Indian Journal of Science and Technology, *Vol 8(15), IPL094, July 2015*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Analysing the quality of Association Rules by Computing an Interestingness Measures

J. Manimaran[1]* and T. Velmurugan[2]

[1]Bharathiar University, Coimbatore - 641 046, Tamil Nadu, India; thavasimaniraj@gmail.com
[2]Research Department of Computer Science, D. G. Vaishnav College, Chennai - 600106, Tamil Nadu, India; velmurugan_dgvc@yahoo.co.in

Abstract

Objective: Association rule mining is one of the data mining process for discovering frequent item set between transaction databases. The main objective of this research work is statistically analyses the quality rules in the apriori algorithm of association rule mining. **Methods:** An Interestingness measures is a subset of statistical method and it can give the solution for splitting interesting rules within huge association rules. Currently, it has shown around hundred and above measures. Specifically, this study is to concentrate on eight measures such as lift, chi-square, hyper-lift, hyper-confidence, conviction, coverage, leverage and cosine. In this analysis is performed in two places of real databases whereas Agriculture and Medical domain. **Findings:** At the experimental results, the proposed system is rectified that the problem many interesting rules are eliminated in satisfying the threshold value of support and confidence. Therefore, the user do not confirm that the strength of interest rules may be least by setting the low threshold value. The comparison and correlation measures also obtained along with the interesting rules. There are some measures outperformed than other and thus measures can mostly correlate with the order lift, chi-squared, hyper-lift, hyper-confidence and conviction. The performance of this work is consistently checking in difference size of transaction databases in addition to we identify the unresolved problem of apriori algorithm. **Conclusion:** Finally, this research concludes that statistical interestingness measures are really helpful for finding interesting rules among large association rules.

Keywords: Apriori Algorithm, Association Rule Mining, Interestingness Measures.

1. Introduction

In DM, Association Rule Mining (ARM) is a well-defined method for discovering interesting relations between variables in large databases. It is commonly used in various domains like financial, meteorological, medical, biology, agriculture etc. Even the databases having different model such as time-series, relational, image, video, audio and so on, ARM is to simply use the transaction database. In supermarket database, the ARM attempts to find groups of items that are mostly occurred together. Usually, Apriori algorithm is one of the widespread functions in the ARM. It needs to settings a two threshold value as support and confidence. As a set of rules (or itemsets) are consider positive if itemsets greater than minimum support and confidence. In consequently, the apriori algorithm possibly defects by generating a lot of association rules (ARs). Among the large collection of association rule, the discovery of interesting association rules has been complicated for making right decision. By using some statistical measure can be effectively deal with the problem of previous statement. As Interestingness Measures (IMs) is a part of statistical measure and it finds an interesting rules along with all ARs. Hence, this work is implementing statistical IMs into huge ARs. Monotonic is a mathematical function which defines to finding the time interval whether increasing, decreasing or equal. Herein, the monotonicity is non-decreasing and anti-monotonicity is non-increasing. The monotonicity is used to prune unnecessary search space efficiently. While setting an anti-monotonicity property into the measure function of apriorialgoritm, the interesting sets or rules can be searched. In this context, statistical measures are not a monotonic property and hence it cannot be used for pruning search space in the strategy of frequency.

*Author for correspondence

Generally, the IMs have divided into two objective and subjective. First, an objective measure is strongly user-independent. It verifies an interesting rule in terms of the pattern structure. In addition to, the objective measure includes such statistical measure as confidence, support, lift, conviction etc. Secondly, subjective measures are employed by domain experts and so it fully dependent upon user-experience. In order to, unexpectedness and actionability are two features of subjective measures. Unexpectedness rules are only useful if there are previously unknown to the user knowledge. For instance a market basket database, the rule bread implies milk may not be useful due to commonly known co-occurring items. Actionability rules are mostly used for direct action that may translate into profitable results. In the impact of subjective measures, different user will be bringing out the different nature of interestingness. So, subjective measures are difficult to determine the interesting rules of different domain. Therefore, this study considers quite well an objective measure. In objective measures, a few IMs will be analyzed enormously like lift, chi-square, hyper-lift, hyper-confidence, conviction, coverage, leverage and cosine. This research discusses about different IMs for identifying strong rules.

ARM is one of the wide-spread concepts in DM. It defines to extracting user interested correlation or finding the frequent sets of items in the transaction database. The usage of ARM declines whilst the user facing the large amount of association rules. So the number of research communities has been concentrated in statistical approach for evaluating the best rules which can be really a challenging and time consuming task. Interestingness measures have an effective way to filter the interesting rule set from the target data set and henceforth this may reduce volume of unwanted rules. Consequently, the related article of interestingness measures denotes shortly in this section. Incremental mining of interesting association patterns approach have discussed with the classical apriori algorithm in order to discover only shocking interesting patterns[1]. In the exemption of support-confidence, it has recommended[2] the rank based system using various Objective Interestingness Measures (OIMs) that perform better than the regular support-confidence OIM. Likewise, the behavior-based clustering of 61 well-known interesting measures was analyzed by using the ranking rules out of 110 datasets[3]. In contrast[4], AROMA (Association Rule Ontology Matching Approach) described an interesting measure in this context of textual taxonomy matching. K. Selvarangam and K. Ramesh kumar have developed a method interesting set of ARs for determining Homogeneity Coefficient (HC). The range of HC varies from 0 to 1. If HC value of a measure close to 1, it leads interesting set of ARs and the knowledge extracted from this set of rules[5]. In the work[6], the authors made an effort to analyze the properties of interestingness measures in taking the rare ARs. For evaluating the IMs in closed itemsets was discussed by Aleksy et al. wherein comparison of IMs performed in artificial datasets without involving experts[7]. As well, it could consider the difference between stability and leverage in the most convenient measure. In the paper [8,9], a comprehensive study of null-invariant IMs were given for mining small probability events. In adding more features to Aprioriaglroithm, the author was created one of the statistical significant algorithm i.e. StatApriori[10]. The research is highly motivates to begin our research in searching quality rules. A potential causal association mining algorithm for screening adverse drug reactions in post-marketing surveillance[11] has suggested using IMs in the field of drug development. In the comparison of IMs[12], the work said that conviction has predicted averagely the best rules out of confidence, lift, chi-square, Laplace, mutual information, cosine, Jaccard and coefficient.

