New Measure of Interestingness for Efficient Extraction of Association Rules

Parvati Bhurani, Mushtaq Ahmed, and Yogesh Kumar Meena

Malaviya National Institute of Technology, Jaipur, India
parvati_bhurani@rediffmail.com, mushtaqahmed@mnit.ac.in,
yogimnit@gmail.com

Abstract. Data Mining helps to uncover the already unknown and non-redundant knowledge in large databases, which can be used for decision making purpose. Association rule mining is one of the key research area in the field of Data Mining. Association rule mining can be considered as unsupervised learning model, it discovers the interesting relationship among large set of data items on the basis of some predefined threshold. Support-confidence is the classical model used for the rule mining purpose, it uses confidence for final rule generation but it has some limitations. As sometimes it can generate those rules which are not positively correlated and thus can mislead the decision maker. In this paper we addressed the problems associated with existing approach and also proposed two new measure of interestingness to deal with these problems. The new measures have been tested for their correctness.

Keywords: Association Rules, Interestingness Measure, Support-confidence, Correlation.

1 Introduction

Data Mining plays an important role to extract out of sight or concealed patterns from large datasets. The aim of Data Mining technique is to uncover the previously unknown, useful, and non-redundant cognition in large databases. It has a potential to help companies to focus on the most important information in their data warehouse. It has been used in many industries like retail market, insurance and banking etc. to increase sales, risk assessment and many more. Data Mining is also referred as knowledge discovery in databases (KDD).

During mining large number of rules are generated but only the small set has significance for the user point of view, so to find whether a rule is of interest or not it must need a suitable metric to measure the degree of rule in which the user is interested. Thus the role of interestingness measure plays an important role in rule extraction process [3]. Association rule mining or frequent pattern mining is a customary and well explored method for discovering interesting relations between itemsets in large data repository under the domain of Data Mining. The concept was first introduced by Agrawal [1,2] to analyze the data of a supermarket containing large collection of customer transaction. It is

S.C. Satapathy et al. (Eds.): Proceedings of the InConINDIA 2012, AISC 132, pp. 103–110.
springerlink.com

a kind of unsupervised learning technique. The following is the general form of an association rule:

$$X \Rightarrow Y$$

In the above rule X is known as antecedent and Y is known as consequent and it can be read as if X then Y. It measures the association between X and Y. Support-confidence model [2] for association rule can be described as given: Let $I = i_1, i_2, ..., i_n$ be a set of n literals called items and T be a set of transactions, where each transaction $t \in T$ is a set of items such that $t \subseteq I$. Each transaction is associated with a unique id TID. An association rule is the implication of the form, $X \Rightarrow Y$, where $X \subseteq I, Y \subseteq I$ and $X \cap Y = \emptyset$. Here X is the antecedent and Y is the consequent. Terms used in rule mining process defined are as follows:

Support(sup): Support of an itemset can be defined as the ratio of transactions containing the items in both antecedent and consequent of the rule to the total number of transaction. It measures the strength of the given itemset. Let $X \Rightarrow Y$ be the rule then support is given as below:

$$sup\left(X \Rightarrow Y\right) = \frac{sup\_count\left(X \cap Y\right)}{T} \tag{1}$$

Here sup_count is the number of transactions containing the given itemset $X \cap Y$.

Confidence(conf): Confidence of an association rule measures how often items in Y appear in transaction that contains X. It measures the strength of a given association rule.

$$conf\left(X \Rightarrow Y\right) = \frac{sup\left(X \cup Y\right)}{sup\left(X\right)} \tag{2}$$

The rule mining process works in two steps:

Frequent itemset identification: The itemset which satisfy the minimum support threshold (σ) are generated in this phase.

Rule extraction: This phase takes frequent itemset as input which has been generated in step 1 along with the value of minimum confidence threshold (τ). The rules which satisfy the minimum confidence are extracted in this phase.
The remaining paper is organized as follows: In section 2, some of the related work in this area is presented. Limitation in existing approaches are discussed in section 3. In section 4 we present the new measures of interestingness while in section 5, new measure are tested on sample dataset. Conclusion is given in last section.

