International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161

Available at http://inpressco.com/category/ijcet

Research Article

A Novel Approach to Rank Association Rules Using Genetic Algorithm

Binay Singh[Å*] and Abhijit Mustafi[Å]

[Å]Computer Science Department, BIT Mesra, Ranchi, India.

Accepted 05 April 2014, Available online 15 February 2014, **Vol.4, No.2 (April 2014)**

Abstract

In this paper we propose a new technique to select the top 'n' *association rules out of a pool of 'k' association rules based on heuristic analysis. The proposed method ranks association rules giving emphasis to a larger set of parameters than used by standard methods. The role of correlation has been emphasized in the proposed method which also tries to eliminate issues faced in incorporating correlation, support and confidence meaningfully into one single fitness function. A genetic algorithm model has been developed to establish the rank of the rules taking into consideration the extended set of parameters. The method allows us to establish the best rules in a set of "good" rules and allows for pruning of misleading rules that are often suggested by standard algorithms like the Apriori method.*

Keywords: Association rules, support, confidence, correlation, strong association rules, weak association rules, genetic algorithm, lift, cosine.

1. Introduction

The evolution of Information Technology has witnessed development of following functionalities: data collection and database creation, data management (including data storage and retrieval and database transaction processing), and advanced data analysis (involving data warehousing and data mining) (J. Han *et al*, 2012). Data can now be stored in many different kinds of databases and information repositories like World Wide Web, data warehouses, etc. The large amount of data makes it tricky to extract relevant information from such repositories. Consequently many techniques have been proposed, all of which fall under the field of 'Data mining'.

Data mining refers to extracting or "mining" knowledge from large amounts of data. Data mining is one of the most important tools which extracts manipulates data and establishes a pattern which helps in decision making (A. Sharma *et al*, 2012). The architecture of typical data mining consists of two major components: 1. Database, data warehouse, World Wide Web and other information repository. 2. Database or data warehouse server (J. Han *et al*, 2012).

Data warehouse collects and stores integrated sets of historical data from multiple operational systems and feeds them to one or more data marts. Data warehouse has the following specific properties: subject-oriented, integrated, non-volatile, time-variant, accessible, process-oriented which is primarily useful for organizational decision making The branch of data mining that deals in discovery of interesting associations and correlations between itemsets in transactional and relational databases is called frequent pattern mining. The most important frequent pattern mining application is mining association rules. In 1993, R. Agrawal and R. Srikant first introduced the association rule mining. Association rule mining (ARM) is a very popular and well researched method for discovering relationship between variables in large databases (M. Renuka Devi *et al*, 2012). Association rules are the rules that correlate the presence of one set of items with that of another set of items. It extracts frequent itemsets, interesting rules and discovers the relationship among items in transactional database or in other data repositories (L. Fang *et al*, 2012). ARM generates the best association rules which qualify the minimum support threshold and minimum confidence threshold. Association rule can be used to improve decision making in various areas such as: market basket strategy, process mining, protein sequences, logistic regression, medical diagnosis, bio-medical literature, web search, CRM of credit card business etc. Many researchers have shown that selecting the right objective measures is a very important factor to be considered (P.N. Tan *et al.*, 2004). A. Silberschatz and A.Tuzhilin proposed an approach about the interestingness pattern (A. Silberschatz *et al*, 1996). Many algorithms have been proposed to generate frequent itemsets: Apriori algorithm, Éclat and FP-Growth.

The Apriori algorithm is an iterative level-wise algorithm which is used to find frequent pattern in data (Shweta *et al*, 2013). Improved Apriori algorithm (M. Dhanda *et al*, 2011), (R. Santhi *et al*, 2012), (J. Singh *et al* , 2013) removes the unnecessary transactional records from the database which reduces scan time in large amount and also reduces the redundant generation of sub-items during pruning the candidate set. However, improved mining algorithms performance and its

*Corresponding author: **Binay Singh**

complexity is subject to research area, as they have to deal with the large set of data items.

Recent works involve different usage of correlation measure (Jun-Sese *et al*, 2002), (H.S. Anand *et al*., 2013). Introduction of new measures like Chi square etc (Yong Xu *et al*, 2005). Also, there have been some soft computing approaches using algorithms like genetic algorithm, ant colony optimization, etc (S. Ghosh *et al*, 2010), (Kannika Nirai Vaani M *et al*, 2013), (P. Mandrai *et al*, 2013), (B. Rani *et al*, 2013).