In dissimilarity to IMs, three multiple testing can use to distinguish such good rules amongst false positive rules[13]. Furthermore, IMs have applied in[14] the financial database. An issue of e-commerce application has also described to[15] execute the IMs of ARs. Similar that, u-commerce recommendation system has newly established by using RFM scoring method of mining ARs in[16]. In the dissimulation, evolutionary method of ARs discovery was done on the total ozone content modeling from satellite observations[17]. As granular ARs introduced in the kind of multi-valued data[18] and it could indicate powerfully the negative rules for filtering uninteresting rules. In the same context, a correlation measure was defined and added to the mining algorithm of ARs wherein it classified discretely the interesting rules and uninteresting rules[19]. In[20] medical domain, the novel method of Swarm Intelligence (SI) used to discover the interesting rules. Apriori algorithm has also applied in[21] the clinical decision support system. Moreover, the ARM is widely used in text application by surveying our previous paper[22]. Hiding sensitive ARM[23] decreases the confidence of sensitive rules to below minimum threshold by removing selective item among items of consequent sensitive rules

(R.H.S) for each selective transaction. As far, this section summarized about some relevant articles in IMs of ARs. Subsequently, this study defines clearly a few IMs in the following section.

2. Materials and Methods

In presence of many association rules cannot be consider as which one interest or significant. So, it explicit the interesting rules by using various IMs. This is efficiently combines the apriori algorithm and some IMs for analyzing the best rules. Subsequently, the required methods of apriori algorithm and IMs are illustrated foremost in the section.

2.1 Apriori Algorithm

In ARM, Apriori algorithm is to find the frequent itemsets in large transactional databases. It is using iteratively bottom-up and level-wise search approach. At the first step, it will be scanning the entire database for counting occurrence of each item by means of this can take the large frequent 1- itemsets. Subsequently, it allows a two pass so as to the first pass L_{k-1} (from k-1 itemsets to generate the superset of all frequent k- itemsets) frequent itemsets are used to generate the candidate C_k itesmsets using the candidate generation process. In the second pass, the database scans again and verifies the support of candidates C_k is counted. The discovered rules should be qualified with minimum value of support and confidence. So that, a user specified threshold value of support and confidence is used to discover all rules. These two basic measures are completely different than IMs. As the support and confidence defines moreover in the following statement.

2.2 Basic Measures

Although numerous IMs have been developed in statistics and data mining to assess object relationships, the association rules cannot implicitly extract rules without using these two basic measures. Hence, the support and confidence is one of the essential task in generating strong association rules from the frequent itemsets. An earlier most research, the interesting rules were considered based on a two basic measure such as support and confidence.

2.2.1 Support

The percentage of transactions is to contain both X and Y in the rules of $X \Rightarrow Y$. It is also called mean value. In easy way to understand, support gives the proportion of transactions which contain X and also it counts the amount of transactions presents Y. As it uses an items count of each transaction.

$$Support\,(X \Rightarrow Y) = P\,(X \cup Y) \tag{1}$$

Support is down-ward closure property (anti-monotonicity) that means all sub sets of a frequent set. This property is used to prune the search space in level-wise algorithms. The disadvantage of support measure is left out the rules less than minimum support although there are interesting and potentially valuable rules.

2.2.2 Confidence

The conditional probability of the occurrence of items in X and Y over the occurrences of items in X is called confidence. Confidence is not down-ward closure in contrast to the support uses down-ward closure property for prune the search space. Hence, it defines the below formula,

$$Confidence\,(X \Rightarrow Y) = \frac{Support(\mathrm{X} \cup \mathrm{Y})}{Support(\mathrm{X})}$$

$$= \frac{P(E_X U\, E_Y)}{P(E_X)} = P(E_Y \mid E_X) \tag{2}$$

Whereas X and Y are two disjoint itemsets. It gives a different value for the rules X → Y and Y → X due to vary the support (X) and support (Y). Also, confidence is easily understood an estimate the conditional probability P (EY | Ex), were Ex (EY) is the event that X (Y) occurs in a transaction.

2.3 Interestingness Measures (IMs)

The problem of common threshold is to find all ARs that satisfy a user-specified minimum support and confidence. In high minimum support, there could be generating a few rules alone and spending more time for scanning the database. For setting the low minimum support, it may produce numerous redundant rules. Therefore, the issue is cleared by using some statistical IMs. A statistical measure is nothing but deriving significant mathematical function into DM applications. It should be accurately determine whether rules are interesting or uninteresting.

The architecture of finding interesting rules using IMs is also shown in Figure 1. There have to process a six steps in orderly database selection, transaction format conversion, apriori algorithm implementation, rule extraction,

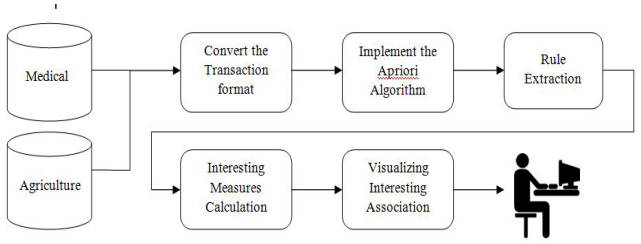

Figure 1. Architecture of finding interesting rules using IMs.

interesting rule calculation and lastly visualize that interesting rules. Herein, the eight important IMs will discuss elaborately likewise lift, chi-square, hyper-lift, hyper-confidence, conviction, coverage, leverage and cosine.

2.3.1 Lift

A lift value is to obtain between 0 and infinity. In lift, the statistical independence can progress in checking how many times more often X and Y occurred together. Occasionally, the threshold level restriction has pruned some rare interest rules but the lift measure can efficiently find out that rare itemsets. So, the disqualified ARs of minimum support (or confidence) using lift value can assess their dependence range. Thus, it is not a downward closure property and also it does not suffer from the rare item problem. The measure lift (also called interest) is defined on rules of the form $X \Longrightarrow Y$ as,

$$Lift\,(X \Rightarrow Y) = \frac{Confidence(X \Rightarrow Y)}{Support(Y)} \tag{3}$$

A lift value greater than 1 indicates that X and Y appear more often together. If the lift value smaller than 1, it denotes that X and Y appear less often together than expected.

