Data Mining and Analysis

Market Basket Analysis: 3

CSCE 676 :: Fall 2019
Texas A&M University
Department of Computer Science & Engineering
Prof. James Caverlee

Resources

MMDS: Mining of Massive Datasets [http://www.mmds.org/mmds/v2.1/ch06-assocrules.pdf]

Tan, Steinbach, Karpatne, Kumar. Introduction to Data Mining [https://www-users.cs.umn.edu/~kumar001/dmbook/slides/chap5-association_analysis.pdf]

Carlos Castillo course on Data Mining [https://github.com/chatox/data-mining-course]

Vagelis Papalexakis course on Data Mining [https://www.cs.ucr.edu/~epapalex/teaching/235_S19/index.html]

Speeding Up Candidate Generation

Speeding Up Candidate Generation

A Naïve Approach

Check all the possible combinations of frequent itemsets

An Example of the Naïve Approach

itemsets: {abc} {bcd} {abd} {cde}

{abc} + {bcd} = {abcd}

{bcd} + {abd} = {abcd}

{abd} + {cde} = {abcde}

….

Introduction of Ordering

- Items in U have a lexicographic ordering
- Itemsets can be ordered as strings

A Better Approach

- Order the frequent k-itemsets
- Merge two itemset if the first k-1 items of them are the same

Example

- k-itemsets: {abc} {abd} {bcd}
  - {abc} + {abd} = {abcd}
- k-itemsets: {abc} {acd} {bcd}
  - No (k+1) -candidates

Early stop is possible

- Do not need to check {abc} +{bcd} after checking {abc} + {acd}

Do we miss {abcd}?

- No, due to the Downward Closure Property

Level-wise pruning trick

Let F_k be the set of frequent k-itemsets

Let C_{k+1} be the set of (k+1)-candidates

$I \in C_{k+1}$ is frequent only if all the k-subsets of I are frequent

Pruning

- Generate all the k-subsets of I
- If any one of them does not belong to F_k, then remove I

Frequent Itemsets in < 2 passes?

Frequent Itemsets in < 2 passes?

Apriori takes k passes to find frequent itemsets of size k

Can we use fewer passes?

Use 2 or fewer passes for all sizes, but may miss some frequent itemsets

- Random sampling
- SON (Savasere, Omiecinski, and Navathe)
- Toivonen

Random Sampling

Take a random sample of the market baskets

Run a-priori or one of its improvements in main memory

- So we don't pay for disk I/O each time we increase the size of itemsets

- Reduce support threshold proportionally to match the sample size

Random Sampling

To avoid false positives: Optionally, verify that the candidate pairs are truly frequent in the entire data set by a second pass (avoid false positives)

But you don’t catch sets frequent in the whole but not in the sample

Smaller threshold, e.g., s/125, helps catch more truly frequent itemsets

But requires more space

SON Algorithm

Repeatedly read small subsets of the baskets into main memory and run an in-memory algorithm to find all frequent itemsets

> Note: we are not sampling, but processing the entire file in memory-sized chunks

An itemset becomes a candidate if it is found to be frequent in any one or more subsets of the baskets.

SON Algorithm

On a second pass, count all the candidate itemsets and determine which are frequent in the entire set

Key “monotonicity” idea: an itemset cannot be frequent in the entire set of baskets unless it is frequent in at least one subset.

SON: Distributed Version

SON lends itself to distributed data mining

Baskets distributed among many nodes

- Compute frequent itemsets at each node
- Distribute candidates to all nodes
- Accumulate the counts of all candidates

Toivonen's Algorithm

Pass 1: Start with a random sample, but lower the threshold slightly for the sample

> Example: If the sample is 1% of the baskets use 1.25% as the minsup threshold instead of 1%

Find frequent itemsets in the sample

Add to the itemsets that are frequent in the sample the negative border of these itemsets

