

CSE5334 DATA MINING

Lecture 12: Association Rule Mining (2) CSE4392/5334 Data Mining, Fall 2009
Department of Computer Science and Engineering, University of Texas at Arlington
Chengkai Li (Slides courtesy of Vipin Kumar and Jiawei Han)

Pattern Evaluation

Pattern Evaluation

- Association rule algorithms tend to produce too many rules
 - ▣ many of them are uninteresting or redundant
 - ▣ Redundant if $\{A,B,C\} \rightarrow \{D\}$ and $\{A,B\} \rightarrow \{D\}$ have same support & confidence
- Interestingness measures can be used to prune/rank the derived patterns
- In the original formulation of association rules, support & confidence are the only measures used

There are lots of measures proposed in the literature

Some measures are good for certain applications, but not for others

What criteria should we use to determine whether a measure is good or bad?

What about Apriori-style support based pruning? How does it affect these measures?

#	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	$\frac{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}{\sum_{i=1}^m \max(P(A_i, B_i), \sum_{j=1}^m P(A_j, B_j)) - \max(P(A_i) - \max_j P(A_j) - \max_k P(B_k))}$
3	Odds ratio (α)	$\frac{P(A,B)P(\bar{A},\bar{B})}{P(A,\bar{B})P(\bar{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(\bar{A},\bar{B}) - P(A,\bar{B})P(\bar{A},B)}{P(A,B)P(\bar{A},\bar{B}) + P(A,\bar{B})P(\bar{A},B)} = \frac{\alpha-1}{\alpha+1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\bar{A},\bar{B})} - \sqrt{P(A,\bar{B})P(\bar{A},B)}}{\sqrt{P(A,B)P(\bar{A},\bar{B})} + \sqrt{P(A,\bar{B})P(\bar{A},B)}} = \frac{\alpha-1}{\alpha+1}$
6	Kappa (κ)	$\frac{P(A,B) - P(A)P(B)}{1 - P(A)P(B) - P(A)P(B)}$
7	Mutual Information (M)	$\frac{\sum_{i,j} P(A_i, B_j) \log \frac{P(A_i, B_j)}{P(A_i)P(B_j)}}{\sum_i P(A_i) \log \frac{P(A_i)}{P(A)} + \sum_j P(B_j) \log \frac{P(B_j)}{P(B)}}$
8	J-Measure (J)	$\max(P(A) \log \frac{P(A,B)}{P(A)P(B)}, P(B) \log \frac{P(A,B)}{P(A)P(B)})$
9	Gini index (G)	$\max(P(A)[P(B A)^2 + P(\bar{B} A)^2] + P(\bar{A})[P(B \bar{A})^2 + P(\bar{B} \bar{A})^2] - P(B)^2 - P(\bar{B})^2)$
10	Support (s)	$P(A, B)$
11	Confidence (c)	$\frac{P(A, B)}{P(A)}$
12	Laplace (L)	$\frac{P(A, B) + 1}{N P(A) + 1}$
13	Conviction (V)	$\frac{P(A) - P(A, B)}{P(A)}$
14	Interest (I)	$\frac{P(A, B)}{P(A)}$
15	cosine (CS)	$\frac{P(A, B)}{\sqrt{P(A)P(B)}}$
16	Plattensky-Shapiro's (PS)	$\frac{P(A, B) - P(A)P(B)}{1 - P(A)}$
17	Certainty factor (F)	$\frac{P(A, B) - P(A)P(B)}{1 - P(A)}$
18	Added Value (AV)	$\frac{P(A, B) - P(A)P(B)}{P(A)}$
19	Collective strength (S)	$\frac{P(A, B) - P(A)P(B)}{P(A)P(B) + P(A)P(B)}$
20	Jaccard (J)	$\frac{P(A, B)}{P(A) + P(B) - P(A, B)}$
21	Klirgen (K)	$\frac{P(A, B)}{\sqrt{P(A)P(B)}}$

Computing Interestingness Measure

□ Given a rule $X \rightarrow Y$, information needed to compute rule interestingness can be obtained from a contingency table

Contingency table for $X \rightarrow Y$

	Y	\bar{Y}	
X	f_{11}	f_{10}	f_{1+}
\bar{X}	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	T

f_{11} : count of X and Y
 f_{10} : count of X and \bar{Y}
 f_{01} : count of \bar{X} and Y
 f_{00} : count of \bar{X} and \bar{Y}

Used to define various measures

- ◆ support, confidence, lift, Gini, J-measure, etc.