2.3.2 Chi-square

Chi-square analysis is a standard technique that allows one to gauge the degree of dependence between the variables A and B. In the field of statistics, $\chi 2$ is widely used method for testing independence or correlation. To computing chi-square measure for the pair of variables of (A, B) requires in constructing two contingency Tables. Technically, $\chi 2$ test is fully based on the comparison of observed frequencies and the corresponding expected frequencies. It is used to test the significance of the derivation from expected values. Let us see the formula,

Table 1. Observed contingency table for (A, B)

	B	$\overline{\mathbf{B}}$
A	n P (A ∩ B)	n P (A ∩ $\overline{B}$)
$\overline{\mathbf{A}}$	n P ($\overline{A}$ ∩ B)	n P (A ∩ $\overline{B}$)

Table 2. Expected contingency table for (A, B)

	B	$\overline{\mathbf{B}}$
A	n P(A) P(B)	n P(A) (1– P(B))
$\overline{\mathbf{A}}$	n (1 – P(A)) P(B)	n (1– P(A)) (1– P(B))

$$\chi^2 = \sum \frac{(f0 - fe)^2}{fe} \tag{4}$$

The observed contingency table for (A, B) has four cells as well as corresponding it to four possible boolean combinations of A, B. Each cell value may be expressed in terms of the total number of samples n, observed frequencies corresponding into the four boolean combinations and it depicts in Table 1. If the variables A and B were statistically independent, chi-square analysis expressed that the observed contingency table should be compared with obtained asymptotically as $n \to \infty$. So that, it showed in Table 2.

2.3.3 Hyper-lift

Hyper-geometric random variables distribution with known parameters can be used to filter random noise. The expected value of random variable C with hyper-geometric distribution is,

$$E(C) = \frac{\kappa\omega}{w + b},$$

Where the parameter k represents the number of trails, w is the number of white balls, and b is the number of black balls. Applying the co-occurrence counts of two itemsets X and Y in a transaction database given as follow,

$$E(C_{XY}) = \frac{C_X C_Y}{m}$$

Where, m considers as the total number of transactions in the database. By using the previous equation, relationship between absolute counts, support and lift can rewritten as,

$$hyper - lift_{\delta}(X \Rightarrow Y) = \frac{C_{XY}}{Q_{\delta}(C_{XY})} \tag{5}$$

Here, the measure interprets the number of times the observed co-occurrence count C_{XY} to be higher than the highest count and expecting at most 99% of the time. $\delta = 0.99$ is compared to the hyper-lift and lift. In a supermarket, the articles offered may have changed or shopping behavior may have changed due to seasonal changes. To address the problem, it can quantify the deviation of the observed co-occurrence count C_{XY} from the independence model dividing by it different location parameter.

2.3.4 Hyper-confidence

The expected value of random variable C with hyper-geometric distribution is, Instead of looking at hyper-geometric distribution to form a lift-like measure, hyper-confidence is to direct calculation the probability of realizing a count smaller than the observed co-occurrences count C_{XY} given the marginal counts Cx and C_Y.

$$hyper\text{-}confidence(X \Rightarrow Y) = P(C_{XY} < C_{XY}) \quad (6)$$

In this case the random variable C_{XY} follows a hyper-geometric distribution with the counts of the itemsets as its parameter. Formally, C_X and C_Y are counting in the r transactions that conditional to two independent itemsets X and Y. Note that hyper-confidence is equivalent to a special case of Fisher's exact test in addition to the one-sided test on 2×2 contingency tables.

2.3.5 Conviction

The implication rules of market basket analysis is based on conviction, which can be a more useful and intuitive measure than confidence and interest. It is normally both the antecedent and the consequent of the rule moreover the statistical notion of correlation. Subsequently, it define as follow,

$$conviction\ (X \rightarrow Y) = \frac{1 - support(Y)}{1 - confidence(X \rightarrow Y)} \quad (7)$$

If the dependency between the actual appearances of X without Y, conviction compares the probability of X appears without Y. It is similar to lift and the lift is only measures co-occurrence but not implication. In contrast, it is a directly measure since it also uses the information of the absence of the consequent. An interesting fact is that conviction is monotone in confidence and lift.

2.3.6 Coverage

Coverage is sometimes called antecedent support. It measures how often a rule $X \rightarrow Y$ is applicable in a database. Therefore it covers LHS support for rules likewise,

$$Coverage(X \rightarrow Y) = supp(X) = P(E_X) \quad (8)$$

2.3.7 Leverage

In leverage measures, the difference of X and Y appearing together in the data set and what would be expected if X and Y where statistically dependent. The rational in a sales setting is to found out how many more units (X and Y item occur together) are sold than expected from the independent sells.

$$leverage(X \rightarrow Y) = supp(X \rightarrow Y) - supp(X)^{*} supp(Y) \quad (9)$$

Using minimum leverage thresholds constraint to 0.01% one first can use an algorithm to find all itemsets with minimum support of 0.01% and then filter the found itemsets. Because of this property leverage also can suffer from the rare item problem.

2.3.8 Cosine

This measure means the geometric between interest factor (*I*) and the support measure, which is a widely used similarity measure for vector –space models. It is used to measure the similarity between LHS and RHS.

$$Cosine = \frac{P(A,B)}{\sqrt{P(A)P(B)}} \quad (10)$$

The closer cosine ($X \Rightarrow Y$) is to 1, the more transactions containing item X also contain item Y. On the contrary, the closer cosine ($X \Rightarrow Y$) is to 0, the more transactions contain item X without containing item Y. This equality shows that transactions are not containing neither item X nor item Y and which have no influence on the result of Cosine ($X \Rightarrow Y$).