Negative Border

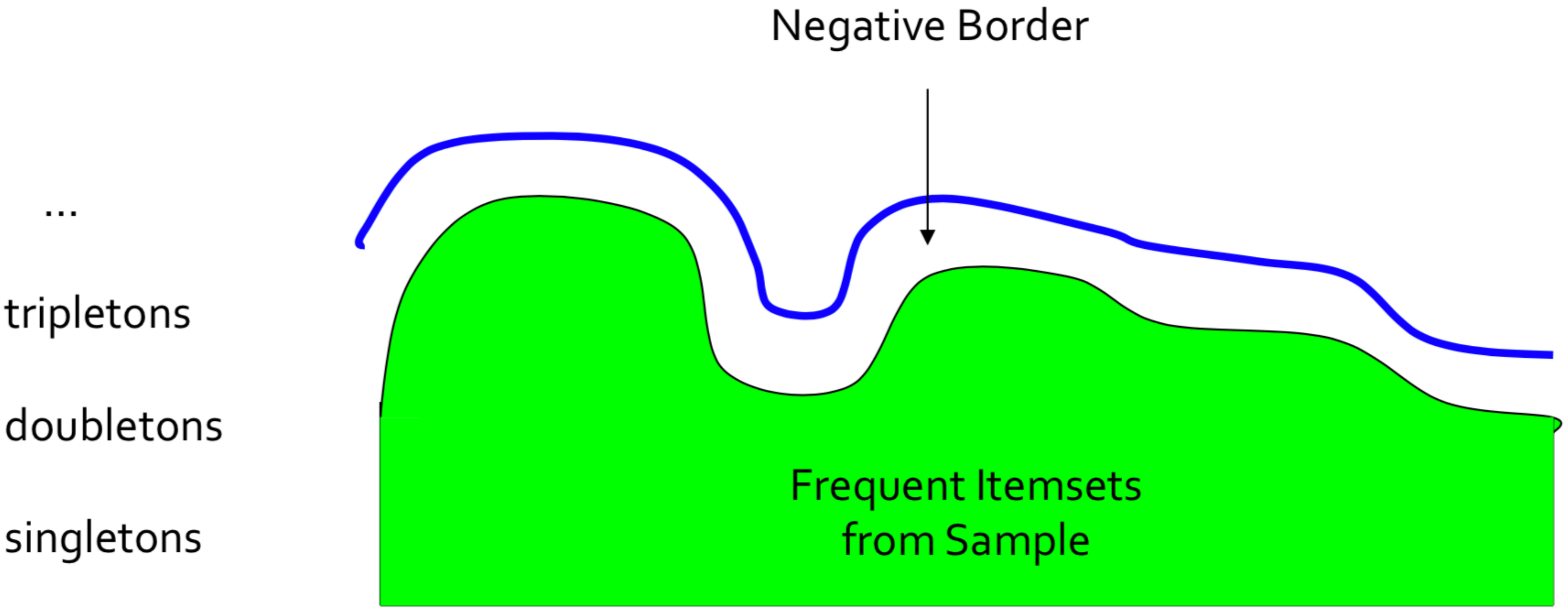
Negative border = an itemset is in the negative border if it is not frequent in the sample but all its immediate subsets are

Immediate subset = “delete exactly one element”

Negative border: Example

{A,B,C,D} is in the negative border if and only if:

1. It is not frequent in the sample, but
2. All of {A,B,C}, {B,C,D}, {A,C,D}, and {A,B,D} are.



Toivonen’s Algorithm

Pass 1:

Start as in the SON algorithm, but lower the threshold slightly for the subset

Add to the itemsets that are frequent in the sample the negative border of these items sets

Pass 2:

Count all candidate frequent itemsets from the first pass, and also count sets in their negative border

Toivonen’s Algorithm

If no itemset from the negative border turns out to be frequent, then we found all the frequent itemsets.

