Pattern Mining: Basic Concepts and Methods

- Basic Concepts
- Frequent Itemset Mining Methods – Apriori Algorithm
- Frequent Itemset Mining Methods – FPGrowth Algorithm
- Which Patterns Are Interesting? - Pattern Evaluation Methods

How to Judge if a Rule/Pattern Is Interesting?

- Pattern-mining will generate a large set of patterns/rules
  - Not all the generated patterns/rules are interesting
- **Objective Measures**: These are quantitative metrics like
  - **Support** (how often a pattern occurs),
  - **Confidence** (how often items in a rule are found together), and
  - **Correlation** (the strength of the relationship between items in a rule).
- **Subjective Measures**: These are qualitative and depend on the user's perspective, needs, or prior knowledge. They include:
  - **Relevance**: Is the pattern relevant to the user's specific query or need?
  - **Unexpectedness**: Does the pattern reveal something surprising against the user's existing knowledge base?
  - **Freshness**: Is the pattern new information, or is it already known?
  - **Timeliness**: Is the pattern currently relevant and timely?

Misleading "Strong" Association Rules

- Sometimes strong rules can be misleading.
- Consider a store with **10,000** total transactions
  - **6000** included **video games**
  - **7500** included **movies**
  - **4000** included **both video games and movies**
- Consider the rule **{video games} ⇒ {Movies}**
- $Support = \frac{4000}{10000}\ (40\%), Confidence = \frac{4000}{6000}\ (66\%)$
- If the parameters are set as min_sup = 30% and min_conf = 60%, this rule is a strong rule.
- But this is misleading because P(movies) = 75% which is larger than 66%
- Therefore, computer games and movies are negatively associated.
  - because the purchase of one of these items actually decreases the likelihood of purchasing the other.

Adding Correlation to the mix

- We saw that support and confidence measures are insufficient
- $A \Rightarrow B$ [support, confidence, **correlation**]
- A correlation rule is measured by support, confidence and the correlation between A and B
- Many different correlation measures
- We will mainly focus on two:
  - Lift
  - $\chi 2$

Limitation of the Support-Confidence Framework

Be careful!

- Are *s* and *c* interesting in association rules: "A $\Rightarrow$ B" [*s*, *c*]?
- 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* $\Rightarrow$ *eat-cereal* [40%, 66.7%] (higher s & c)
- But this strong association rule is misleading: The overall % of students eating cereal is 75% > 66.7%, a more telling rule:
  - ¬ *play-basketball* $\Rightarrow$ *eat-cereal* [35%, 87.5%] (high s & c)

Interestingness Measure: Lift

- The occurrence of itemset A is independent of the occurrence of itemset B if $P(A \cup B\ ) = P(A)P(B)$
- else, itemsets A and B are dependent and correlated
- $\boldsymbol{Lift(A,B)} = \frac{\boldsymbol{c(A \to B)}}{\boldsymbol{s(C)}} = \frac{\boldsymbol{s(A,B)}}{\boldsymbol{s(A) \times s(B)}} = \frac{\boldsymbol{P(A \cup B)}}{\boldsymbol{P(A)P(B)}}$

Lift is more telling than s & c

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

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

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

- ❑ 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
- ❑ B and C are **negatively** correlated
- ❑ B and ¬C are **positively** correlated

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

- For the table on the right,

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

- Lookup χ^2 distribution table → B, C are correlated
- Because the χ2 value is greater than 1,
  - and the observed value (400) < expected value (450), buying game and buying video are negatively correlated
- Thus, χ^2 is also more telling than the support-confidence framework

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

- Null transactions: Transactions that contain neither B nor C
- Let's examine the new 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	100000	101000
$\Sigma_{col.}$	1100	101000	102100

null transactions

Contingency table with expected values added

	B	¬B	Σ_{row}
C	100 (11.85)	1000	1100
¬C	1000 (988.15)	100000	101000
$\Sigma_{col.}$	1100	101000	102100