3. Experimental Results

The apriori algorithm with some IMs is analyzed in various real world dataset. Consequently, this work described clearly about selecting data repositories and their experimental results. The prerequisites item of experimental is database repositories. In addition to, the comparative study and correlation measures also performed in this section.

Table 3. The details of two data repositories

Database	S.No	Tuple	Transactions	Items
Agriculture	1	12295	227	11
	2	56996	13821	213
	3	1560	224	11
Medical	4	1524	569	261
	5	40906	12967	141

3.1 Data Repositories

Database is one of the main key factors in all kind of research. It may obtain anyone domains. In this work, we are targeted to carry out the transaction database in the field of Agriculture and Medical. The Agriculture source have commonly many food items as rice, paddy, sorghum etc and the database is availed an online in FAO (Food and Agriculture Organization of the United Nations) statistical world food and agriculture. In order to, it divided into different sizes for scaling the performance of apriori algorithm and IMs. As notice that in table 3. As well, the medical database has collected from various hospitals in Tamil Nadu. The patient-wise diagnostics services (Blood and Radiography Tests) are enumerated and which is decomposed into two sizes. The above mentioned databases have used a two standard column attributes as Transaction Identifier (TID) and Item name or Service name.

Hence, the two attributes is only chosen in this work. In table 3, it shows separately the list of databases, number of tuple (that is number of rows), amount of the transaction and distinct items or service names. All of these databases have arranged in the same format as CSV (comma delimited) and so it can directly apply to the apriori algorithm.

3.2 Experimental Results

The practical implementation of our research is to play a vital role in ARM. Apriori algorithm has applied on descried transaction databases of Table 3. Although the apriori algorithm have made many rules in the large transaction databases, there are limited some uninteresting rules. Searching statistically significant association rules is an important because the resulting rules may be spurious[10]. In traditional statistical approach, each interesting measures techniques having discrete future that all explained in section III. Therefore, this research work is discussed the various interestingness measures of apriori algorithm. Clearly see that, Table 4 and 5 are classified the interesting rules and uninteresting rules. The apriori algorithm cannot be executed without the parameter value of support and confidence. Both parameter values adjusted until return the associated rules among the chosen databases. In the prior knowledge, agriculture database-1 sets the parameter values of 0.6/0.1 where the value 0.6 is support and confidence is 0.1. As a result shown in table 4, the total rules have exposed 26 out of 227 transactions and the appreciated primary interesting rules equal to 9.

Unfortunately, it is uninteresting for 17 rules and depicted in Table 5. Among that, we measured an interesting rules by some statistical methods in orderly lift, chi-squared, hyper-lift, hyper-confidence, conviction, coverage, leverage and cosine. In database-2 of Agriculture is assigned to the threshold value 0.01/0.1 that contains less support values compare than testing database-1. And so, the confidence value does not change in the connection of database-1. By setting these threshold values, the result found totally 20 rules of 13821 transactions. It considers the interesting rules of 10 and an uninteresting rule is 10. As agriculture database-3 is to attain the support and confidence value as 0.6/0.1. Those values are same as that database-1. It could be yield out 27 rules among the transactions of 1560. In addition to, the irredundant rules is to sum of 9 while the redundant rules seized from the remaining 18 rules. In medical database, the database-4 gives 1524 tuple and 569 transactions. Threshold value is 0.05/0.1. Meanwhile, association rules are found at 13 in the middle of 1524 transactions. In these rules, there are five rules acceptable and other 8 rules are unacceptable. Finally, database-5 has to increase rapidly the amount of tuple and transactions. In this case, the predefined threshold value is 0.01/0.1. Support value is decreased than database-4, the value of confidence does not changed and resulting rules 11 amongst the transactions of 12967. After analyzed these 11 rules, it can be appreciate only 5 rules and another 6 rules consider as uninterested.

In Table 4 and 5 more are symmetric in the column of support/confidence value. In subsequently, the database-1 and 3 are similar because of support/confidence value range is equal. Database-2, 4, 5 have almost behaved in the same way as point out in last sentence. In deeply, this work discusses about how to work basic threshold and statistical methods of each interesting measures among different size of real world database. As counting the column of support an interesting rules is zero and inversely assumes that all rules are uninteresting. In confidence, the range of

Table 4. The number of interesting rules for support, confidence and IMs

DB No.	Supp/Confi. Val.	Assoc. Rules	Supp.	Confi.	IMs								Interest Rules
					Lift	Chi.	H.Lift	H.Confi.	Convi.	Cov.	Lev.	Cosi.	
1	0.6/0.1	26	0	10	26	18	26	18	26	6	0	2	9
2	0.01/0.1	20	0	0	20	20	0	0	20	0	0	0	10
3	0.6/0.1	27	0	10	27	18	27	18	27	7	0	2	9
4	0.05/0.1	13	0	2	13	10	13	10	13	3	0	0	5
5	0.01/0.1	11	0	0	11	10	11	10	11	1	0	0	5

Table 5. The number of uninteresting rules for support, confidence and IMs

DB No.	Supp/Confi. Val.	Assoc. Rules	Supp.	Confi.	IMs								Interest Rules
					Lift	Chi.	H.Lift	H.Confi.	Convi.	Cov.	Lev.	Cosi.	
1	0.6/0.1	26	26	16	0	8	0	8	0	20	26	24	17
2	0.01/0.1	20	20	20	0	0	20	20	0	20	20	20	10
3	0.6/0.1	27	27	17	0	9	0	9	0	20	27	25	18
4	0.05/0.1	13	13	11	0	3	0	3	0	10	13	13	8
5	0.01/0.1	11	11	11	0	1	0	1	0	10	11	11	6

interesting rules has smaller than range of uninteresting rules. Appending all interesting rules of lift has qualified without missing any rule and hence no one qualified uninteresting rules. The sum of interesting rules on chi-squared has almost larger than uninteresting rules. In except to database-2, the hyper-lift sustains high range in interesting rules as well as uninteresting rules get down in all other database. Summarizing the attribute of hyper-confidence among interesting rules is higher than uninteresting rules except database-2. Conversely, conviction measure is always higher than uninteresting rules in counting the table of interesting rules. Amongst the rules of both interesting and uninteresting, the behavior of coverage, leverage and cosine measures are getting the same result.