2 Related Work

Association rule mining is one of the actively researched area in the Data Mining community. Lot of work has been reported in this domain and used in various applications in real world. The algorithms discussed here are generally variation of classical apriori approach and involves a time consuming candidate generation process. Apriori [1] is one of important algorithm for mining the frequent itemset. The process of rule discovery is divided into two subproblems the first is to find the frequent itemset with minimum support threshold while the second is to find the association rules from the frequent itemset generated in first phase with the help of minimum confidence. Some variations of apriori are also proposed like partition based [6], sampling technique [7] to deal with the associated problems. Two major issues are observed in above methods as all the addressed approaches are variation of classical apriori algorithm thus involves the time consuming candidate generation process and these approaches also suffer from a rare item problem because of single minimum support threshold. These two issues are discussed by many researchers and solutions are also proposed to deal with these issues. FP-Growth [9], it uses a new tree data structure for storing compressed information regarding the frequent patterns. FP-growth is better than other approaches as FP-tree is constructed which removes the requirement of costly database scan and a pattern growth approach avoids costly candidate generation and thus it is time efficient. CFP-Growth, it is the extension of FP-growth algorithm and uses CFP-tree structure in place of FP-tree for storing the frequent itemset. The method discussed in this section works as per traditional support-confidence model and thus suffers from limitation like generation of rules which are not important as per the user criteria. We discuss this problem in section 3 with some example.

3 Limitation of Support-Confidence Framework

Support-confidence framework uses confidence measure for the rule generation process, as in some cases it may discover uncorrelated rules which can mislead the decision maker. Let us consider the $\sigma = 20\%$ and $\tau = 40\%$. Following is the sample dataset taken from supermarket illustrates the above addressed limitation: From the above data we can have following set of association rules:

Table 1. Sample Dataset

Item/Tran_No	1	2	3	4	5	6	7	8	9	10
Bread	1	1	1	1	0	0	0	0	0	0
Jam	1	1	1	0	1	1	1	1	1	0
Egg	0	1	1	0	1	1	1	0	0	0
Butter	1	0	0	1	0	1	1	1	1	1

Rule1: bread $\Rightarrow jam\,[30\%, 75\%]$
Rule2: bread $\Rightarrow egg\,[20\%, 50\%]$
Rule3: jam $\Rightarrow egg\,[50\%, 62.5\%]$
Rule4: jam $\Rightarrow butter\,[50\%, 62.5\%]$

In the above mentioned rules, first parameter is the support of the rule while the second parameter is the confidence of the rule. As per the confidence value the first rule is very strong but in reality there is negative correlation between bread and jam as support of jam alone is 80% which is greater than the confidence value, while the third rule has lower confidence than the first rule but it has a positive correlation between jam and egg as the support of egg alone is 50% which is lower than the confidence value and thus if we associate egg with jam then support of egg will be increased. Now consider second rule then we can observe that there is no association between the two variables i.e. bread and egg are independent to each other as support of egg is 50% which is equal to the confidence value.

Suggested Solution: The above example suggests that it is not necessary that all strong rules i.e. the rules with high confidence value are interesting. The problem occurs because in confidence we do not consider the baseline frequency of consequent. So to overcome this problem base line frequency of consequent is to be taken into consideration while measuring the correlation among different itemset. In following section we will see the new interest measure with their properties.

4 Proposed Objective Interest Measure

In this section we introduce two new measure of interestingness namely ratioPS and ratioLEV. We discuss these two new measures along with their set of properties and possible range. Let us consider $X \Rightarrow Y$ as an association rule.

ratioPS: It is an asymmetric interest measure and hence the value of ratioPS(XY) is different from ratioPS(YX). Following is the formula for this new measure:

$$ratioPS = \frac{P(XY) - P(X) * P(Y)}{1 - P(X)} \tag{3}$$

The possible range of this measure is from -1 to +1. We can consider this measure as the ratio of PS and the probability of $\bar{X}$.

Table 2. Correlation criteria between two itemset X and Y

S.No.	Value	Correlation
1	If ratioPS($X \Rightarrow Y \prec 0$)	Negatively Correlated
2	If ratioPS($X \Rightarrow Y = 0$)	No Correlation
3	If ratioPS($X \Rightarrow Y \succ 0$)	Positively Correlated

ratioLEV: This measure can be considered as the ratio of leverage to the product of probability of $\bar{X}$ and Y.

$$ratioLEV = \frac{\frac{P(XY)}{P(X)} - P(X) * P(Y)}{(1 - P(X)) * P(Y)} \tag{4}$$

The possible range of this measure can vary from 0 to ∞. Let us consider a transaction dataset of 10 transactions, this example illustrate how the new measure identify the true associations between two itemsets namely X and Y:

