- Two measures of association, M1 and M2, are similar to each other with respect to the data set D if the correlation between Om1(T) and Om2(T) is greater than or equal to some positive threshold t.

Desired properties of a measure

three key properties

- 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.

Other properties of a measure

- Property 1: [symmetry under variable permutation]
- A measure O is symmetric under variable permutation, A $\longleftrightarrow$ B,if $O(M^T)=O(M)$ for all contingency matrices M

	B	$\overline{B}$
A	p	q
$\overline{A}$	r	s

$\Longrightarrow$

	A	$\overline{A}$
B	p	r
$\overline{B}$	q	s

(a) Variable Permutation Operation

- Property 2:[Row/Column scaling invariance]
- Let R=C=[k1 0 ;0 k2] be a 2*2 square matrix.
- A measure O is invariant under row and column scaling if O(RM)=O(M) and O(MC)=O(M) for all contingency matrices,M

	B	$\overline{B}$
A	p	q
$\overline{A}$	r	s

$\Longrightarrow$

	B	$\overline{B}$
A	$k_3 k_1 p$	$k_4 k_1 q$
$\overline{A}$	$k_3 k_2 r$	$k_4 k_2 s$

(b) Row & Column Scaling Operation

Property 3: Antisymmetry under Row/Column permutation.

- Let S=[0 1; 1 0] be a 2*2 permutation matrix. A normalized measure O is antisymmetric under the row permutation operation.
- O(SM)= - O (M).
- Under the column permutation operation
- O(MS)=-O(M)

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

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

(c) Row & Column Permutation Operation

Property 4: Inversion Invariance

- Let S=[0 1;1 0] be a 2*2 permutation matrix . A measure O is invariant under the inversion operation , if O(SMS)=O(M) for all contingency matrices M.

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

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

(d) Inversion Operation

- Property 5: Null Invariance
- A binary measure of association is null-invariant if O(M+C)=O(M) where C=[0 0; 0 k] and is a positive constant.

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

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

(e) Null Addition Operation

Table 6 properties of interestingness measures

Symbol	Measure	Range	P1	P2	P3	O1	O2	O3	O3'	O4
ϕ	ϕ-coefficient	$-1 \cdots 0 \cdots 1$	Yes	Yes	Yes	Yes	No	Yes	Yes	No
λ	Goodman-Kruskal's	$0 \cdots 1$	Yes	No	No	Yes	No	No*	Yes	No
α	odds ratio	$0 \cdots 1 \cdots \infty$	Yes*	Yes	Yes	Yes	Yes	Yes*	Yes	No
Q	Yule's Q	$-1 \cdots 0 \cdots 1$	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Y	Yule's Y	$-1 \cdots 0 \cdots 1$	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
κ	Cohen's	$-1 \cdots 0 \cdots 1$	Yes	Yes	Yes	Yes	No	No	Yes	No
M	Mutual Information	$0 \cdots 1$	Yes	Yes	Yes	No**	No	No*	Yes	No
J	J-Measure	$0 \cdots 1$	Yes	No	No	No**	No	No	No	No
G	Gini index	$0 \cdots 1$	Yes	No	No	No**	No	No*	Yes	No
s	Support	$0 \cdots 1$	No	Yes	No	Yes	No	No	No	No
c	Confidence	$0 \cdots 1$	No	Yes	No	No**	No	No	No	Yes
L	Laplace	$0 \cdots 1$	No	Yes	No	No**	No	No	No	No
V	Conviction	$0.5 \cdots 1 \cdots \infty$	No	Yes	No	No**	No	No	Yes	No
I	Interest	$0 \cdots 1 \cdots \infty$	Yes*	Yes	Yes	Yes	No	No	No	No
IS	Cosine	$0 \cdots \sqrt{P(A,B)} \cdots 1$	No	Yes	Yes	Yes	No	No	No	Yes
PS	Piatetsky-Shapiro's	$-0.25 \cdots 0 \cdots 0.25$	Yes	Yes	Yes	Yes	No	Yes	Yes	No
F	Certainty factor	$-1 \cdots 0 \cdots 1$	Yes	Yes	Yes	No**	No	No	Yes	No
AV	Added value	$-0.5 \cdots 0 \cdots 1$	Yes	Yes	Yes	No**	No	No	No	No
S	Collective strength	$0 \cdots 1 \cdots \infty$	No	Yes	Yes	Yes	No	Yes*	Yes	No
ζ	Jaccard	$0 \cdots 1$	No	Yes	Yes	Yes	No	No	No	Yes
K	Klosgen's	$(\frac{2}{\sqrt{3}}-1)^{1/2}[2-\sqrt{3}-\frac{1}{\sqrt{3}}] \cdots 0 \cdots \frac{2}{3\sqrt{3}}$	Yes	Yes	Yes	No**	No	No	No	No