Drawback of Confidence

	Coffee	$\bar{\text{Coffee}}$	
Tea	15	5	20
$\bar{\text{Tea}}$	75	5	80
	90	10	100

Association Rule: Tea \rightarrow Coffee

Confidence = $P(\text{Coffee}|\text{Tea}) = 0.75$
 but $P(\text{Coffee}) = 0.9$
 \Rightarrow Although confidence is high, rule is misleading
 $\Rightarrow P(\text{Coffee}|\bar{\text{Tea}}) = 0.9375$

Statistical Independence

- Population of 1000 students
 - 600 students know how to swim (S)
 - 700 students know how to bike (B)
 - 420 students know how to swim and bike (S,B)
- $P(S \cap B) = 420/1000 = 0.42$
- $P(S) \times P(B) = 0.6 \times 0.7 = 0.42$
- $P(S \cap B) = P(S) \times P(B) \Rightarrow$ Statistical independence
- $P(S \cap B) > P(S) \times P(B) \Rightarrow$ Positively correlated
- $P(S \cap B) < P(S) \times P(B) \Rightarrow$ Negatively correlated

Statistical-based Measures

- Measures that take into account statistical dependence

$$Lift = \frac{conf(X \Rightarrow Y)}{sup(Y)} = \frac{P(Y|X)}{P(Y)}$$

= 1, independent
> 1, positively correlated
< 1, negatively correlated

$$Interest_factor = \frac{P(X,Y)}{P(X)P(Y)}$$

$$\phi - coefficient = \frac{P(X,Y) - P(X)P(Y)}{\sqrt{P(X)[1-P(X)]P(Y)[1-P(Y)]}}$$

= 0, independent
> 0, positively correlated
< 0, negatively correlated

Example: Lift/Interest

	Coffee	Coffee	
Tea	15	5	20
Tea	75	5	80
	90	10	100

Association Rule: Tea \rightarrow Coffee

Confidence = $P(\text{Coffee}|\text{Tea}) = 0.75$
 but $P(\text{Coffee}) = 0.9$
 $\Rightarrow Lift = 0.75/0.9 = 0.8333$ (< 1, therefore is negatively associated)

Drawback of Lift & Interest

	Y	Y	
X	10	0	10
X	0	90	90
	10	90	100

	Y	Y	
X	90	0	90
X	0	10	10
	90	10	100

$$Lift = \frac{0.1}{(0.1)(0.1)} = 10$$

$$Lift = \frac{0.9}{(0.9)(0.9)} = 1.11$$

Statistical independence:
 If $P(X,Y) = P(X)P(Y) \Rightarrow Lift = 1$

Example: ϕ -Coefficient

	Coffee	Coffee	
Tea	15	5	20
Tea	75	5	80
	90	10	100

Association Rule: Tea \rightarrow Coffee

$$\phi = \frac{0.15 - 0.9 \times 0.2}{\sqrt{0.9 \times 0.1 \times 0.2 \times 0.8}}$$

$$= -0.25$$
 (< 0, therefore is negatively correlated)

Drawback of ϕ -Coefficient

	Y	Y	
X	60	10	70
X	10	20	30
	70	30	100

	Y	Y	
X	20	10	30
X	10	60	70
	30	70	100

$$\phi = \frac{0.6 - 0.7 \times 0.7}{\sqrt{0.7 \times 0.3 \times 0.7 \times 0.3}} = 0.5238$$

$$\phi = \frac{0.2 - 0.3 \times 0.3}{\sqrt{0.7 \times 0.3 \times 0.7 \times 0.3}} = 0.5238$$

ϕ Coefficient is the same for both tables