Table 3. Correlation between two itemset X and Y

S.No.	Value	Correlation
1	If ratioLEV$(X \Rightarrow Y \prec 1)$	Negatively Correlated
2	If ratioLEV$(X \Rightarrow Y = 1)$	No Correlation
3	If ratioLEV$(X \Rightarrow Y \succ 1)$	Positively Correlated

Table 4. Results of the new interest measure

S.No.	P(XY	P(X)	P(Y)	Conf.	ratioPS	ratioLEV	Remark
1	0.3	0.4	0.8	0.75	-0.33	0.89	Negative Correlated
2	0.2	0.4	0.5	0.5	0.0	1.0	Independent
3	0.4	0.4	0.7	1.0	1.7	0.2	Positive Correlated

From the above example we can see that both the proposed measures are capable to detect the negatively correlated itemset.

4.1 Properties Followed by the New Measures

In this section we see the set of properties which should be followed by various interest measures and then we will show the properties satisfied by new measures. Following is the description of properties:

Property 1: This property checks whether the two itemset are independent or not. If both X and Y are independent then the value of P=0.

Property 2: If value of joint probability i.e. P(X,Y) increases when the antecedent (P(X)) and consequent (P(Y)) probability remain unchanged. Then the value of P increases.

Property 3: If P(X) decreases when P(Y) and P(X,Y) remain unchanged or P(Y) decreases when P(X) and P(X,Y) remain unchanged. Then the value of P increases.
The above properties proposed by Piatetsky-Shapiro [4] the first property can be relaxed as some measure has P=1 when the two itemset are independent.

Property 4: The property is known as symmetry under variable permutation, If the itemset in the antecedent and consequent are exchanged then there is no change in value of P before and after the permutation.

Property 5: The property is known as scaling invariance under row or column, If row or column values in the contingency table are multiplied by some factor then no change in value of P after row or column scaling.

Property 6: The property is known as antisymmetry under row or column permutation, If the row or column values in the contingency table are swapped then P becomes -P.

Property 7: The property is known as invariance under inversion, If both row and column values in the contingency table are swapped at the same time then no change in value of P.

Property 8: The property is known as invariance under the addition of null record, In this new transactions are added which do not contain any of the itemset under consideration. No change in value of P after null addition.
The above properties(4-8) proposed by Tan et al. [5] which are based upon operations for 2x2 contingency matrix.

Next we present the set of properties followed by the measure ratioPS and ratioLEV in the following table:

Table 5. Properties followed by proposed measures

Measure/Property	P1	P2	P3	P4	P5	P6	P7	P8
ratioPS	Y	Y	Y	Y	Y	N	N	N
ratioLEV	N	Y	Y	Y	Y	N	N	N

Here Y means the property is satisfied while N represents that the property is not satisfied by the particular measure.

5 Testing of New Measures on Sample Dataset

Sample Dataset: Here, we present the interest value of the new measures by using pearson's correlation coefficient measure along with some of the existing measures on sample dataset. Let us consider the following dataset:

Table 6. Sample Dataset For Illustration of Proposed Approach

Tran No	Itemset
1	Beef, Chicken, Milk
2	Beef, Cheese
3	Cheese, Boots
4	Beef, Chicken, Cheese
5	Beef, Chicken, Milk, Cheese, Clothes
6	Chicken, Milk, Clothes
7	Chicken, Milk, Clothes
8	Beef, Chicken

Following is the formula for pearson correlation coefficient:

$$r = \frac{\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left(\sum X^2 - \frac{\sum X^2}{N}\right)\left(\sum Y^2 - \frac{\sum Y^2}{N}\right)}} \tag{5}$$

Here r can take three values and on the basis of these values we can identify whether the two itemset (X and Y) are independent, positively correlated or negatively correlated. If r=0 means both X and Y are independent, if r is greater than 0 means X and Y are positively correlated otherwise X and Y are negatively correlated.