Table 6 properties of interestingness measures

- where: P1: O(M) = 0 if det(M) = 0, i.e. , whenever A and B are statistically independent.
- P2: O(M2) > O(M1) if M2 = M1+ [k –k;-k k]
- P3: O(M2) < O(M1) if M2=M1+ [0 k;0 -k] or M2=M1+ [0 0;k -k] .
- O1: Property1:symmetry under variable permutation
- O2: Property2: Row/Column scaling invariance
- O3: Property3:Antisymmetry under Row/Column permutation.
- O3’:Property4: inversion invariance.
- O4:: Property5: Null invariance
- Yes*: yes if measure is normalized.
- No*:Symmetry under row or column permulation.
- No**:No unless the measure is symmetrized by taking max(M(A,B),M(B,A)).

Summary

- The discussion in this section suggests that there is no measure that is better than others in all application domains .
- Thus, in order to find the right measure, one must match the desired properties of an application against the properties of the existing measures.

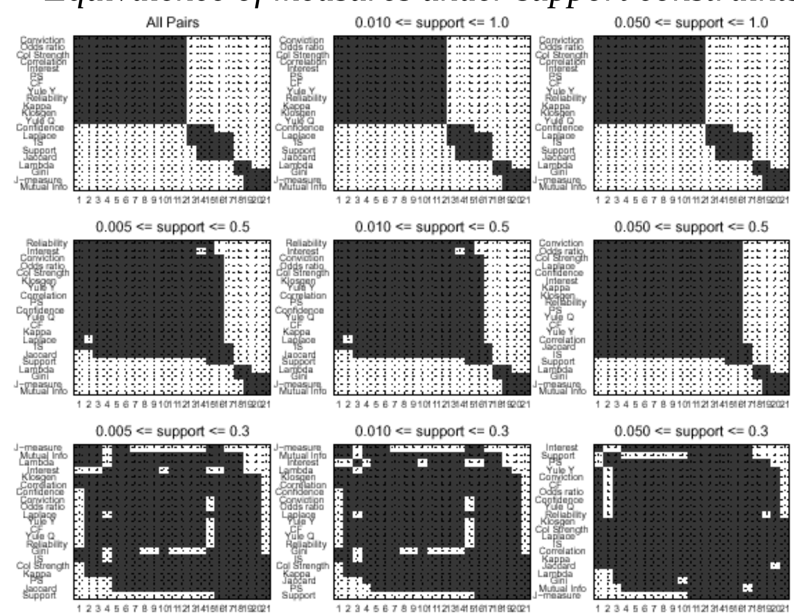
Effect of support-based pruning

- Support is a widely-used measure in association rule mining because it represents the statistical significance of a pattern.
- We now describe two additional consequences of using the support measure.

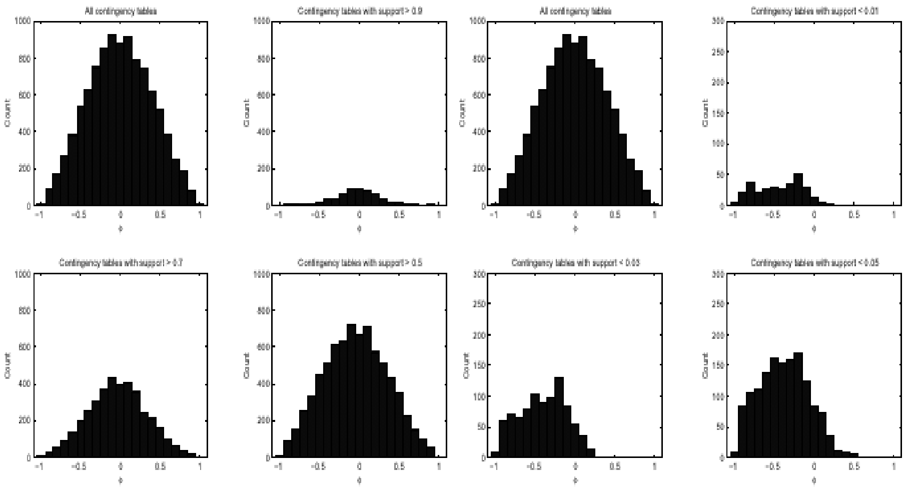
1: Equivalence of measures under support constraints.

2: Elimination of poorly correlated tables using support-based pruning.

Equivalence of measures under support constraints



Elimination of poorly correlated tables using support-based pruning.