本來是用這個線來區分，但如果出現頻繁，則線除了問題

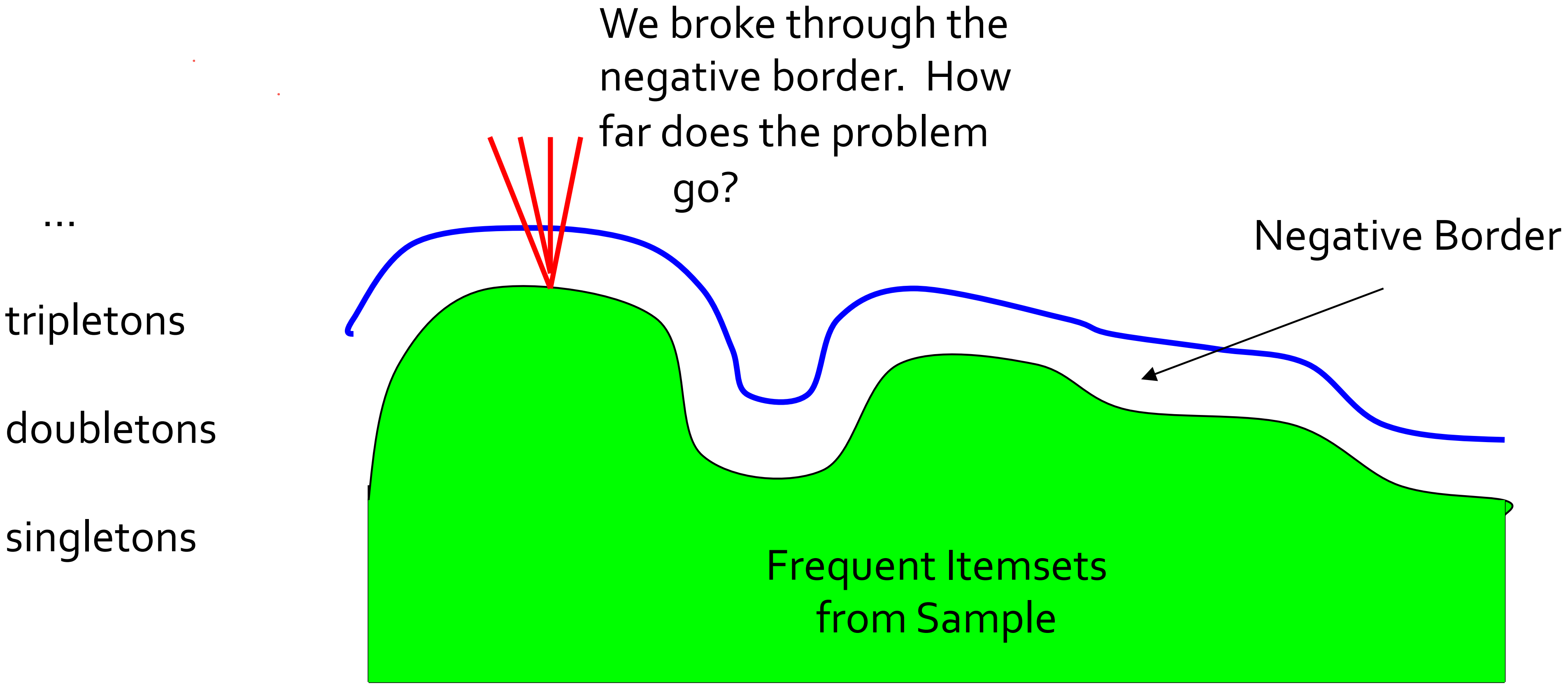
What if we find that something in the negative border is frequent?

We must start over again with another sample!

Try to choose the support threshold so the probability of failure is low, while the number of itemsets checked on the second pass fits in main memory

If something in the negative border is frequent …

參考鏈接: https://blog.csdn.net/pipisorry/article/details/48914067



We broke through the negative border. How far does the problem go?

…

tripletons

doubletons

singletons

Negative Border

Frequent Itemsets from Sample

如果一个项集在整个数据集上是频繁的，而在样本中不是频繁的，那么negative border中一定有一个成员是频繁的。（这样就要重新运行算法了）

Theorem

也就是说如果在negative border中没有成员在整个数据集中是频繁的，那么在整个数据集中再也不存在我们没有计数的频繁项了。

If there is an itemset S that is frequent in full data, but not frequent in the sample, then the negative border contains at least one itemset that is frequent in the whole.

Proof by contradiction:

Suppose not; i.e.;

1. There is an itemset S frequent in the full data but not frequent in the sample, and
2. Nothing in the negative border is frequent in the full data

Let T be a smallest subset of S that is not frequent in the sample (but every subset of T is)

T is frequent in the whole (S is frequent + monotonicity).

But then T is in the negative border (contradiction)

Evaluating Patterns

Interesting Association Rules

Not all high-confidence rules are interesting

The rule X → milk may have high confidence for many item sets X, because milk is just purchased very often (independent of X) and the confidence will be high

One idea: Lift

Lift

The lift of the rule X ⇒ Y is:

$$\text{lift}(X \Rightarrow Y) = \frac{\sup(X \cup Y)}{\sup(X)\sup(Y)}$$

This is the ratio between the observed support and the expected support if X and Y were independent

Lift

解釋的不錯: https://youtu.be/WxDV9WEYqPw?t=362

Lift(X,Y) = 1

X and Y are independent

Lift(X,Y) > 1

X and Y are positively correlated

Lift(X,Y) < 1

X and Y are negatively correlated

$$\text{lift}(X \Rightarrow Y) = \frac{\sup(X \cup Y)}{\sup(X)\sup(Y)}$$

Lift: Example

Five Baskets:

