Additional Association Rules Topics

Rule Pattern Evaluation

Association rule algorithms can produce large number of rules

Interestingness measures can be used to prune/rank the patterns

- In the original formulation, **support** and **confidence** are the only measures used

So to compute an interestingness measure, use a contingency table for X →Y or {X,Y}

Contingency table

	Y	$\overline{Y}$	
X	f_{11}	f_{10}	f_{1+}
$\overline{X}$	f_{01}	f_{00}	f_{o+}
	f_{+1}	f_{+0}	N

f_{11}: support of X and Y
f_{10}: support of X and $\overline{Y}$
f_{01}: support of $\overline{X}$ and Y
f_{00}: support of $\overline{X}$ and $\overline{Y}$

Used to define various measures

- support, confidence, Gini, entropy, etc.

Example drawback on confidence

Customers	Tea	Coffee	…
C1	0	1	…
C2	1	0	…
C3	1	1	…
C4	1	0	…
…			

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

Association Rule: **Tea → Coffee**

Confidence **P(Coffee|Tea) = 15/20 = 0.75**

The threshold is met, confidence > 50%, meaning people who drink tea are more likely to drink coffee than not drink coffee
So rule seems reasonable.

but **P(Coffee) = 0.9**, which means knowing that a person drinks tea reduces the probability that the person drinks coffee!

Note that **P(Coffee | ~Tea) = 75/80 = 0.9375**

Measures for Association Rules

So, what kind of rules do we really want?

Confidence(X→ Y) should be sufficiently high

> To ensure that people who buy X will more likely buy Y than not buy Y

Confidence(X → Y) > support(Y)

- Otherwise, rule will be misleading because having item X actually reduces the chance of having item Y in the same transaction
- There are many more measures that capture this constraint.

Statistical Independence

The criterion for statistical independence is that **confidence(X→ Y) = support(Y)**

which is is equivalent to: P(Y|X) = P(Y) and P(X,Y) = P(X) ***** P(Y)

- If P(X,Y) > P(X) * P(Y) : X & Y are positively correlated
- If P(X,Y) < P(X) *****P(Y) : X & Y are negatively correlated

$$Lift = \frac{P(Y \mid X)}{P(Y)}$$

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

lift is used for rules while interest is used for itemsets

$$PS = P(X,Y) - P(X)P(Y)$$

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

All of these measures are reported by Weka and RapidMiner.

Lift computations

Conf (X→Y) =σ(X,Y) / σ(Y)

$$\text{Lift } (X \rightarrow Y) = \frac{P(Y|X)}{P(Y)} = \frac{\text{conf}(X \rightarrow Y)}{\sigma(X)} = \frac{\sigma(X,Y)}{\sigma(X)*\sigma(Y)}$$

Lift/Interest example

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

Association Rule: Tea → Coffee

Confidence= P(Coffee | Tea) = 15/20 = 0.75
but P(Coffee) = 0.9

Thus, Lift = **0.75/0.9= 0.8333** = 0.15 / (0.9*0.2) = 0.8333

Lift here is < 1, and therefore negatively associated So,
is it enough to use confidence and/or lift for pruning?

	Y	$\bar{Y}$	
X	10	0	10
$\bar{X}$	0	90	90
	10	90	100

	Y	$\bar{Y}$	
X	90	0	90
$\bar{X}$	0	10	10
	90	10	100

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

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

Set of 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)	$\frac{\sum_j \max_k P(A_j,B_k)+\sum_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-\max_k P(B_k)}$
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	$\frac{P(A,B)P(\overline{AB})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{AB})+P(A,\overline{B})P(\overline{A},B)} = \frac{\alpha-1}{\alpha+1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\overline{AB})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{AB})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$
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)	$\max\left(P(A,B)\log(\frac{P(B\|A)}{P(B)})+P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),\right.$ $\left.P(A,B)\log(\frac{P(A\|B)}{P(A)})+P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})\right)$
9	Gini index (G)	$\max\Big(P(A)[P(B\|A)^2+P(\overline{B}\|A)^2]+P(\overline{A})[P(B\|\overline{A})^2+P(\overline{B}\|\overline{A})^2]$ $-P(B)^2-P(\overline{B})^2,$ $P(B)[P(A\|B)^2+P(\overline{A}\|B)^2]+P(\overline{B})[P(A\|\overline{B})^2+P(\overline{A}\|\overline{B})^2]$ $-P(A)^2-P(\overline{A})^2\Big)$
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)	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})}\times\frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
21	Klosgen (K)	$\sqrt{P(A,B)}\max(P(B\|A)-P(B),P(A\|B)-P(A))$

Simpson Paradox

Buy HDTV	Buy Exercise Machine		
	Yes	No	
Yes	99	81	180
No	54	66	120
	153	147	300

c({HDTV=Yes} -> {Exercise Machine=Yes} = 99/180 = 55%
c({HDTV=No} -> {Exercise Machine=Yes} = 54/120 = 45%

Thus the conclusion is that customers who buy HDTV are more likely to buy exercise machines.

Customer Group	Buy HDTV	Buy Exercise Machine		Total
		Yes	No	
College Students	Yes	1	9	10
	No	4	30	34
Working Adult	Yes	98	72	170
	No	50	36	86

College students:

$$c(\{\text{HDTV} = \text{Yes}\} \rightarrow \{\text{Exercise Machine} = \text{Yes}\}) = 1/10 = 10\%$$
$$c(\{\text{HDTV} = \text{No}\} \rightarrow \{\text{Exercise Machine} = \text{Yes}\}) = 4/34 = 11.8\%$$

Working adults:

$$c(\{\text{HDTV} = \text{Yes}\} \rightarrow \{\text{Exercise Machine} = \text{Yes}\}) = 98/170 = 57.7\%$$
$$c(\{\text{HDTV} = \text{No}\} \rightarrow \{\text{Exercise Machine} = \text{Yes}\}) = 50/86 = 58.1\%$$

Observed relationship in data may be influenced by the presence of other confounding factors (hidden variables).

Hidden variables may cause the observed relationship to disappear or reverse its direction.

Proper stratification is needed to avoid generating spurious patterns.

Mathematically

a/b < c/d and **p/q < r/s**, where a/b and p/q represent the confidence of the rule A-> B in two different strata, accounting for an additional variable and c/d and r/s represents the confidence of the rule ~A -> B in the same strata.

The paradox occurs when **(a+p)/(b+q) > (c+r)/(d+s),** which can lead to a flawed conclusion.

Effect of Support Distribution on Association Mining

Many real data sets have skewed support distribution

Support distribution of a retail data set