(a) Distribution of ϕ-coefficient for contingency tables that are removed by applying a maximum support threshold.

(b) Distribution of ϕ-coefficient for contingency tables that are removed by applying a minimum support threshold.

TABLE STANDARDIZATION

- Standardization is a widely-used technique.
- standardization is needed to get a better idea of the underlying association between marginals are variables by transforming an existing table so that their equal.

$f_{1+}^* = f_{0+}^* = f_{+1}^* = f_{+0}^* = N/2$

$$f_{1+}^* = f_{0+}^* = f_{+1}^* = f_{+0}^* = N / 2$$

Table 7: Table Standardization

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

$\longrightarrow$

	B	$\overline{B}$	
A	f^*_{11}	f^*_{10}	f^*_{1+}
$\overline{A}$	f^*_{01}	f^*_{00}	f^*_{0+}
	f^*_{+1}	f^*_{+0}	N

(a)

$\longrightarrow$

	B	$\overline{B}$	
A	x	$N/2-x$	$N/2$
$\overline{A}$	$N/2-x$	x	$N/2$
	$N/2$	$N/2$	N

(b)

- Row scaling: $f_{ij}^{(k)} = f_{ij}^{(k-1)} \times \frac{f_{i+}^*}{f_{+j}^{(k)}}$
- Column scaling: $f_{ij}^{(k+1)} = f_{ij}^{(k)} \times \frac{f_{+j}^*}{f_{+i}^{(k)}}$

k=0

15	10	25
35	40	75
50	50	100

Original Table

k=1

30	20	50
23	27	50
53	47	100

k=2

28	21	50
22	29	50
50	50	100

k=3

28	22	50
22	28	50
50	50	100

k=4

28	22	50
22	28	50
50	50	100

k=5

28	22	50
22	28	50
50	50	100

Table 8:Rankings of contingency table after IPF standardization

Table 8: Rankings of contingency tables after IPF standardization.

Example	ϕ	λ	α	Q	Y	κ	M	J	G	s	c	L	V	I	IS	PS	F	AV	S	ζ	K
E1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
E2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
E3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
E4	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
E5	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
E6	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
E7	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
E8	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
E9	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
E10	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6

Three equation for fix the standardized table

- 1 $f_{11}^{*} = f_{00}^{*}$
- 2 $f_{10}^{*} = f_{01}^{*}$
- 3 $f_{11}^{*} + f_{10}^{*} = N / 2$

Example

- Odds ratio : $\dfrac{P(A, B)\, P(\overline{A}, \overline{B})}{P(A, \overline{B})\, P(\overline{A}, B)}$

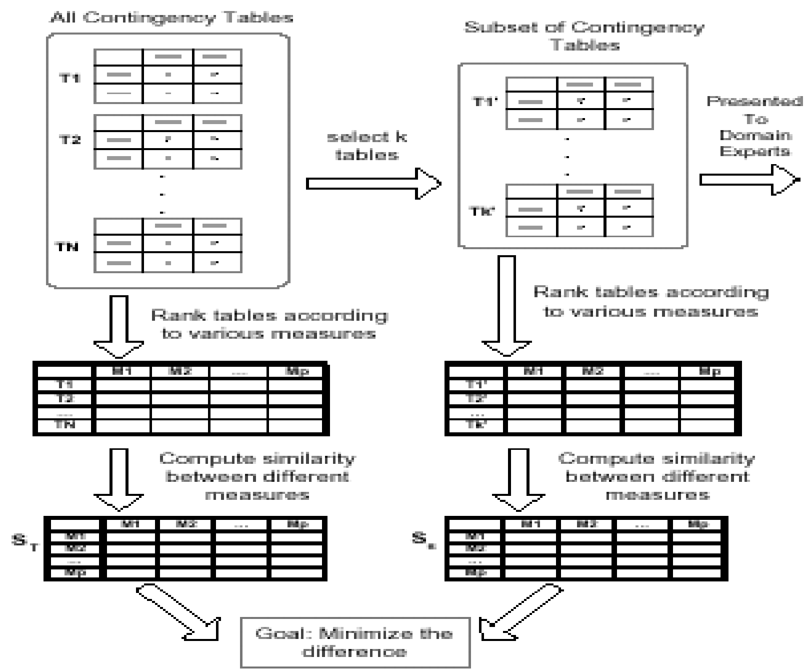
Fourth equations: $\dfrac{f_{11}\, f_{00}}{f_{10}\, f_{01}} = \dfrac{f_{11}^{*}\, f_{00}^{*}}{f_{10}^{*}\, f_{01}^{*}}$

$$f_{11}^{*} = f_{00}^{*} = \frac{N\sqrt{f_{11}\, f_{00}}}{2(\sqrt{f_{11}\, f_{00}} + \sqrt{f_{10}\, f_{01}})}$$