{A, B, C}, {A, C, D}, {B, C, D}, {A, D, E}, {B, C, E}

Association Rules:

A—>D

C—>A

A—>C

{B,C} —>D

Selecting the Right Interestingness Measure for Association Patterns

Pang-Ning Tan
Department of Computer Science and Engineering
University of Minnesota
200 Union Street SE
Minneapolis, MN 55455
ptan@cs.umn.edu

Vipin Kumar
Department of Computer Science and Engineering
University of Minnesota
200 Union Street SE
Minneapolis, MN 55455
kumar@cs.umn.edu

Jaideep Srivastava
Department of Computer Science and Engineering
University of Minnesota
200 Union Street SE
Minneapolis, MN 55455
srivasta@cs.umn.edu

ABSTRACT

Many techniques for association rule mining and feature selection require a suitable metric to capture the dependencies among variables in a data set. For example, metrics such as support, confidence, lift, correlation, and collective strength are often used to determine the interestingness of association patterns. However, many such measures provide conflicting information about the interestingness of a pattern, and the best metric to use for a given application domain is rarely known. In this paper, we present an overview of various measures proposed in the statistics, machine learning and data mining literature. We 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. We show that each measure has different properties which make them useful for some application domains,

For instance, the central task of association rule mining [2] is to find sets of binary variables that *co-occur* together frequently in a transaction database, while the goal of feature selection problems is to identify groups of variables that are strongly *correlated* with each other or with a specific target variable. Regardless of how the relationships are defined, such analysis often requires a suitable metric to capture the dependencies among variables. For example, metrics such as support, confidence, lift, correlation, and collective strength have been used extensively to evaluate the interestingness of association patterns [9, 14, 1, 15, 11]. These metrics are defined in terms of the frequency counts tabulated in a 2×2 contingency table as shown in Table 1. Unfortunately, many such metrics provide conflicting information about the interestingness of a pattern, and the best metric to use for a given application domain is rarely known.

KDD 2002

Table 5: 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)	$\max\Big(P(A,B)\log(\frac{P(B\mid A)}{P(B)}) + P(A\overline{B})\log(\frac{P(\overline{B}\mid A)}{P(\overline{B})}),$ $P(A,B)\log(\frac{P(A\mid B)}{P(A)}) + P(\overline{A}B)\log(\frac{P(\overline{A}\mid B)}{P(\overline{A})})\Big)$
9	Gini index (G)	$\max\Big(P(A)[P(B\mid A)^2 + P(\overline{B}\mid A)^2] + P(\overline{A})[P(B\mid\overline{A})^2 + P(\overline{B}\mid\overline{A})^2]$ $-P(B)^2 - P(\overline{B})^2,$ $P(B)[P(A\mid B)^2 + P(\overline{A}\mid B)^2] + P(\overline{B})[P(A\mid\overline{B})^2 + P(\overline{A}\mid\overline{B})^2]$ $-P(A)^2 - P(\overline{A})^2\Big)$
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B\mid A), P(A\mid B))$
12	Laplace (L)	$\max\Big(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2}\Big)$
13	Conviction (V)	$\max\Big(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})}\Big)$
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)	$\max\Big(\frac{P(B\mid A)-P(B)}{1-P(B)}, \frac{P(A\mid B)-P(A)}{1-P(A)}\Big)$
18	Added Value (AV)	$\max(P(B\mid A) - P(B), P(A\mid B) - P(A))$
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)	$\sqrt{P(A,B)}\max(P(B\mid A) - P(B), P(A\mid B) - P(A))$

How do we know if a rule is interesting?

Objective direct measures:

support, confidence, correlation, …

issues?

Subjective direct measures:

User-based — let users decide if a rule is unexpected, fresh, timely, etc.?

issues?

Objective indirect measure?

Put rule into practice (like A/B testing)

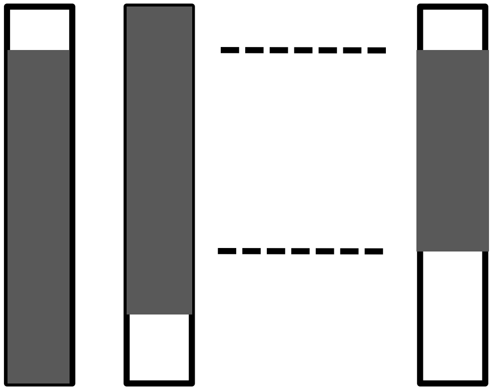
e.g., put beer next to diapers and measure sales

issues?