The proposed work assesses the traditional way of frequent pattern mining using Apriori algorithm and introduces the concept of F-measure by using the notion of correlation i.e., association rule is generated by considering three factors, support, confidence and correlation:

A=>B [support, confidence, correlation].

Correlation is calculated by using the "Lift" measure. F-measure is the linear summation of the support, confidence and correlation of each rule with the unknown coefficient α, β, and γ. The values of unknown coefficient are generated by using the Genetic Algorithm. According to F-measure values, best association rules will be generated. Higher the F-measure value, better the association rule will be.

1.1. Motivating Example

Table 1: 2× 2 Transaction summary for purchase of Egg and Butter

	Butter	$\overline{\text{Butter}}$	
Egg	6000	7000	13000
$\overline{\text{Egg}}$	3000	1000	4000
∑	9000	8000	17000

In table 1, the transactions of an item are summarized by their occurrences of Egg and Butter. Egg and Butter is two itemsets. Butter here refers to the transactions which contain butter. $\overline{\text{Butter}}$ refers to the transactions which do not contain butter. Similarly, Egg refers to the transactions which contain egg and $\overline{\text{Egg}}$ refers to the transactions which do not contain egg. As we can see from the table that, the probability of purchasing Egg is P {Egg} = 0.76 and probability of purchasing computer butter is P {Butter} =0.53 and the probability of purchasing both the item P {Egg, Butter} =0.35. The correlation is calculated by the lift:

$$\text{Lift} = \frac{P(X \cup Y)}{P(X)P(Y)}$$

Here the lift of association rule is $P\ (Egg \cup Butter)\ /P\ (Egg) \times P(Butter)$=0.87 which is less than 1 , thus it signifies that both the item are negatively correlated mean purchase of one item decreases the purchase of other item. Support-confidence framework does not give such information about negative correlation between the itemset (J. Han *et al*, 2012).

2. Association Rule Mining

Relationship between the data is called association. Association rule shows attribute value conditions which occur most frequently in the given dataset. In general, Association rules are expressed in the form X→Y, where *X* and *Y* are itemsets (collection of items) representing the antecedent and the consequent part of the rule and both *X* and *Y* do not intersect each other (disjoint), they do not have common items. Association rule may have more than one item in antecedent (*X*) and consequent(*Y*) part. The complexity of rules depends upon the number of items it contains. Association rule mining (ARM) finds interesting associations and correlation among the data in a given dataset (B.Ramasubbareddy *et al*, 2010). Support and confidence are two measures or rule interestingness.

The strength of association rule depends upon following factors:-

1). *Support or prevalence:* - It is simply the number of transactions that contain all the items in the antecedent and consequent parts of rule. Thus, the rule has support S in dataset D, if S% of the transactions in D contains both X and Y i.e. ($X \cup Y$).

$$\text{Supp}\ (X \rightarrow Y) = P\ (X \cup Y) \qquad (2.1)$$

2). *Confidence or predictability:* - It is a ratio of the number of transactions that contain all items in the consequent as well as in the antecedent(namely, support) to the number of transactions that contain all items in antecedent. A rule is said to hold on D, if the confidence of the rule is greater than or equal to confidence threshold. Thus, the rule has confidence C, if C% of the transactions in dataset D that contain *X* also contains *Y*.

$$\text{Conf}\ (X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)} \qquad (2.2)$$

3). *Correlation:* It finds the actual relationship between two or more items whether it is negatively or positively associated. It measures the strength of the implication between *X* and *Y*. It prunes out the large number of negatively associated rules. Thus, actual interesting association rules are generated based upon support, confidence and correlation value.

Some common terms which is mostly used are:

1. Transaction Database: -
   It stores transaction data.
2. Itemset: -
   Set of certain items in the transactions.
3. Frequent-itemset:-
   Itemset that appears frequently in a dataset.
4. Candidate set: - It is the name given to a set of itemsets that is used for testing to meet the certain requirement.

Strong and weak association rules: According to Apriori algorithm the strength of a rule is measured by its support and confidence value. Each rule must qualify user's

specified constraints: - the support of each rule must be greater than or equal to the minimum support threshold (measure statistical significance) and confidence of each rule must be greater than or equal to minimum confidence threshold(measure goodness). The rule which qualifies minimum support and minimum confidence threshold is known as strong association rule otherwise it is weak association rule.

