The Evolution of the Association Rules

Varshali Jaiswal and Jitendra Agarwal

***Abstract*—Association rules is a popular and well researched method for discovering interesting relation between variables in large databases and association rules, is one of the most important tasks in data mining. The generated strong association rules is depend on the association rule extraction by any algorithm, for example Apriory algorithm or Fp growth etc and the evolution of the rules by interestingness measure, for example support and confidence, lift or interest, correlation coefficient, statistical correlation, leverage, conviction etc.**

The association rules mining are dependent on both steps equally. The classical model of association rules mining is support-confidence, the interestingness measure of which is the confidence measure. The classical interestingness measure in Association Rules have existed some disadvantage.

This paper present measurements that are support and confidence, interest or lift, chi-square test for independency, correlation coefficient, statistical correlation to calculate the strength of association rules. There are other interestingness measures, besides support and confidence which include generality, reliability, peculiarity, novelty, surprising ness, utility, and applicability. This paper investigates the evolution association rule mining.

***Index Terms*—Association rules, support /confidence, interest/lift, chi-square test for independency, correlation coefficient, statistical correlation.**

I. Introduction

In the previous few years a lot of work is done in the field of data mining especially in finding association between items in a data base of customer transaction. Association rule mining, one of the most important and well researched techniques of data mining [1]. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories [2].

Nowadays, association rules mining from large databases is an active research field of data mining motivated by many application areas such as telecommunication networks, market and risk management, inventory control etc.

II. Rules Measurement and Selection

One challenge for the association rule mining is the rules measurement and selection. Since the data mining methods are mostly applied in the large datasets, the association mining is very likely to generate numerous rules from which it is difficult to build a model or summarize useful information. A simple but widely used approach to help mitigate this problem is to gradually increase the threshold value of support and confidence until a manageable size of rules is generated. It is an effective way to reduce the number of rules; however it may cause problems in the results as well. The major concern is that by increasing the minimum support and confidence value, some important information may be filtered out while the remaining rules may be obvious or already known. The data mining is a process involving interpretation and evaluation as well as analysis. For association rule mining, the evaluation is an even more important phase of the process.

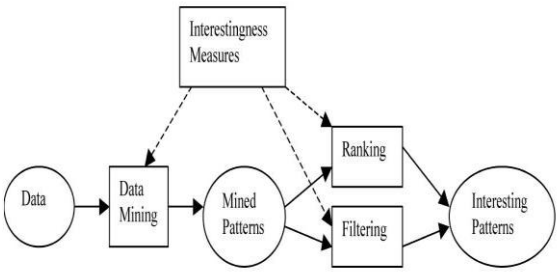

Fig. 1. The evaluation

This paper is divided in to three sections the first section gives the formal definition and some explanation of each measure. The second section gives us the calculation of each measure on our sample data and the last section contains our recommendation on using which measure for discovering the interesting rules.

III. The Basic Concepts

Support and Confidence: Support [1] is defined as the percentage of transactions in the data that contain all items in both the antecedent and the consequent of the rule,

$$S=P(X\cap Y) = \{X\cap Y\}/\{D\}$$

Confidence is estimates of the conditional probabilities of Y given X, i.e. $P(X\cap Y)/P(X)$.

$$C= P(X\cap Y)/P(X)$$

The support of a rule is also important since it indicates how frequent the rule is in the transactions. Rules, which have very small support, are often uninteresting since they do not describe significantly large populations.

A rule that has a very high confidence (i.e., close to 1.0) is very important because it provides an accurate prediction on the association of the items in the rule.

The disadvantage of this, it is not trivial to set good values for the minimum support and confidence thresholds.

Fundamental critique in so far that the same support threshold is being used for rules containing a different number of items.

Lift or Interest: A few years after the introduction of association rules, researchers [3] started to realize the disadvantages of the confidence measure by not taking

Manuscript received July 24, 2012; revised August 28, 2012.