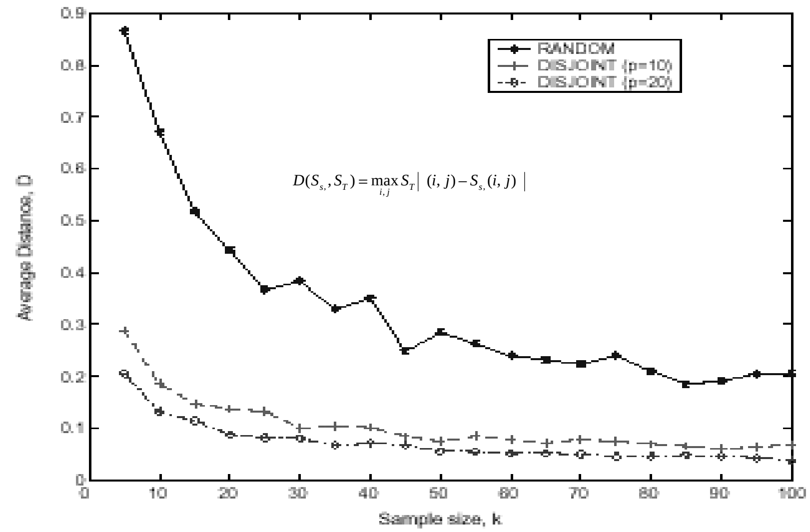
$$f_{10}^{*} = f_{01}^{*} = \frac{N\sqrt{f_{10}\, f_{01}}}{2(\sqrt{f_{11}\, f_{00}} + \sqrt{f_{10}\, f_{01}})}$$

Measure Selection Based on bankings by experts

- 1:Random :randomly select k out of the overall N tables and present them to the experts.
- 2:Disjoint: select k tables that are “furthest” Apart according to their average ranking and would produce the largest amount of ranking conflicts.



$$D(S_{s,}, S_T) = \max_{i,j} \left| \; S_T(i,j) - S_{s,}(i,j) \; \right|$$



re0	Q	Y	κ	PS	F	AV	K	I	c	L	IS	ξ	s	S	λ	M	J	G	α	V
All tables	8	7	4	16	15	10	11	9	17	18	2	12	19	3	20	5	1	13	6	14
k=20	6	6	5	16	13	10	11	12	17	18	2	15	19	4	20	3	1	9	6	14

la1	Q	Y	κ	PS	F	AV	K	I	c	L	IS	ξ	s	S	λ	M	J	G	α	V
All tables	10	9	2	7	5	3	6	16	18	17	13	14	19	1	20	12	11	15	8	4
k=20	13	13	2	5	8	3	6	16	18	17	10	11	19	1	20	9	4	12	13	7

Product	Q	Y	κ	PS	F	AV	K	I	c	L	IS	ξ	s	S	λ	M	J	G	α	V
All tables	12	11	3	10	8	7	14	16	17	18	1	4	19	2	20	5	6	15	13	9
k=20	13	13	2	7	11	10	9	17	16	18	1	4	19	3	20	6	5	8	13	11

S&P500	Q	Y	κ	PS	F	AV	K	I	c	L	IS	ξ	s	S	λ	M	J	G	α	V
All tables	9	8	1	10	6	3	4	11	15	14	12	13	19	2	20	16	18	17	7	5
k=20	7	7	2	10	4	3	6	11	17	18	12	13	19	1	20	15	14	16	7	4

E-Com	Q	Y	κ	PS	F	AV	K	I	c	L	IS	ξ	s	S	λ	M	J	G	α	V
All tables	9	8	3	7	14	13	16	11	17	18	1	4	19	2	20	6	5	12	10	15
k=20	7	7	3	10	15	14	13	11	17	18	1	4	19	2	20	6	5	12	7	15

Census	Q	Y	κ	PS	F	AV	K	I	c	L	IS	ξ	s	S	λ	M	J	G	α	V
All tables	10	10	2	3	7	5	4	11	13	12	14	15	16	1	20	19	18	17	10	6
k=20	6	6	3	2	9	5	4	11	13	12	14	15	16	1	17	18	19	20	6	9

All tables: Rankings when all contingency tables are ordered.

k=20 : Rankings when 20 of the selected tables are ordered.

Conclusions

- 1:Describe several key properties.
- 2:There are situations in which many of these measure that is consistently with each other
- 3:Present an algorithm to select a small set of tables that an expert can find the most appropriate measure by looking at this small set of table.