2.1 Association Rule Example

The problem of ARM is given as: Given *a frequent set { Milk, Diaper, Beer}. What association rules have minsup=30% and minconf= 60% ?*

Table 2: Transaction database

TID	Items
1	Jacket, Jeans
2	Jacket, Shirt, Sock, Shawl
3	Jeans, Shirt, Sock, Sweater
4	Jacket, Jeans, Shirt, Sock
5	Jacket, Jeans, Shirt, Sweater

Minimum support threshold= 30% and minimum confidence threshold= 60%

$$S= \frac{\sigma(\text{Jeans, Shirt, Sock})}{T} = \frac{2}{5} = 0.40$$

$$C= \frac{\sigma(\text{Jeans, Shirt, Sock})}{\sigma(\text{Jeans, Shirt})} = \frac{2}{3} = 0.67$$

Rule : { Jeans, Shirt }→Sock

Here, S represent for Support, C represent for Confidence and T represent for transactions. We have calculated support and confidence value from frequent itemsets. Table 2 shows a transactional database in which each transaction is a non-empty itemset. Each transaction is associated with an identifier known as transaction identifier (TID). {Jacket, Jeans, Shirt, Sock, Shawl, and Sweater} is an itemset present in transactions. Support and confidence value is calculated by using the equation 2.1 and 2.2. As the support and confidence of rule: {Jeans, Shirt} → {Sock}, support (40%) confidence = (67%) is greater than given minimum support threshold (30%) and minimum confidence threshold (60%). Thus, it is a strong association rule.
Discovering of all association rules can be viewed as two-step process (R. Agrawal *et al*, 1993):
1) Finds the frequent itemsets.
2) Use the frequent itemsets to generate the strong association rules.

2.2 Drawbacks of Association rules

Many researchers have given the drawbacks of association rules in their paper (E. Garcia *et al*, 2007):-
1).*Discovering too many association rules:* The traditional association rules mining (ARM) algorithms were very simple and efficient. However, ARM algorithms generate a large number of association rules and it does not give the actual information that the rules generated are relevant or not.
2).*Strong rules generated can be misleading and uninteresting:* The traditional ARM algorithm is based upon a support-confidence framework. A large number of association rules is generated by using low support thresholds. Although minimum support and minimum confidence threshold helps to prune out a good number of rules, many rules found are still not interesting to the users. This truly happens when mining for long patterns or when mining at low support thresholds.
3).*Does not considers effect of correlation*: The traditional ARM algorithm does not measure the strength of the correlation and implication between *X* and *Y*. It does not give any information about negative association among items which leads to unwise decisions based on rules.
In this paper, we have used three parameters:- support, confidence and correlation in order to remove the drawbacks of association rules to a large extent. It also gives negative correlation which is not identified by the traditional ARM following support-confidence framework.

3. Apriori Algorithm

It is also known as level-wise algorithm. It was introduced by R.Agrawal and R.Srikant in 1994 (R. Agrawal *et al*, 1994). It is the most popular algorithm for mining frequent itemsets for Boolean association rules. Apriori consists of two important steps: the first step is to find the frequent itemsets among the given number of transactions, and second step is to extract the rules from the mined frequent itemsets. It requires the prior knowledge of frequent itemsets. It uses the downward closure property. Apriori algorithm uses the bottom-up search method, moving towards upward level-wise in the lattice. Before reading the database at every level, it prunes out the infrequent sets. If there is any itemset which is infrequent, then its superset should not be tested /generated.

Method

1. Initially scan the database DB to accumulate the count for each item and retain those that satisfy minimum support, to generate frequent 1-itemset.
2. Frequent *k-itemsets* is used to generate *(k+1) candidate itemsets*.
3. Test the candidates against DB.
4. Terminate when no candidate set can be generated or it is unlikely to be frequent (fails to meet the minimum support threshold).

Apriori property

All non empty subsets of frequent itemsets must also be frequent (J. Han *et al*, 2012). It is an *anti-monotone* property: if a set cannot pass a test, then all its supersets will fail the same test as well. An itemset I is not frequent, if it fails to meet the minimum support threshold
Apriori algorithm viewed as a two-step process to find frequent itemsets:

1. *Self-Join*: 1-frequent sets join with itself to generate 2-itemsets. Apriori employs an iterative approach, in which *k-itemsets* are used to explore *(k+1) itemsets*. However, itemsets can be joinable only if there is at least a one common item.
2. *Prune:* In this, the itemsets generated from *Join, is pruned* out which fails to qualify the minimum support threshold. An itemset *I*, which qualify the minimum support threshold, is known as *Frequent Itemset.*
In this paper, Apriori algorithm is used to generate association rules using the support and confidence threshold.