Varshali Jaiswal is with RKDF, India (e-mail: varshalijaswal@gmail.com)

into account the baseline frequency of the consequent. Therefore, lift, originally called Interest, it measures the number of times X and Y occur together compared to the expected number of times if they were statistically independent. It is presented as:

$$\mathbf{I= P\ (X\cap Y)/P(X)\ P(Y)}$$

Since *P(Y)* appears in the denominator of the interest measure, the interest can be seen as the confidence divided

If I <1, then X and Y appear less frequently together in the data than expected under the assumption of conditional independence. *X* and *Y* are said to be negatively interdependent.

If I =1, then X and Y appear as frequently together as expected under the assumption of conditional independence. *X* and *Y* are said to be independent of each other.

If I >1, then X and Y appear more frequently together in the data than expected under the assumption of conditional independence. *X* and *Y* are said to be positively interdependent.

Advantage:

The difference between confidence and lift lies in their formulation and the corresponding ding limitations. Confidence is sensitive to the probability of consequent (Y). Higher frequency of Y will ensure a higher confidence value even if there is not true relationship between X and Y. But if we increase the threshold of the confidence value to avoid this situation, some important pattern with relatively lower frequency may be lost. In contrast to confidence, lift is not vulnerable to the rare items problem. It is focused on the ratio between the joint probability of two itemsets with respect to their expected probabilities if they are independent. Even itemsets with lower frequency together can have high lift values [4].

Disadvantages:

The first one is related to the problem of sampling variability (see section Empirical Bayes Estimate). This means that for low absolute support values, the value of the interest measure may fluctuate heavily for small changes in the value of the absolute support of a rule. This problem is solved by introducing a Empirical Bayes estimate of the interest measure.

The second problem is that the interest measure should not be used to compare the interestingness of itemsets of different size. Indeed, the interest tends to be higher for large itemsets than for small itemsets [6].

Chi-square Test for Independency: A natural way to express the dependence between the antecedent and the consequent ofan association Rule XUY is the correlation measure based on the Chi-square test for independence [3].

$$x=\sqrt{\sum x\sum y(O_{XY}-E_{XY})^2/E_{XY}}$$

The chi-square test for independence is calculated as follows, with Oxy the observed frequency in the contingency table and Exy the expected frequency (by multiplying the row and column total divided by the grand total) Therefore, the $\chi 2$is a summed normalized square deviation of the observed values from the expected values. It can then be used to calculate the p-value by comparing the value of statistics to a chi-square distribution to determine the significance level of the rule. For instance, if the p-value is higher than 0.05 (when $\chi 2$ value is less than 3.84), we can tell X and Y are significantly independent, and therefore the rule X => Y can be pruned from the results.

Advantages:

The advantage of the chi-square measure, on the other hand, is that it takes into account all the available information in the data about the occurrence or non-occurrence of combinations of items, whereas the lift/interest measure only measures the co-occurrence of two itemsets, corresponding to the upper left cell in the contingency table.

Disadvantages:

First of all, the Chi-square test rests on the normal approximation to the Binomial distribution. Thiapproximation breaks down when the expected values (Exy) are small [5].

The Chi-square test should only be used when all cells in the contingency table have expected values greater than 1 and at least 80% of the cells have expected values greater than 5.

The Chi-square test will produce larger values when the data set grows to infinity. Therefore, more items will tend to become significantly interdependent if the size of the dataset increases. The reason is that the Chi-square value depends on the total number of transactions, whereas the critical cutoff value only depends on the degrees of freedom (which is equal to 1 for binary variables) and the desired significance level. Therefore, whilst comparison of Chi-squared values within the same data set may be meaningful, it is certainly not advisable to compare Chi-squared values across different data sets.

