Correlation Analysis

Shubham kumar
Dept. of CS&IT
MGCUB

- The rules derived after Association may or may not be interesting to the users.
- This is especially true when mining at low support threshold or mining for long patterns.
- Let us take an example of uninteresting strong association rules where we have to analyze transactions at AllElectronics with respect to the purchase of computer games and videos. Of the 10,000 transactions analyzed, the data show that 6000 of the customer transactions included computer games, while 7500 included videos, and 4000 included both computer games and videos.

Let minimum support = 30%

minimum confidence= 60%

The association rule discovered is:

buys(X, ” computer games”) => buys(X, “videos”)

[support= 40%, confidence= 66%]

Since it satisfies the minimum support and minimum confidence threshold.

Therefore this derived association rule is strong.

Even the above association rule is strong but it is misleading because the probability of purchasing videos is 75 %, which is even larger than 66%.

Or, we can say that computer games and videos are negatively associated because the purchase of one of these items decreases the chance of purchasing the other.

From this example we can say that strong association rule may be uninteresting.

- Now we can say that, the **support and confidence** measures are **not** sufficient to find interesting association rules.

- To tackle this weakness, a correlation measure can be used.

This leads to correlation rules of the form:

$A \Rightarrow B$ [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.

- There are many different correlation measures like Lift, X^2 measures etc.

Lift: The lift between the occurrence of A and B can be measured as:

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

- If the resulting value is less than 1, then the occurrence of A is negatively correlated with the occurrence of B, means the occurrence likely leads to the absence of the other one.

- 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.
- The lift can also be calculated as:

$$lift(A, B) = P(B/A) / P(B)$$

or, **confidence(A=> B)/ support(B).**

In other words, lift accesses the degree to which the occurrence of one “lifts” the occurrence of the other.

Example- let us take the same example that are taken in slide no. 2.

2x2 contingency table for the transactions is:

	$game$	$\overline{game}$	Σ_{row}
$video$	4000	3500	7500
$\overline{video}$	2000	500	2500
Σ_{col}	6000	4000	10,000

From the table we can find:

Probability of purchasing a computer game is

P(game)=6000/10000 = 0.60

Probability of purchasing a video is:

P(video)=7500/10000 = 0.75

Probability of purchasing both is:

P({ game, video})= 4000/10000 = 0.40

Now, we have to find the lift of the association rule

buys(X, ” computer games”) => buys(X, “videos”)

[support= 40%, confidence= 66%]

lift(computer games, video) =

$$P(\{games, video\})/P(games)P(videos)$$

$$= 0.40/(0.60*0.75)$$

$$= 0.89$$

since the lift value of the rule is less than 1, hence we can say that there is a negative correlation between the occurrence of computer games and videos.

- X^2 (chi-square) measure:

We need the observed value and the expected value for each slot of the contingency table to compute the correlation using X^2 .

The expected values are calculated based on the data distributions.

2x2 contingency table with expected values in parenthesis.

	game	$\overline{game}$	Σ_{row}
video	4000 (4500)	3500 (3000)	7500
$\overline{video}$	2000 (1500)	500 (1000)	2500
Σ_{col}	6000	4000	10,000

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected} = \frac{(4000-4500)^2}{4500} + \frac{(3500-3000)^2}{3000}$$
$$+ \frac{(2000-1500)^2}{1500} + \frac{(500-1000)^2}{1000} = 555.6.$$

since the value is greater than 1, and the observed value of (game, video) = 4000 which is less than the expected value i.e. 4500.

Hence buying games and buying video are negatively correlated.

Assignment: (Deadline 04/05/20)

The following contingency table summarizes supermarket transaction data, where *hot dogs* refers to the transactions containing hot dogs, $\overline{hot\ dogs}$ refers to the transactions that do not contain hot dogs, *hamburgers* refers to the transactions containing hamburgers, and $\overline{hamburgers}$ refers to the transactions that do not contain hamburgers.

	hot dogs	$\overline{hot\ dogs}$	Σ_{row}
hamburgers	2000	500	2500
$\overline{hamburgers}$	1000	1500	2500
Σ_{col}	3000	2000	5000

(a) Suppose that the association rule "*hot dogs* $\Rightarrow$ *hamburgers*" is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%, is this association rule strong?

(b) Based on the given data, is the purchase of *hot dogs* independent of the purchase of *hamburgers*? If not, what kind of *correlation* relationship exists between the two?

Reference

- Jiawei Han, Micheline kamber and Jian pei. “DATA MINING concepts and Techniques” 3/e, Elsevier, 2012