4. Discussion

During the previous work, apriori algorithm is exposed many rules. Among these rules, a few interestingness measures have included on seeing the quality rules. Although preceding a lot of difference between each measure, most similar measures of the interesting rules can probably assume by correlating symmetric measure.

Moreover, this discussion describes the merit and demerit points of all measure. Herein, the support value is to define shortly an itemset percentage. It is used to find the high proportion of itemsets. In database-1, maize and potatoes presented together at 174 times out of 227 transactions wherein the support range is very high compare than others. So, this rule has been adopted for setting the minimum support threshold as 0.6. In case, the minimum support threshold will be assigning above 0.8 in this circumstance the support criterion does not seize that rule. If user expected rule exists less than support threshold, it may be affect in eliminating unsatisfied rules. Confidence is also same in lacking some interesting rules.

For example, the limitation of confidence threshold will explain in Table 6. There is same kind of itemset divided into two rules. Note that, wheat and sorghum to be left when minimum confidence value is above 0.67. To address the above mentioned problem of confidence threshold is solved by hyper-confidence. The hyper-confidence value of 1.67 is to stable at rule

Table 6. The limitation of confidence threshold

TID	Items	Rules	Confidence Val.
1	Rice, Maize	Sorghum ⇒ Wheat	1.00
2	Rice, Maize, Sorghum, Beans		
3	Wheat, Sorghum	Wheat ⇒ Sorghum	0.67
4	Wheat, Sorghum		
5	Chick peas, Cabbages		

1 and 2 in Table 6. So, it does not change the value of interchanged itemset. In discrete feature of hyper-confidence is to avoid uninteresting rules in concerning RHS (Right hand side) rule alone. Lift also leaves out the difficulty of confidence. The lift measure inversely decreases to the hyper-confidence feature. In the impact, the unwanted rules shown redundantly. In testing chi-square method, the rule is positive if support rate of LHS itemsets can be greater than or equal to support value of RHS. It also rectifies the issue of confidence. As good as hyper-confidence, it does not consider those rules wherein RHS only present without LHS. The problem of chi-square is, suppose LHS have 2 items and RHS to be a single item. Along with this the chi-square certainly negative whenever as percentage of anyone item in LHS less than RHS. Hyper-lift is more over same as lift. In order to seeing the difference between lift and hyper-lift, there are no one exist in our experiments. Conviction refers to user expected rule from the condition as LHS higher than RHS. Sometimes, conviction rule is not available where as anyone item of LHS smaller than RHS. In this context, the production rules can be very efficient in comparing other measures. The conviction approach has suffered in a single RHS rule. Consequently, medium range of interested itemset may not cover in the case same percentage of LHS and RHS is less by dividing total transaction. Coverage measure does not have any distinct feature.

In most cases, the coverage rate approximately takes one for a single item. At the leverage, it is always return small value along with interested or uninterested itemset. Hence, that is not useful to predict frequent itemset. Finally, Cosine can be used to find similarity between LHS and RHS. In this task is simply checking as LHS is equal to RHS. Henceforth, the least informative rules are only carried out by allowing same weight of LHS and RHS. So far, this work is analyzed to the behaviors of all measure. Besides, this research will discuss about correlating those interesting measures.

In the above mentioned table 4 is to completely construct from the list of Table as 7, 8, 9, 10 and 11. The interesting rules and their individual value of each measure have also referenced in the following tables. In agriculture database, the obtained result enumerates in the Table orderly 7, 8 and 9. Here, the database-1 belongs to 9 interesting rules and which shown in table 7. In Table 8, the database-2 is contained the interesting rules at 10. As well, database-1 and database-3 are symmetrically equal because finding the quality rules from duplicate and induplicate transactions. In comparing database-1 and database-3, database-1 has more duplicate transaction than database-3. As Table 9, the database-3 also to be obtains the same number of interesting rules as 9. Therefore, the apriori algorithm does not affect by any the duplicate transactions. In excepting to agriculture database, the two medical dataset has used to measure the performance of IMs. There are yielding totally 5 interesting rules both databse-4 and database-5. Between these two databases, the support threshold is only changed from 0.05 to 0.01. Hence, the two results of the medical datasets are depicted on table 10 and table 11.

Further, the comparison of each measure is seen in the following graphical representation of all databases. Note that, the basic threshold and IMs annotates on the numerical order of 1-Support, 2-Confidence, 3-Lift, 4-Chi-square, 5-Hyper-Lift, 6-Hyper-Confidence, 7-Conviction, 8-Coverage, 9-Leverage and 10-Cosine.

Table 7. The interesting rules of database-1

Rules	Supp.	Confi.	IMs							
			Lift	Chi.	H.Lift	H.Confi.	Convi.	Cov.	Lev.	Cosi.
Rice, paddy ⇒ Maize	0.66	0.98	1.10	39	1.06	1.00	5.61	0.67	0.0	0.85
Sorghum ⇒ Maize	0.62	0.97	1.10	33	1.05	1.00	5.32	0.63	0.0	0.82
Beans, dry ⇒ Maize	0.66	0.95	1.07	23	1.03	0.99	2.48	0.69	0.0	0.84
Beans, dry ⇒ Potatoes	0.66	0.94	1.10	35	1.06	1.00	2.78	0.69	0.06	0.85
Wheat ⇒ Maize	0.63	0.91	1.02	2.64	0.98	0.91	1.25	0.70	0.01	0.80
Wheat ⇒ Potatoes	0.69	0.99	1.15	80	1.11	1.00	22.41	0.70	0.09	0.89
Maize ⇒ Potatoes	0.76	0.86	1.00	0.08	0.97	0.51	1.01	0.88	0.00	0.87
Beans, dry, Maize ⇒ Potatoes	0.62	0.94	1.10	28.83	1.05	0.99	2.66	0.66	0.05	0.83
Maize, Wheat ⇒ Potatoes	0.63	0.99	1.15	59.58	1.09	1.00	20.44	0.63	0.08	0.85

*Note that measure value is above .90 which also consider as interesting rule.