Correlation Coefficient: The [7] correlation coefficient (also known as the Φ-coefficient) measures the degree of linear interdependency between a pair of random variables. It is defined by the covariance between the two variables divided by their standard deviations:

$$\rho_{XY}=(p(x\cap y)-p(x)^{*}p(Y))/\sqrt{p(x)(1-p(x))}*\sqrt{p(y)(1-p(y))}$$

where ρXY = 0 when X and Y are independent and ranges from [-1, +1].

Statistical Correlation: To [8] get the association rules with real correlation; this measure put forward statistical correlation from the view point of statistics to compensate the deficiency of support-confidence. Statistical correlation is defined as equation, which is

$$S_{corr(x\cup y)}=\left|D\right|\sup(x\cup y)\left|D\right|\pi_{(x\cup y)}\sup(i)$$
$$/\sqrt{\left|D\right|\pi_{(x\cup y)}\sup(i)\left|D\right|(1-\pi_{(x\cup y)}\sup(i))}$$

If Scorrelation {XUY}<0, it denotes that the items in antecedent X and the consequent Y of an association rule are negative correlation, and the items have a relationship of restricting each other.

If Scorrelation {XUY}=0, it means that the items in

antecedent X and the consequent Y of an association rule are independent, and the items are not mutually influence.

If Scorrelation {XUY}>0, it represents that the items in antecedent X and the consequent Y of an association rule have some degree correlation, and correlation is more and more strong with the Scorrelation increase.

Advantages:

Scorrelation, which can enhance the correlation degree of items in association rule and cut negative correlation rules.

IV. Experiment and Ananlsis

The sample data (Table I) for the analysis purpose is taken from a store database of customer transaction there are six different types of items and a total of ten transactions. In each transaction a 1 represents the presence of an item while a 0 represents the absence of an item from the market basket.

A. Experiment Process

TABLE I: Sample Transactions

Tid	Items					
	soap	shampoo	hair oil	tooth paste	tooth brush	comb
T1001	1	1	0	1	0	1
T1002	1	0	1	1	0	1
T1003	1	0	1	1	0	1
T1004	0	1	1	1	0	0
T1005	0	1	0	1	1	0
T1006	1	0	0	0	1	1
T1007	1	0	1	0	1	1
T1008	0	0	1	0	0	0
T1009	0	1	1	1	0	0
T1010	1	1	0	1	1	0
TOTAL	6	5	6	7	4	5

The frequent item set generated by the sample data using A-priori algorithm is shown in the following Table II.

TABLE II: The Frequent Item

Itemsets	Support
{ soap, tooth paste }	40%
{ soap, comb }	50%
{ shampoo, tooth paste }	50%
{ hair oil, tooth paste }	40%

All measures are calculated for each rule in table II, which is output of the A-priori algorithm. The results are shown in table III

B. Experiment Results

TABLE III: Calculation of Measure on Sample Datasets

Rules	Support	confidence	Lift	Chi-square test	Correlation	Statistical Correlation
soap → tooth paste	0.40	0.66	0.955	5.86	-0.089	-0.040
Tooth paste → soap	0.40	0.57	0.955	5.86	-0.089	-0.040
soap → comb	0.50	0.83	1.66	0.915	+0.8179	+0.522
comb→ soap	0.50	1.00	1.66	0.915	+0.8179	+0.522
Shampoo → tooth paste	0.50	1.00	1.42	1.713	+0.655	+0.315
tooth past → shampoos	0.50	0.71	1.42	1.713	+0.655	+0.315
hair oil → tooth paste	0.40	0.66	0.95	8.613	-0.089	-0.040

