Mining Closed patterns
&
From Association Analysis to correlation Analysis

Mining closed Patterns

- An itemset X is closed in dataset D if there exists no proper super-itemset Y such that Y has the same support count as X in D.
- Closed frequent itemsets can reduce the number of patterns generated in frequent itemsets mining but preserves the information regarding set of frequent itemsets.
- From the closed frequent itemsets we can derive the frequent itemsets and their support.
- Search for closed frequent itemsets directly during mining process which prune the search space.
- Different pruning strategies : **Item Merging, sub-item pruning and item skipping**

Mining closed Patterns

- **Itemset merging:** if every transaction containing frequent itemset x also contains an itemset Y but not any proper superset of Y, then Y is merged with X forms frequent closed itemset and there is no need to search for any itemset containing X but no Y.

The projected conditional database for prefix item {I5:2} is {{I2,I1},{I2,I1,I3} , each of its transaction dataset contain itemset{I1,I2} and merged with {I5} to form closed itemset {I5,I1,I2:2} but we do not need to mine for closed itemsets that contain {I5} but not{I2,I1}

Mining closed Patterns

- **Sub-itemset pruning:** if Y ɔ X, and sup(X) = sup(Y), X and all of X's descendants in the set enumeration tree cannot be frequent closed itemsets and thus can be pruned
- Eg: {<a1,a2…. a100>,<a1,a2,……. a50>} min_supp=2
- Supp{a2}= supp{a1,a2…. a50}=2 since a2 is the proper subset of {a1,a2…a50} then a2 and its projected db cannot be examined
- **Item skipping:** In depth first mining, if a local frequent item has the same support in several header tables at different levels, one can prune it from the header table at higher levels.
- Eg: Because a2 has the same support in a1's projected and in the global header table a2 can be pruned from header

Which Patterns Are Interesting?—Pattern Evaluation Methods

- Pattern-mining will generate a large set of patterns/rules
  - Not all the generated patterns/rules are interesting
- Interestingness measures: Objective vs. subjective
  - Objective interestingness measures (statistics "behind" data)
    - Support, confidence, correlation,…..
  - Subjective interestingness measures: One man's trash could be another man's treasure
    - Query-based: Relevant to a user's particular request
    - Against one's knowledge-base: unexpected, freshness, timeliness
    - Visualization tools: Multi-dimensional, interactive examination

Limitation of the Support-Confidence Framework

Be careful!

- Are s and c interesting in association rules: "A => B" [s
- Example: Suppose one school may have the following statistics on # of students who may play basketball and/or eat cereal:

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

2-way contingency table

- Association rule mining may generate the following:

  –play-basketball => eat-cereal [40%, 66.7%] (higher s & c)

- But this strong association rule is misleading: The overall % of students eating cereal is 75% is more larger than 66.7%
- play basketball => not eat cereal [20%, 33.3%] is more accurate, although with lower support and confidence

Which Patterns Are Interesting?—Pattern Evaluation Methods

- Play basket ball and eating cereal are negatively associated the occurrence of one item actually decreases the likehood of other items.
- Without understanding there is possibility of making unwise decisions.
- The confidence rule A=>B can be deceiving, it does not measure the real strength of the correlation.
- Support -confidence measures are insufficient at filtering uninteresting rules.
- Leads to correlation rules
  - A=>B(s,c,corr)
- Lift is simple correlation measure, the occurrence of an itemset A is independent of occurrence of B if P(AUB)=P(A) P(B)

Interestingness Measure: Lift

- Measure of dependent/correlated events: **lift**

$$lift(B,C) = \frac{c(B \rightarrow C)}{s(C)} = \frac{P(B \cup C)}{P(B) \times P(C)}$$