Association rules vs. correlation

If {coffee, milk} is a “large itemset” does this mean that there is a positive correlation between coffee and milk sales?

NO!!



‘coffee’ and ‘milk’
ANTI-correlated,
yet {coffee, milk}: frequent

Example from C. Faloutsos

	play-basketball	not play-basketball	sum (row)
eat-cereal	400	350	750
not eat-cereal	200	50	250
sum(col.)	600	400	1000

play basketball —> eat cereal [40%, 66.7%] is misleading

> The overall % of students eating cereal is 75% > 66.7%.

play basketball —> not eat cereal [20%, 33.3%] is more accurate,

> Although with lower support and confidence

	play-basketball	not play-basketball	sum (row)
eat-cereal	400	350	750
not eat-cereal	200	50	250
sum(col.)	600	400	1000

$$\boxed{\frac{0.4}{0.6 \times 0.75}}$$

lift(play basketball, eat cereal) = 0.89

lift(play basketball, not eat cereal) = 1.33

$$\boxed{\frac{0.2}{0.6 \times 0.25} =}$$

Null Invariance

Null transaction/itemset: an itemset that does not contain the ones we're interested in

- e.g., for the Coffee —> Milk, all the transactions that do not contain {Coffee, Milk}

Null invariance: measure value does not change with #null transactions

Lift and c2 are not null invariant!

- They can mislead us to think that an association rule is strong when it may be neutral