Interestingness Measures & Null-Invariance

- *Null invariance* means: The number of null transactions does not matter. Does not change the measure value.
- A few interestingness measures: Some are null invariant
- If you care about the null values (if huge imbalance?) use Null Invariant

Measure	Definition	Range	Null-Invariant?
$\chi^2(A,B)$	$\sum_{i,j} \frac{(e(a_i,b_j)-o(a_i,b_j))^2}{e(a_i,b_j)}$	$[0, \infty]$	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
$Kulczynski(A,B)$	$\frac{1}{2}(\frac{s(A \cup B)}{s(A)} + \frac{s(A \cup B)}{s(B)})$	$[0, 1]$	Yes
$MaxConf(A,B)$	$max\{\frac{s(A \cup B)}{s(A)}, \frac{s(A \cup B)}{s(B)}\}$	$[0, 1]$	Yes

Let

$$p = \frac{s(A \cup B)}{s(A)} = P(B|A)$$

$$q = \frac{s(A \cup B)}{s(B)} = P(A|B)$$

p, q are null invariant

Essentially min, max, mean variants of p, q

Null Invariance: An Important Property

- Why is null invariance crucial for the analysis of massive transaction data?
  - Many transactions may contain neither milk nor coffee!

milk vs. coffee contingency table

	$milk$	$\neg milk$	Σ_{row}
$coffee$	mc	$\neg mc$	c
$\neg coffee$	$m\neg c$	$\neg m\neg c$	$\neg c$
Σ_{col}	m	$\neg m$	Σ

- Lift and χ^2 are not null-invariant: not good to evaluate data that contain too many or too few null transactions!
- Many measures are not null-invariant!

Null-transactions w.r.t. m and c

Data set	mc	$\neg mc$	$m\neg c$	$\neg m\neg c$	χ^2	$Lift$
D_1	10,000	1,000	1,000	100,000	90557	9.26
D_2	10,000	1,000	1,000	100	0	1
D_3	100	1,000	1,000	100,000	670	8.44
D_4	1,000	1,000	1,000	100,000	24740	25.75
D_5	1,000	100	10,000	100,000	8173	9.18
D_6	1,000	10	100,000	100,000	965	1.97

Comparison of Null-Invariant Measures

- Not all null-invariant measures are created equal
- Which one is better?
  - D_4—D_6 differentiate the null-invariant measures

2-variable contingency table

	$milk$	$\neg milk$	Σ_{row}
$coffee$	mc	$\neg mc$	c
$\neg coffee$	$m\neg c$	$\neg m\neg c$	$\neg c$
Σ_{col}	m	$\neg m$	Σ

All 5 are null-invariant

Data set	mc	$\neg mc$	$m\neg c$	$\neg m\neg c$	$AllConf$	$Jaccard$	$Cosine$	$Kulc$	$MaxConf$
D_1	10,000	1,000	1,000	100,000	0.91	0.83	0.91	0.91	0.91
D_2	10,000	1,000	1,000	100	0.91	0.83	0.91	0.91	0.91
D_3	100	1,000	1,000	100,000	0.09	0.05	0.09	0.09	0.09
D_4	1,000	1,000	1,000	100,000	0.5	0.33	0.5	0.5	0.5
D_5	1,000	100	10,000	100,000	0.09	0.09	0.29	0.5	0.91
D_6	1,000	10	100,000	100,000	0.01	0.01	0.10	0.5	0.99

Subtle: They disagree on those cases

Summary

- Basic Concepts
  - What Is Pattern Discovery?   Why Is It Important?
  - Basic Concepts: Frequent Patterns and Association Rules
  - Compressed Representation: Closed Patterns and Max-Patterns
- Efficient Pattern Mining Methods
  - The Downward Closure Property of Frequent Patterns
  - The Apriori Algorithm
  - The FPGrowth Algorithm
  - Extensions or Improvements of Apriori
- Pattern Evaluation
  - Interestingness Measures in Pattern Mining
  - Interestingness Measures: Lift and χ^2
  - Null-Invariant Measures
  - Comparison of Interestingness Measures