Table 8. The interesting rules of database-2

Rules	Supp.	Confi.	IMs							
			Lift	Chi.	H.Lift	H.Confi.	Convi.	Cov.	Lev.	Cosi.
Artichokes ⇒ Onions, shallots, green	0.01	0.41	4.71	570	3.44	1.00	1.56	0.02	0.00	0.24
Chick peas ⇒ Cabbages and other brassicas	0.01	0.32	4.88	567	3.51	1.00	1.38	0.03	0.00	0.24
Cloves ⇒ Onions, shallots, green	0.01	0.40	4.55	724	3.45	1.00	1.52	0.04	0.01	0.27
Onions, dry ⇒ Onions, shallots, green	0.01	0.28	3.20	301	2.43	1.00	1.27	0.04	0.00	0.20
Nuts, nes ⇒ Cashew nuts, with shell	0.01	0.17	2.19	128	1.76	1.00	1.11	0.06	0.00	0.16
Taro (cocoyam) ⇒ Onions, shallots, green	0.01	0.18	2.08	106	1.67	1.00	1.11	0.06	0.00	0.15
Raspberries ⇒ Onions, shallots, green	0.01	0.17	1.95	81.15	1.55	1.00	1.10	0.06	0.00	0.14
Beans, dry ⇒ Spinach	0.01	0.12	1.52	30.06	1.24	0.99	1.04	0.08	0.00	0.12
Spinach ⇒ Cashew nuts, with shell	0.01	0.13	1.66	48.28	1.35	1.00	1.06	0.08	0.00	0.13
Cashew nuts, with shell ⇒ Onions, shallots, green	0.01	0.14	1.67	53.79	1.36	1.00	1.06	0.08	0.00	0.14

*Note that measure value is above .90 which also consider as interesting rule.

Table 9. The interesting rules of database-3

Rules	Supp.	Confi.	IMs							
			Lift	Chi.	H.Lift	H.Confi.	Convi.	Cov.	Lev.	Cosi.
Rice, paddy ⇒ Maize	0.65	0.98	1.10	38.42	1.05	1.00	5.58	0.66	0.06	0.85
Sorghum ⇒ Maize	0.62	0.97	1.10	32.75	1.06	1.00	5.31	0.63	0.05	0.82
Beans, dry ⇒ Maize	0.66	0.95	1.07	22.40	1.03	0.99	2.47	0.69	0.04	0.84
Beans, dry ⇒ Potatoes	0.66	0.95	1.10	36.68	1.06	1.00	3.06	0.69	0.06	0.85
Wheat ⇒ Maize	0.64	0.91	1.02	2.86	0.98	0.92	1.25	0.70	0.01	0.81
Wheat ⇒ Potatoes	0.70	0.99	1.15	78.43	1.10	1.00	21.86	0.70	0.09	0.89
Maize ⇒ Potatoes	0.76	0.86	1.00	0.11	0.97	0.53	1.02	0.88	0.00	0.87
Beans, dry, Maize ⇒ Potatoes	0.62	0.95	1.10	30.35	1.06	0.99	2.92	0.66	0.06	0.83
Maize, Wheat ⇒ Potatoes	0.63	0.99	1.15	58.42	1.10	1.00	19.92	0.64	0.08	0.85

*Note that measure value is above .90 which also consider as interesting rule.

Table 10. The interesting rules of database-4

Rules	Supp.	Confi.	IMs							
			Lift	Chi.	H.Lift	H.Confi.	Convi.	Cov.	Lev.	Cosi.
PPBS ⇒ FBS-BSF	0.07	0.97	7.72	312	3.90	1.00	38.43	0.07	0.06	0.76
Lipid Profile ⇒ FBS-BSF	0.05	0.71	5.61	151	2.90	1.00	3.02	0.07	0.04	0.56
Urea ⇒ Creatinine	0.08	0.98	8.85	430	4.54	1.00	45.35	0.08	0.07	0.88
Urine Routine ⇒ FBS-BSF	0.05	0.55	4.37	102	2.38	1.00	1.95	0.09	0.04	0.48
FBS-BSF ⇒ Creatinine	0.05	0.40	3.63	71.40	2.07	1.00	1.48	0.12	0.03	0.43

*Note that measure value is above .90 which also consider as interesting rule.

Table 11. The interesting rules of database-5

Rules	Supp.	Confi.	IMs							
			Lift	Chi.	H.Lift	H.Confi.	Conv.	Cov.	Lev.	Cosi.
MRI Both Knee ⇒ MRI Angio 3T	0.01	0.47	3.68	558	2.87	1.00	1.65	0.03	0.01	0.26
MRI Brain 3T ⇒ MRI Angio 3T	0.01	0.33	2.63	280	2.14	1.00	1.31	0.05	0.01	0.21
MRI Brain (Spectro) ⇒ MRI Angio 3T	0.01	0.25	1.98	111	1.62	1.00	1.17	0.05	0.00	0.16
MRA Carotid & Vertebral ⇒ MRI Brain	0.01	0.27	4.80	696	3.59	1.00	1.30	0.05	0.01	0.27
MRI Both Ankle ⇒ MRI Limb Venogram 3T	0.01	0.19	2.19	141	1.76	1.00	1.12	0.07	0.00	0.17

*Note that measure value is above .90 which also consider as interesting rule.

In Figure 2 of database-1, the chi-square and conviction are outperforming than other measures. Nevertheless, these two measures could not stable in the interesting rules. Consistently, the number of interesting rule values are sustained in same level at lift, hyper-lift, hyper-confidence and cosine. In addition to, the performance of support, coverage and leverage are very low. Also, the IMs of chi-square, lift, hyper-lift, hyper-confidence and conviction are situated high interesting rules in the database-2.

The remaining measure is low range and hence the chances to be miss the interesting rules. It is shown in figure 3. In the simulation of database-1, the same result is obtained in database-3 so that notice at Figure. 4.