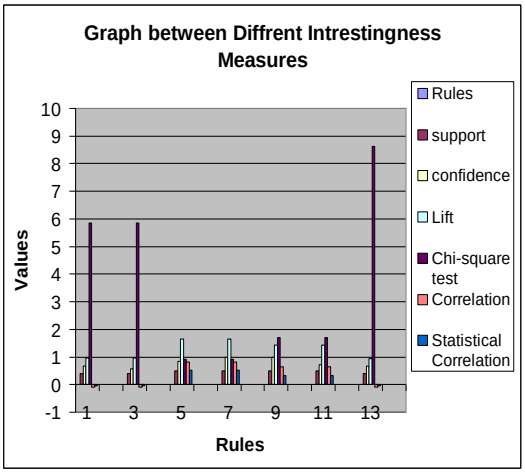

Fig. 2. Different interestingness measures

V. Conclusions

It is generally accepted that there is no single measure that is perfect and applicable to all problems. Usually different measures are complementary and can be applied at different applications or phases. The works of this dissertation propose a new model to measure the confidence of association rules based on sufficiency measure of uncertain reasoning, which overcome some shortages of the metric of the classical model support-confidence framework. The new measure not only captures correlation but can also detect negative implication. We have study some properties of the new measure and test its validity.

The following suggestions can be formulated based on the analysis of the different interestingness measures discussed in the previously with example:

• Confidence is never the preferred method to compare association rules since it does not account for the baseline frequency of the consequent.

• The lift/interest value corrects for this baseline frequency but when the support threshold is very low, it may be instable due to sampling variability. However, when the data set is very large, even a low percentage support threshold will yield rather large absolute support values. In that case, we do not need to worry too much about sampling variability. A drawback of the interest measure is that it cannot be used to compare itemsets or rules of different size since it tends to overestimate the interestingness for large itemsets.

• When association rules need to be compared between data sets of different sizes, the Chi-square test for independence and Correlation analysis are not preferred since they are highly dependent on the dataset size. Both measures tend to overestimate the interestingness of itemsets in large datasets.

REFERENCES

[1] C. C. Aggarwal and P. S. Yu, "A New Framework for Item Set Generation," in *Proceedings of the ACM PODS Symposium on Principles of Database Systems*, Seattle, Washington (USA), pp. 18-24, 1998.

[2] A. Agresti, *An Introduction to Categorical Data Analysis*, Wiley Series in Probability and Statistics, 1996.

[3] T. Brijs, G. Swinnen, K. Vanhoof, and G. Wets, "The use of association rules for product assortment decisions: a case study," in *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*, San Diego (USA), August 15-18, pp. 254-260, 1999.

[4] R. Agrawal, T. Imielinski, and A. N. Swami, "Mining Association Rules between Sets of Items in Large Databases," in *Proceedings of the* 1993ACM *SIGMOD Conference*, pp. 207–216, 1993.

[5] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation[A]," in *Proceeding of 2000 ACM-SIGMOD International Conference on Management of Data*[C], pp. 1–12, 2000.

[6] R. Agrawal R. S. "Fast Algorithms for Mining "Association Rules," in *Proc. 20th Int. Conf. On Ver Large DataBases*, 1994, pp. 487-499.

[7] J. H. Liu, "A New Interestingness Measure of Association Rules," *Second International Conference on Genetic and Evolutionary Computing*, 2008.

[8] J. Hu and X. Y. Li, *Association Rules Mining Based on Statistical Correlation*, 2008.

[9] A. Silberschatz A T. "What Makes pattern interesting in kownledge discovery systems," *IEEE Transactions on Knowledge and Data* Engineering, 1996, vol. 8, no. 6, pp. 970-974

[10] T. Brijs, K. Vanhoof, and G. Wets, "Defining Interestingness For Association Rules," *International Journal Information Theories & Applications*, vol.10.

[11] A. S. V. Semester, "Data warehousing and data mining," *Raipur Institute of Technology*, Raipur.

Varsshali Jaiswal was born in India on 21/12/1982 and she is pursuing PhD in information Technology from SGSITS, Indore MP, India. Her areas of interest include Data Mining, Design and Analysis of Algorithm, image processing. Received her Master of Technology in Information Technology from SOIT, UIT, RGPV, Bhopal MP India in 2009 and done Bachelor of Engineering in Information Technology from SGSITS, Indore MP India in 2005. She is currently Assistant Professor of Information Technology Department in RKDF, Indore.