4. Correlation measures

Many researchers have proposed a different pattern evaluation measures: Lift, Cosine, Chi-square, Max_confidence, All_confidence, and Kulczynski. In these six pattern evaluation measures , four measures (Cosine, Max_confidence, All_confidence, Kulczynski) values are influenced by the supports of X, Y and X ∪ Y or it is more likely, by the conditional probabilities of *P(X/Y)* and *P(Y/X)* but not by the total number of transactions.
For a given itemset *X* and *Y* correlation measures used:
1)*Lift* : It is the simple correlation measure of how much better the rule is doing . If $P(X \cup Y) = P(X).P(Y)$, then the occurrence of an itemset *X* is independent of the occurrence of an itemset *Y*, else they are correlated or dependent. The Lift value can be computed by:

$$\text{Lift}(X, Y) = \frac{P(X \cup Y)}{P(X)P(Y)} \quad (4.1)$$

If the resulting value is less than1, then *X* and *Y* are *negatively correlated*. If it is greater than 1, then *X* and *Y* are *positively correlated*, meaning that the occurrence of one will implies the occurrence of the other and if resulting value is equal to 1 , then *X* and *Y* are independent mean there is no correlation between them. It is also referred as the lift of association rule $X \rightarrow Y$, as it tend to lift the occurrence of an item with the other items.
2).*Chi-Square***:** The squared difference between the observed value snd expected value of each slot in the contigency table is required in order to compute the Chi-Square($\chi 2$):

$$\chi 2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} \quad (4.2)$$

3).*Cosine***:** It is a harmonized lift measure. Lift and cosine are very similar to each other, except that in cosine, square root is taken upon the product of probabilities of *X* and *Y*. Thus, because of square root the cosine value is only influenced by the number of transactions which contain X, *Y*, and *X* ∪*Y*, and not by the total number of transactions.

$$\text{Cosine}(X, Y) = \frac{P(X \cup Y)}{\sqrt{P(X)P(Y)}} \quad (4.3)$$

The other correlation measures in practice are: All_confidence, Max_confidence and Kulczynski. The proposed algorithm uses the lift and cosine correlation measure.

5. Genetic Algorithm

Genetic algorithms (GA) are search algorithms based on natural selection and natural genetics. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems. It simulates the survival of the fittest among individuals over consecutive generations for solving a problem. After an initial population is randomly generated, the algorithm evolves through three operators: - selection, crossover and mutation.
Some common terms:
1). *Chromosome:* It is also sometimes called as Genome. It is a set of parameters which define a proposed solution to the problem which GA is trying to solve. It is represented as a simple string.
2). *Gene:* It is a part of chromosome. It contains a part of solutions. For example if 82596 is a chromosome, then 8, 2, 5, 9, 6 are its gene.
3). *Fitness*: It is a central idea in evolutionary theory. It describes the ability to both survive and reproduce, and is equal to the average contribution to the gene pool of the next generation that is made by an average individual of the specified genotype or phenotype. If differences between alleles at a given gene affect fitness, then the frequencies of the alleles will change over generations
Genetic algorithms differ from traditional search and optimization methods in four significant points:
1). Genetic algorithms search parallel from a population of points. Therefore, it has the ability to avoid being trapped in local optimal solution like traditional methods, which search from a single point.
2). Genetic algorithms use probabilistic selection rules, not deterministic ones.
3).Genetic algorithms work on the Chromosome, which is encoded version of potential solutions' parameters, rather the parameters themselves.
4).Genetic algorithms use fitness score, which is obtained from objective functions, without other derivative or auxiliary information.

Method:
Step 1: Identify the genes which contribute the chromosomes (feasible solutions).
Step 2: Start with an initial population of 'p' chromosomes.
Step 3: Repeat step 4 to 7.
Step 4: Evaluate the p-chromosomes based on F-measure and its constraints.
Step 5: Select 'n' best chromosomes.
Step 6: Perform 'crossover' and 'mutation' on n chromosome.
Step 7: Generate next generation of 'p' chromosome using the chromosome from step 5 and step 6.
Step 8: Until < termination condition >

In the proposed work, 'p' is taken as 1000. Also, the termination condition is 1000 generation. Genetic

algorithm is used to get the value of unknown coefficients α, β and γ in order to calculate F-measure. Chromosomes are chosen consisting of two genes- α and β. Evaluation of chromosomes chooses the best possible value of α, β and γ (=1- α -β. See the constraint 1 below) which meets the criteria used in the F-measure. Constraint used in F-measure:

1). $\alpha + \beta + \gamma = 1$

2). $\gamma > \alpha > \beta$

α_i	β_i

Fig. 1 Chromosome

Here in figure 1, it depicts that chromosome is made up of genes (α, β) and γ value can be calculated by 1-(α+β).