Lift is more telling than s & c

	B	¬B	Σ_{row}
C	400	350	750
¬C	200	50	250
$\Sigma_{col.}$	600	400	1000

- Lift(B, C) may tell how B and C are correlated
  - Lift(B, C) = 1: B and C are independent
  - > 1: positively correlated
  - < 1: negatively correlated
- For our example,

$$lift(B,C) = \frac{400/1000}{600/1000 \times 750/1000} = 0.89$$

$$lift(B,\neg C) = \frac{200/1000}{600/1000 \times 250/1000} = 1.33$$

- Thus, B and C are negatively correlated since lift(B, C) < 1;
  - B and ¬C are positively correlated since lift(B, ¬C) > 1

Interestingness Measure: χ^2

- Another measure to test correlated events: χ^2

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

	B	¬B	Σ_{row}
C	400 (450)	350 (300)	750
¬C	200 (150)	50 (100)	250
Σ_{col}	600	400	1000

Expected value

Observed value

- General rules
  - $\chi^2 = 0$: independent
  - $\chi^2 > 0$: correlated, either positive or negative, so it needs additional test

$$\chi^2 = \frac{(400-450)^2}{450} + \frac{(350-300)^2}{300} + \frac{(200-150)^2}{150} + \frac{(50-100)^2}{100} = 55.56$$

- χ^2 shows B and C are negatively correlated since the expected value is 450 is less than the observed value 400

Lift and χ^2 : Are They Always Good Measures?

- Null transactions: Transactions that contain neither B nor C
  - Let's examine the dataset D
  - BC (100) is much rarer than B¬C (1000) and ¬BC (1000), but there are many ¬B¬C (100000)
  - Unlikely B & C will happen together! But, Lift(B, C) = 8.44 >> 1 (Lift shows B and C are strongly positively correlated!)
  - χ^2 = 670: Observed(BC) >> expected value (11.85) *Too many null transactions may "spoil the soup"!*

	B	¬B	Σ_{row}
C	100	1000	1100
¬C	1000	10000 0	101000
$\Sigma_{col.}$	1100	1[illegible]	[illegible]

null transactions

Contingency table with expected values added

	B	¬B	Σ_{row}
C	100 (11.85)	1000	1100
¬C	1000 (988.15)	10000 0	10100 0
$\Sigma_{col.}$	1100	10100 0	10210 0

Interestingness Measures & Null-Invariance

- *Null invariance:* Value does not change with the # of null-transactions

Measure	Definition	Range	Null-Invariant
$\chi^2(A,B)$	$\sum_{i,j=0,1} \frac{(e(a_i b_j) - o(a_i b_j))^2}{e(a_i b_j)}$	$[0,\infty]$	No
$Lift(A,B)$	$\frac{s(A \cup B)}{s(A) \times s(B)}$	$[0,\infty]$	No
$AllConf(A,B)$	$\frac{s(A \cup B)}{max\{s(A), s(B)\}}$	$[0,1]$	Yes
$Jaccard(A,B)$	$\frac{s(A \cup B)}{s(A) + s(B) - s(A \cup B)}$	$[0,1]$	Yes
$Cosine(A,B)$	$\frac{s(A \cup B)}{\sqrt{s(A) \times s(B)}}$	$[0,1]$	Yes
$Kulczynski(A,B)$	$\frac{1}{2}\left(\frac{s(A \cup B)}{s(A)} + \frac{s(A \cup B)}{s(B)}\right)$	$[0,1]$	Yes
$MaxConf(A,B)$	$max\{\frac{s(A)}{s(A \cup B)}, \frac{s(B)}{s(A \cup B)}\}$	$[0,1]$	Yes

X^2 *and lift are not null-invariant*

Jaccard, consine, AllConf, MaxConf, **and** ***Kulczynski*** **are** ***null-invariant***

Imbalance Ratio with Kulczynski Measure

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

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

- Null value cases are predominant in many large datasets
- Lift, **χ²** and cosine are good measures if null transactions are not predominant
- Otherwise, *Kulczynski* + *Imbalance Ratio* should be used to judge the interestingness of a pattern