Business Intelligence

G. Mudare

Mining Frequent Itemsets without Candidate Generation

Association Analysis to Correlation Analysis

- Sometimes the support and confidence measures are insufficient at filtering out uninteresting association rules.
- To tackle this weakness, a correlation measure can be used to augment the support-confidence framework for association rules.

$$A \Rightarrow B \ [support, confidence. correlation].$$

A Correlation rule is measured not only by its support and confidence but also By the correlation between itemsets A and B.

Correlation Measures

- Lift is a simple correlation measure: and states:
- The occurrence of itemset A is independent of the occurrence of itemset B if $P(A \cup B) = P(A)P(B)$; otherwise, itemsets A and B are dependent and correlated as events.
- The lift between the occurrence of A and B can be measured by computing

$$lift(A, B) = \frac{P(A \cup B)}{P(A)P(B)}.$$

- If the resulting value of this Equation is less than 1, then the occurrence of A is *negatively correlated with* the occurrence of B.
- If the resulting value is greater than 1, then A and B are *positively correlated*, meaning that the occurrence of one implies the occurrence of the other.
- If the resulting value is equal to 1, then A and B are *independent* and there is no correlation between them.

LIFT

- The Equation is equivalent to $P(B/A)/P(B)$, or $conf\,(A \rightarrow B)/sup(B)$, which is also referred as the *lift* of the association (or correlation) rule $A \rightarrow B$.
- The Lift it assesses the degree to which the occurrence of one “lifts” the occurrence of the other.
- eg
- if A corresponds to the sale of computer games and B corresponds to the sale of videos, then given the current market conditions, the sale of games is said to increase or “lift” the likelihood of the sale of videos by a factor of the value returned by Equation

Correlation analysis using lift

	game	$\overline{game}$	Σ_{row}
video	4,000	3,500	7,500
$\overline{video}$	2,000	500	2,500
Σ_{col}	6,000	4,000	10,000

A 2 x 2 contingency table summarizing the transactions with respect to game and video purchases.

We want to study how the two itemsets, A and B, are correlated. Let refer to the transactions that do not contain computer games, and refer to those that do not contain videos.

$$lift(A, B) = \frac{P(A \cup B)}{P(A)P(B)}.$$

From the table
$P(game)$ = 0:60,
$P(video)$ = 0:75,
$P(\{game; video\})$ = 0:40.

By Equation the lift is $P(game, video)/(P(game) \times P(video)) = 0.40/(0.60 \times 0.75) = 0.89$.
Because this value is less than 1, there is a negative correlation between the occurrence of **{*game*}** and {*video*}.
The numerator is the likelihood of a customer purchasing both, while the denominator is what the likelihood would have been if the two purchases were completely independent.
Such a negative correlation cannot be identified by a support confidence framework.

Correlation analysis

- The second correlation measure that we study is the χ_2 measure
- The χ_2 statistic tests the hypothesis that A and B are independent.
- The test is based on a significance level, with (r-1)x(c-1) degrees of freedom.
- To compute the χ_2 value, we take the squared difference between the observed and expected value for a slot (A and B pair) in the contingency table, divided by the expected value.
- This amount is summed for all slots of the contingency table.

Contingency table, with the expected values

	game	$\overline{game}$	Σ_{row}
video	4,000 (4,500)	3,500 (3,000)	7,500
$\overline{video}$	2,000 (1,500)	500 (1,000)	2,500
Σ_{col}	6,000	4,000	10,000

To compute the correlation using $\chi 2$ analysis, we need the observed value and expected value

$$\chi^2 = \Sigma \frac{(observed - expected)^2}{expected} = \frac{(4{,}000 - 4{,}500)^2}{4{,}500} + \frac{(3{,}500 - 3{,}000)^2}{3{,}000} + \frac{(2{,}000 - 1{,}500)^2}{1{,}500} + \frac{(500 - 1{,}000)^2}{1{,}000} = 555.6.$$

Because the $\chi 2$ value is greater than one, and the observed value of the slot (*game*, *video*) = 4,000, which is less than the expected value 4,500, *buying game* and *buying video* are *negatively correlated*.

How to calculate Expected

- Correlation analysis of categorical attributes using χ_2. Suppose that a group of 1,500 people was surveyed. The gender of each person was noted

  Each person was polled as to whether their preferred type of reading material was fiction or nonfiction. Thus, we have two attributes, *gender* and *preferred reading*.

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{N},$$

$$e_{11} = \frac{count(male) \times count(fiction)}{N} = \frac{300 \times 450}{1500} = 90,$$

Other correlation measures, *all confidence*

- *all confidence:*
- *Given an itemset* $X = \{i_1, i_2, : : : , i_k\}$, *the all confidence of X is defined as*

$$all\_conf(X) = \frac{sup(X)}{max\_item\_sup(X)} = \frac{sup(X)}{max\{sup(i_j)|\forall i_j \in X\}},$$

Constraint-Based Association Mining

- A data mining process may uncover thousands of rules from a given set of data, most of which end up being unrelated or uninteresting to the users
- The best is to have the users specify their expectations as *constraints* to confine the search space
- This strategy is known as constraint-based mining..

Mining constraints

1. **Knowledge type constraints:** These specify the type of knowledge to be mined, such as association or correlation.
2. **Data constraints:** These specify the set of task-relevant data.
3. **Dimension/level constraints:** These specify the desired dimensions (or attributes) of the data, or levels of the concept hierarchies, to be used in mining.
4. **Interestingness constraints:** These specify thresholds on statistical measures of rule interestingness, such as support, confidence, and correlation.
5. **Rule constraints:** These specify the form of rules to be mined.

Metarule-Guided Mining of Association Rules

- Metarules allow users to specify the syntactic form of rules that they are interested in mining.
- The rule forms can be used as constraints to help improve the efficiency of the mining process.

$$P_1(X, Y) \wedge P_2(X, W) \Rightarrow buys(X, \text{“office software”}),$$

In general, a metarule forms a hypothesis regarding the relationships that the user is interested in probing or confirming.

$$age(X, \text{“30...39”}) \wedge income(X, \text{“41K...60K”}) \Rightarrow buys(X, \text{“office software”})$$