Null Invariance

Total # transactions at a store may fluctuate every day

We don’t want our interestingness measure to be sensitive to that

Null Invariance guarantees that – Offers stability

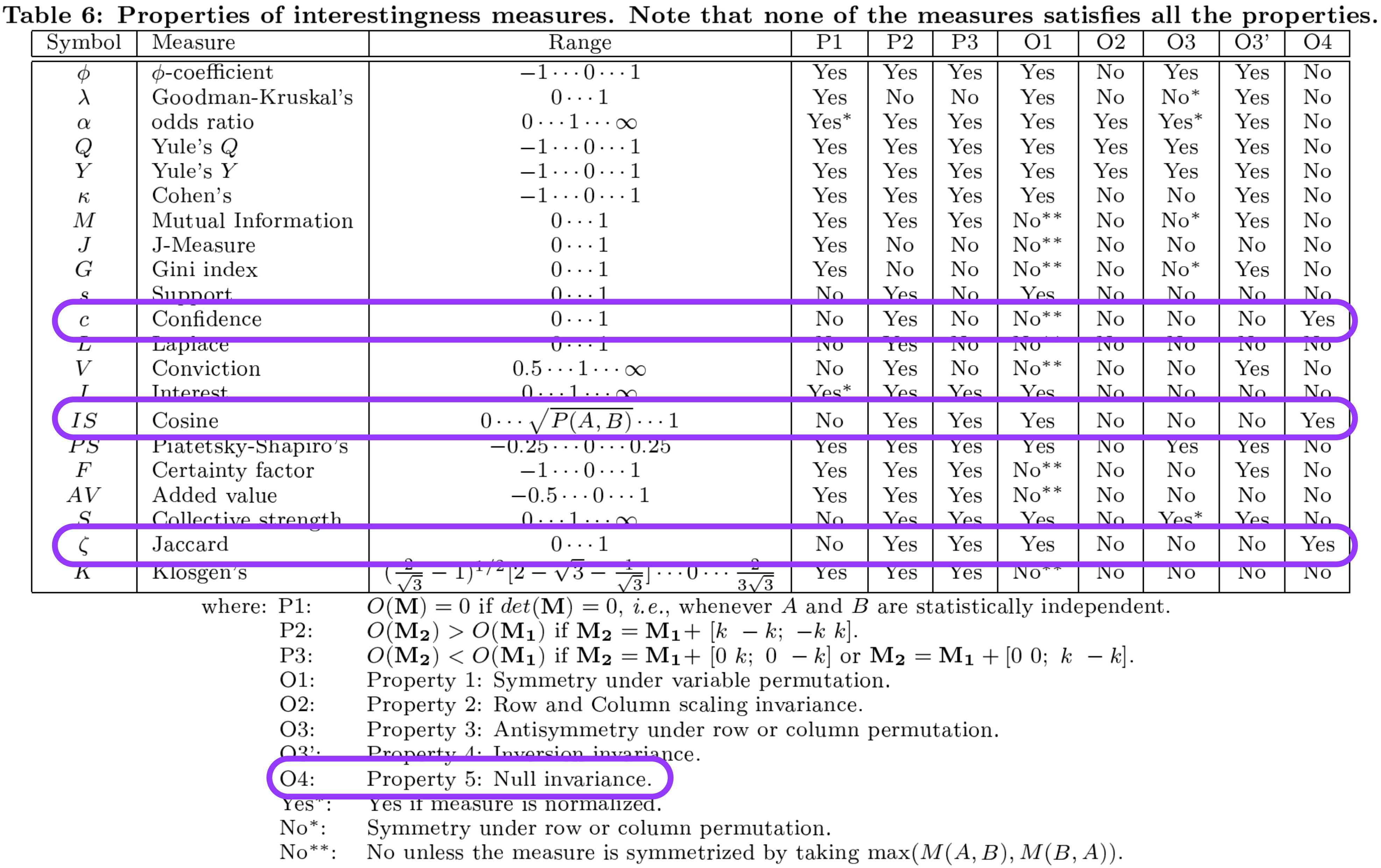


Table 6: Properties of interestingness measures. Note that none of the measures satisfies all the properties.

Symbol	Measure	Range	P1	P2	P3	O1	O2	O3	O3'	O4
ϕ	ϕ-coefficient	$-1\cdots0\cdots1$	Yes	Yes	Yes	Yes	No	Yes	Yes	No
λ	Goodman-Kruskal's	$0\cdots1$	Yes	No	No	Yes	No	No*	Yes	No
α	odds ratio	$0\cdots1\cdots\infty$	Yes*	Yes	Yes	Yes	Yes	Yes*	Yes	No
Q	Yule's Q	$-1\cdots0\cdots1$	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Y	Yule's Y	$-1\cdots0\cdots1$	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
κ	Cohen's	$-1\cdots0\cdots1$	Yes	Yes	Yes	Yes	No	No	Yes	No
M	Mutual Information	$0\cdots1$	Yes	Yes	Yes	No**	No	No*	Yes	No
J	J-Measure	$0\cdots1$	Yes	No	No	No**	No	No	No	No
G	Gini index	$0\cdots1$	Yes	No	No	No**	No	No*	Yes	No
s	Support	$0\cdots1$	No	Yes	No	Yes	No	No	No	No
c	Confidence	$0\cdots1$	No	Yes	No	No**	No	No	No	Yes
L	Laplace	$0\cdots1$	No	Yes	No	No**	No	No	No	No
V	Conviction	$0.5\cdots1\cdots\infty$	No	Yes	No	No**	No	No	Yes	No
I	Interest	$0\cdots1\cdots\infty$	Yes*	Yes	Yes	Yes	No	No	No	No
IS	Cosine	$0\cdots\sqrt{P(A,B)}\cdots1$	No	Yes	Yes	Yes	No	No	No	Yes
PS	Piatetsky-Shapiro's	$-0.25\cdots0\cdots0.25$	Yes	Yes	Yes	Yes	No	Yes	Yes	No
F	Certainty factor	$-1\cdots0\cdots1$	Yes	Yes	Yes	No**	No	No	Yes	No
AV	Added value	$-0.5\cdots0\cdots1$	Yes	Yes	Yes	No**	No	No	No	No
S	Collective strength	$0\cdots1\cdots\infty$	No	Yes	Yes	Yes	No	Yes*	Yes	No
ζ	Jaccard	$0\cdots1$	No	Yes	Yes	Yes	No	No	No	Yes
K	Klosgen's	$(\frac{2}{\sqrt{3}}-1)^{1/2}[2-\sqrt{3}-\frac{1}{\sqrt{3}}]\cdots0\cdots\frac{2}{3\sqrt{3}}$	Yes	Yes	Yes	No**	No	No	No	No

where: P1: $O(\mathbf{M})=0$ if $det(\mathbf{M})=0$, *i.e.*, whenever A and B are statistically independent.

P2: $O(\mathbf{M_2})>O(\mathbf{M_1})$ if $\mathbf{M_2}=\mathbf{M_1}+[k\ \ -k;\ -k\ \ k]$.