In database-4, the high level value of measure can be lift, chi-square, hyper-lift, hyper-confidence and conviction. Inconsistently, the lift, hyper-lift and conviction could not stable that shown in Figure 5. Finally, the Figure 6 of database-5 is shown high level measure as lift, chi-square, hyper-lift, hyper-confidence and conviction. It is keep on same range of all interesting rules likewise chi-square, hyper-confidence and conviction. The lift and hyper-lift could not steady in all stage and the rest of measure support, confidence, coverage, leverage, cosine to be very low.

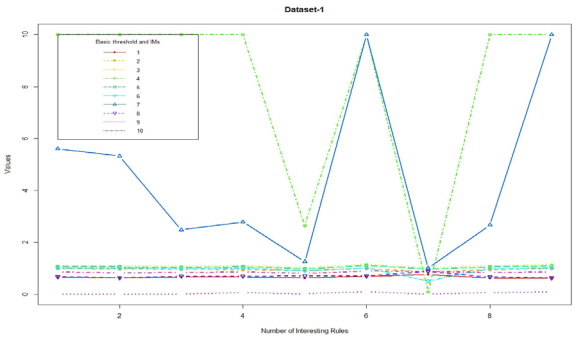

Figure 2. The comparison of IMs in dataset-1.

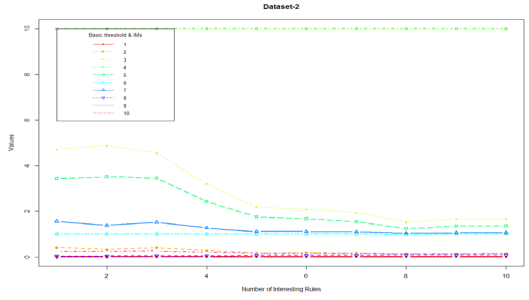

Figure 3. The comparison of IMs in dataset-2

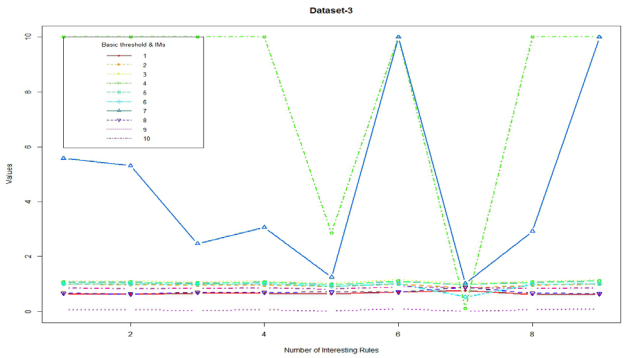

Figure 4. The comparison of IMs in dataset-3.

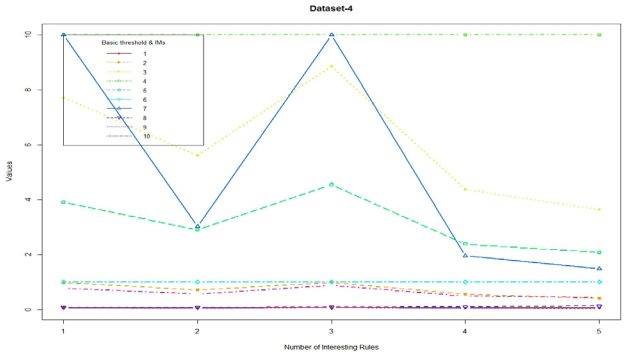

Figure 5. The comparison of IMs in dataset-4

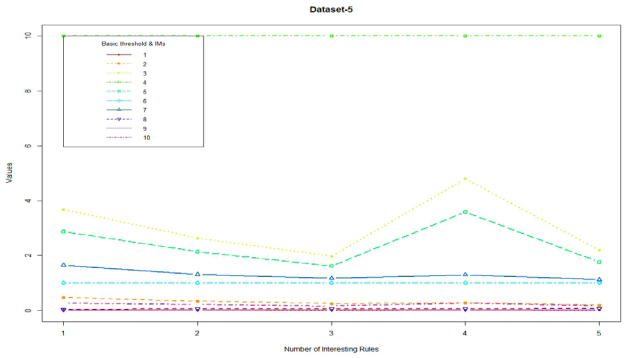

Figure 6. The comparison of IMs in dataset-5.

3.4 Correlation of Interesingness Measures

As correlating one measure to another is specifically takes into the interesting rules. Using this correlation approach, perhaps the simulation measure will be compress as one. Confidence, lift, chi-squared, hyper-lift, hyper-confidence, conviction are sustained an interesting rules in database-1. In this case, cosine is medium and support, coverage, leverage are to hold very less value. In the experimental of database-2, confidence value decreases so as to lift,

chi-squared, hyper-lift, hyper-confidence and conviction to be acceptable. The order of low proportion in interesting rules is support, confidence, coverage, leverage and cosine. In database-3, the confidence is added in the group of standard interesting measures as lift, chi-squared, hyper-lift, hyper-confidence and conviction. As database-1, the cosine is alone middle level. The remaining measures of support, coverage and leverage have to minimal. Due to an inadequacy value of confidence, it terminates in the interesting rules of database-4. Still, the measures of lift, chi-squared, hyper-lift, hyper-confidence and conviction are continuing greater than enough value within interesting rules. As well, the less than threshold measures have been included orderly support, coverage, leverage and cosine. Lastly, the database-5 contains the high value of interesting rules at lift, chi-squared, hyper-lift, hyper-confidence and conviction. Oppositely, support, confidence, coverage, leverage, cosine does not cover certain point on the interesting rules. By the result, the correlation has to present positively in those measures as lift, chi-squared, hyper-lift, hyper-confidence and conviction. In connecting unacceptable measures, the support, confidence, coverage, leverage and cosine consistently yields less value at the interesting rules. This correlation measures is fully based on our experiments and it illustrates the observational point of correlation measure in the next few lines. In the observation, the symmetric measure cannot accurately define on the same value of support and confidence, and also it does not have any co-occurrence between the same size of assembled and unassembled itemset. Herein, the IMs look different result on different transactional databases. As a result, the interesting measure performs whether high or low depend upon the size of each items and transaction.