Table 7. Interest values by different measures

Association Rule	Conf.	Lift	PS	CR	AV	ratioPS	ratioLEV
Beef ⇒ Chicken	.66	.88	-.62	-.33	-.083	-.25	0.55
Chicken ⇒ Beef	.66	.88	-.62	-.33	-.083	-.25	0.55
Beef ⇒ Cheese	0.5	1.0	0.0	0.0	0.0	0.0	1.0
Cheese ⇒ Beef	0.75	1.0	0.0	0.0	0.0	0.0	1.0
Chicken ⇒ Milk	0.66	1.33	0.125	0.57	0.16	0.5	2.33
Milk ⇒ Chicken	1.0	1.33	0.125	0.57	0.16	0.5	1.66
Clothes ⇒ Chicken	1.0	1.33	0.093	0.447	0.25	0.15	1.53
Milk ⇒ Clothes	0.75	2.0	0.18	0.77	0.375	.375	3.0
Clothes ⇒ Milk	1.0	2.0	0.18	0.77	0.5	0.3	2.6
{Beef,Milk} ⇒ Chicken	0.5	0.66	-.125	-0.57	-.25	-.25	0.33
{Beef,Cheese} ⇒ Chicken	0.5	0.66	-.125	-0.57	-.25	-.25	0.33
{Chicken,Cheese} ⇒ Beef	0.5	0.66	-.125	-0.57	-.25	-.25	0.33
Milk ⇒ {Chicken,Clothes}	0.75	2.0	0.18	0.77	0.375	0.375	3.0
Clothes ⇒ {Chicken,Milk}	1.0	2.0	0.18	0.77	0.5	0.3	2.6
{Chicken,Milk} ⇒ Clothes	0.75	2.0	0.18	0.77	0.375	0.375	3.0
{Chicken,Clothes} ⇒ Milk	1.0	2.0	0.18	0.77	0.5	0.3	2.6
{Milk,Clothes} ⇒ Chicken	1.0	1.33	0.093	0.447	0.25	0.15	1.53

From the table 7, it can be seen that the proposed measures are equivalent to some of the existing measures, like lift, PS, correlation, which also taken consideration of the negative correlation. As per the given dataset we can see that some rules are negatively correlated like, $beef \Rightarrow chicken$, $chicken \Rightarrow beef$ and $beef, chicken \Rightarrow cheese$ etc., all these negatively correlated rules are not detected by the classical support-confidence measure and thus we infer that the proposed measure are capable to detect all kinds of correlation hence can help in correct decision making.

6 Conclusion

The work proposed in this paper introduces a new measure of interest to extract the association rules. It deals with the limitation of existing support-confidence framework. In this paper we have introduced two measure of interestingness and have seen that both are capable to capture negative correlation among different itemsets. We have seen some properties followed by the proposed measure and correctness of new measures also checked on sample dataset.

References

1. Imielinski, T., Agrawal, R., Swami, A.N.: Mining association rules between sets of items in large databases. In: Proceedings of the ACM SIGMOD Conference on Management of Data, vol. 22, pp. 207–216 (1993)
2. Agrawal, R., Srikant, R.: Fast algorithm for mining association rules. In: Proceeding of 20th International Conference on Very Large Databases, pp. 487–499 (2003)
3. Silberschatz, A., Tuzhilin, A.: What makes pattern interesting in knowledge discovery systems. IEEE Transactions on Knowledge and Data Engineering, 970–974 (1996)
4. Piatetsky-Shapiro, G.: Discovery, analysis and presentation of strong rules. In: Piatetsky-Shapiro, G., Frawley, W. (eds.) Knowledge Discovery in Databases, pp. 229–248 (1991)
5. Kumar, V., Tan, P., Srivastva, J.: Selecting the right interestingness measure for association patterns. In: Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining, pp. 32–41 (2002)
6. Omiecinski, E., Savasere, A., Navathe, S.: An efficient algorithm for mining association in large databases. In: Proceedings of the 21st International Conference on Very Large Databases, pp. 432–444 (1995)
7. Toivonen, H.: Sampling large databases for association rules. VLDB Journal, 134–145 (1996)
8. Ullman, J.D., Brin, S., Motwani, R., Tsur, S.: Dynamic itemset counting and implication rules for market basket data. In: Proceedings of ACM SIGMOD International Conference Management of Data, vol. 8, pp. 255–264 (1997)
9. Pei, J., Han, J., Yin, Y.: Mining frequent patterns without candidate generation. In: Proc. ACM-SIGMOD International Conference on Management of Data, pp. 1–12 (2000)
10. Zhang, C., Wu, X., Zhang, S.: Efficient mining of both positive and negative association rules. ACM Transaction on Information Systems 22, 381–405 (2004)
11. Geng, L., Hamilton, H.J.: Interestingness measures for Data Mining. A survey. ACM Computing Surveys 38 (2006)
12. Vanhoof, K., Brijs, T., Vets, G.: Defining interestingness for association rules. International Journal on Information Theories Applications 10, 370–375 (2010)