P3: $O(\mathbf{M_2})<O(\mathbf{M_1})$ if $\mathbf{M_2}=\mathbf{M_1}+[0\ k;\ 0\ \ -k]$ or $\mathbf{M_2}=\mathbf{M_1}+[0\ 0;\ k\ \ -k]$.

O1: Property 1: Symmetry under variable permutation.

O2: Property 2: Row and Column scaling invariance.

O3: Property 3: Antisymmetry under row or column permutation.

O3': Property 4: Inversion invariance.

O4: Property 5: Null invariance.

Yes*: Yes if measure is normalized.

No*: Symmetry under row or column permutation.

No**: No unless the measure is symmetrized by taking $\max(M(A,B),M(B,A))$.

Kulczynski measure

$\mathrm{Kulc}(A,B) = 0.5 * [\, P(A|B) + P(B|A) \,]$

Captures the balance between
A —> B and B —> A

In other words, it tells us whether both directions of the rule are similar

Or whether there is a direction that is far dominating

Kulczynski measure

- When Kulczynski is constant (0.5)
  - We call this rule “neutral”
  - Both directions equally important
  - Not enough to decide if the rule is interesting or not
- When it is far from 0.5, it shows skew of the rule
  - Potentially interesting

Imbalance ratio

IR (Imbalance Ratio): measure the imbalance of two itemsets A and B in an association rule

$$IR(A, B) = \frac{|sup(A) - sup(B)|}{sup(A) + sup(B) - sup(A \cup B)}$$

If IR is close to 0 then we have perfect balance

If IR close to 1 then we have heavy imbalance

We are typically interested in imbalanced rules, therefore we want IR to large

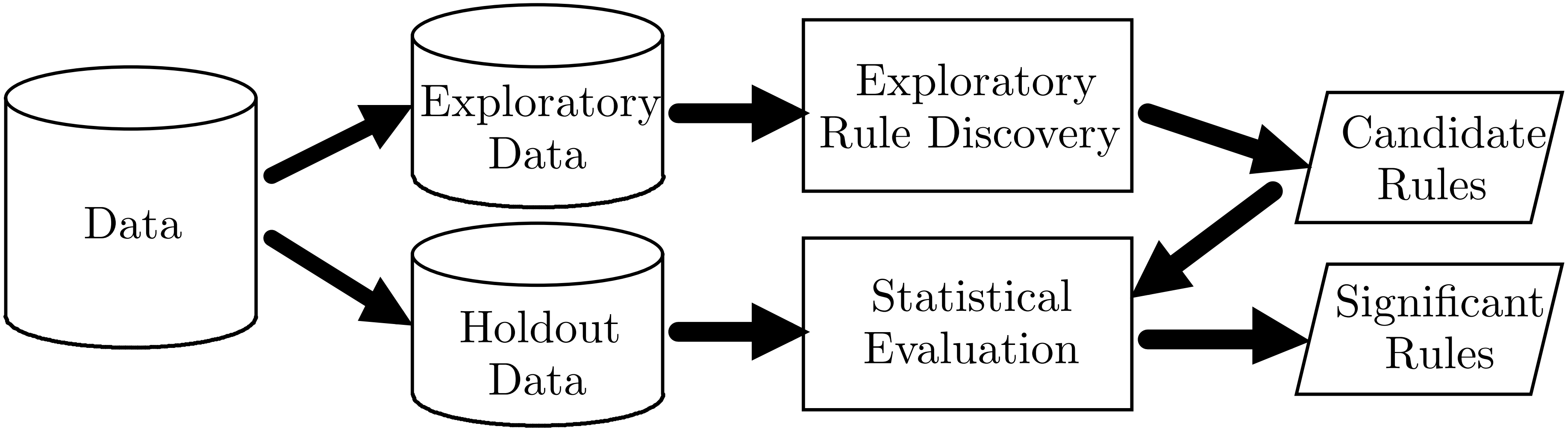
Mach Learn (2007) 68: 1–33
DOI 10.1007/s10994-007-5006-x

Discovering Significant Patterns

Geoffrey I. Webb

Received: 23 May 2005 / Revised: 13 February 2007 /
Accepted: 14 February 2007 / Published online: 14 April 2007

Discovering Significant Patterns


Data
Exploratory Data
Holdout Data
Exploratory Rule Discovery
Statistical Evaluation
Candidate Rules
Significant Rules