3.5 Complication of Apriori Algorithm

Even though, the Apriori algorithm plays a central role in this research. There are facing such difficulties in the following scenario as,

- Sometimes, it shows a single item redundantly.
- Duplicate rule generates on interchanged items of same itemset. For example, the result is moreover same where Sorghum ⇒ Maize and Maize ⇒ Sorghum.
- The chance of missing many rules while inaccurately changing the threshold value as high and low. Hence, association rules will be lost if a threshold is set out insufficiently.
- The different user may be use a different minimum support and confidence threshold. As a result, the discovered ARs will change across user and which could not know whether interesting or not.
- As search time is very high whenever setting the low minimum support and confidence.

In the above notified difficulties are motivating to improve the apriori algorithm in the way without constraints based. By this concept, many researchers persistently to be achieving various aspects within association rule of data mining.

4. Conclusion

In ARM, the user faced the problem of assembling unwanted rule where as the apriori algorithm is used to find many association rules. In the problem can be addressing by various interestingness measures approach. Although, a lot of interestingness measures presents in the ARM, this research takes only eight measures for analyzing their distinct feature. In this work, interesting measures has been calculated statistically within the resulting rules of apriori algorithm. Initially, the different transaction database is chosen for finding quality rules in distributed rules of apriori algorithm in addition to it checks the stability of interestingness measures. In the eight interesting measures, the outperforming measures are lift, chi-squared, hyper-lift, hyper-confidence and conviction. The rest of the measures have always decreased an interesting rule. As a result, this research is to identify some interesting correlation measures as lift, chi-squared, hyper-lift, hyper-confidence and conviction.

5. References

1. Yafi E, Al-Hegami A, Alam A. YAMI: Incremental Mining of Interesting Association Patterns. The International Arab Journal of Information Technology. 2012; 9(6):504–10.
2. Joest B, Quix, Anwar. Automated Interestingness measure selection for Exhibition Recommender Systems. Intelligent Information and Database Systems. 2014; 8397:221–31.
3. Tew C, Giraud-Carrier C, Tanner K, Burton S. Behavior-based clustering and analysis of interestingness measures for association rule mining. Data Mining and Knowledge Discovery. 2013; 28(4):1004–45.

4. David J, Guillet F, Gras R, Briand H. Comparison of Interestingness Measures applied to textual taxonomies matching. Revue des Nouvelles Technologies de l'Information; 2008 p. 1–8.
5. Selvarangam K, Kumar RK. Interesting set of Association Rules. International Journal of Fuzzy Mathematical Archive 2015; 6(2):133–8.
6. Surana A, Kiran RU, Reddy PK. Selecting a Right Interestingness Measure for Rare Association Rules. 15th International Conference on Management of Data; 2010. p. 115-124.
7. Buzmakov, Kuznetsov, Napoli On Evaluating Interestingnesss Measures for Closed Itemsets. 7th European Starting AI Research Symposium (STAIRS). 2014; 264:71–80.
8. Wu T, Chen Y, Han J. Assocition Mining in Large Databases: A Re-Examination of Its measures. In knowledge Discovery in Databases: PKDD; 2007. p. 621-628.
9. Chen Y, Han J. Re-examination of Interestingness measures in pattern mining: a unified framework. Data Min Knowl Disc. 2010; 21(3):371–97. doi 10.1007/s10618-009-0161-2.
10. Hamalainen W. StatApriori: an efficient algorithm for searching statistically significant association rules. Knowledge and Information Systems. 2010; 23(3):373–99.
11. Yanqing Ji, et al. A Potential Causal Association Mining Algorithm for Screening Adverse Drug Reactions in Postmarketing Surveillance. IEEE Transactions on Information Technology in Biomedicine;. 2011; 15(3):428–37.
12. Azevedo PJ, Jorge AM. Comparing Rule Measures for Predictive Association Rules. Machine Learning: ECML. Springer Berlin Heidelberg; 2007. p. 510–7.
13. Liu G, Zhang H, Wong L. Controlling False Positives in Association Rule Mining. Proceedings of the VLDB Endowment. 2011; 5(2):145–56.
14. Soldacki P. Discovering interesting rules from financial data. Intelligent Information Sytems; 2002. p. 109–19.
15. Natarajan R, Shekar B. Interestingness of association rules in data mining: Issues relevant to e-commerce. Sadhana. 2005; 30:291–309.
16. Cho YS, Moon SC, Ryu KH. Mining Association Rules using RFM Scoring Method for Personalized u-Commerce Recommendation System in emerging data. Computer Applications for Modeling Simulation and Automobile. 2012; 341:190–8.
17. Martinez-Ballesteros M, Salcedo-Sanz S, Casanova-Mateo C, Camacho JL. Evolutionary association rules for total ozone content modeling from satellite observations. Chemometrics and Intelligent Laboratory Systems; 2011. p. 217–27. doi: 10.1016/j.chemolab.2011.09.011.
18. Min F, Zhu W. Granular association rules for multi-valued data. Electrical and Computer Engineering (CCECE) and 26th Annual IEEE Canadian Conference; 2013. p. 1–5.
19. Piao X, Wang Z, Liu G. Research on Mining Positive and Negative Association Rules Based on Dual Confidence. Fifth International Conference on Internet Computing for Science and Engineering; 2010. p. 102–5.
20. Mangat V. Swarm Intelligence Based Technique for Rule Mining in the Medical Domain. International Journal of Computer Applications, 2010; 4(1):19–24.
21. Nagarajan E, SravaniVS. Knowledge Abstraction from MIMIC II using Apriori Algorithm for Clinical Decision Support System. Indian Journal of Science and Technology. 2015; 8(8):728–30.
22. Manimaran J, Velmurugan T. A survey of Association Rule Mining in Text Applications. IEEE International Conference on Computational Intelligience and Computing Resarch; 2013. p. 698–702.
23. Ramezani A, Dehkordi MN, Safi Esfahani F. Hiding Sensitive Association Rules by Elimination Selective Item among R.H.S Items for each Selective Transaction. Indian Journal of Science and Technology. 2014; 7(6):826–32.