6. Proposed Algorithm

6.1Block Diagram

Method employed in this paper:-

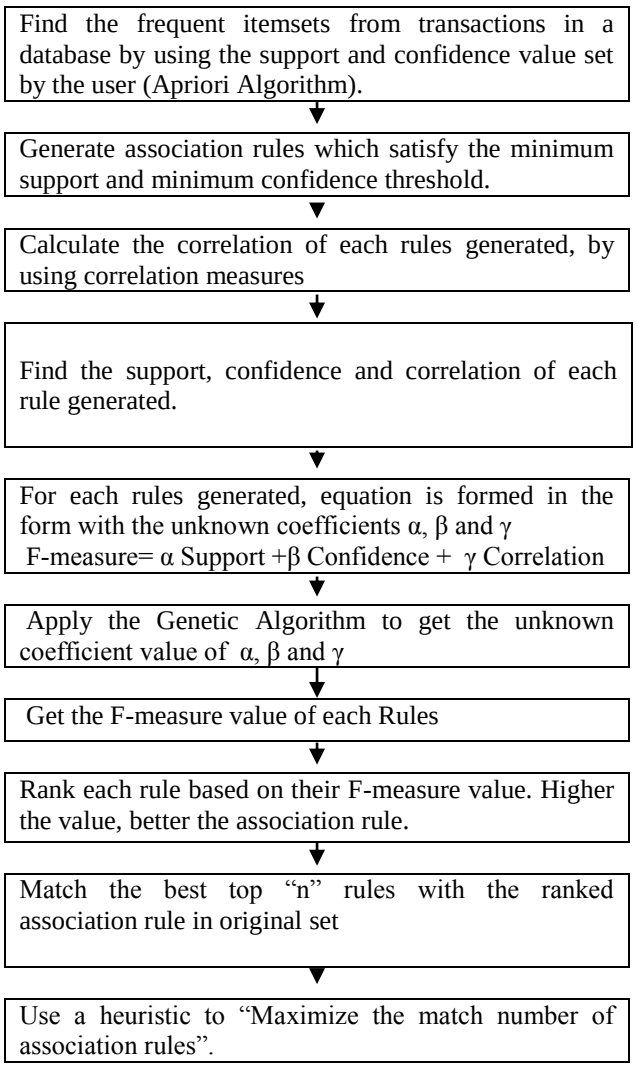

Fig.2 Block Diagram of proposed method

6.2 Algorithm

1) Generate Association Rules with the minimum support and minimum confidence.

2) For each rule generated:

Find the support, confidence and correlation of each rule.

Find F-measure =α support + β confidence + γ correlation.

3) apply the Genetic Algorithm to get the unknown coefficient value α, β and γ.

4) Rank the association rules according to the F-measure value, higher the value better is the association rule.

5) Match the best top "n" association rule generated by using support, confidence and correlation (F-measure) with the association rule generated by the support-confidence.

Step 6: Use a heuristic to "Maximize the match number of association rule".

It gives best association rules, using three parameters: support, confidence and correlation. Thus, each rule has X->Y {support, confidence and correlation}.

6.3 Generate Association Rules

In this Apriori Algorithm has been considered for generating association rules (R. Agrawal *et al*, 1993). Association rule generated upon transactions in database can be considered as a two step process:

1) Find all sets of items (Itemsets) whose support satisfies the minimum support threshold set by the user. These itemsets are known as frequent itemsets.

2) Generate the association rules by using these frequent itemsets.

6.4 Linear summation of support, confidence and correlation with the unknown coefficients

α, β and γ are the unknown coefficients which have been used in order to get the F-measure value of each rule. Each Rule A => B has: {Support, Confidence, and Correlation} By using these three parameters with the unknown coefficient, equation is formed:

F-measure =α Support+ β Confidence + γ Correlation (6.1)

By integrating the correlation with support confidence, it generates the best and interesting rules. F-measure value of each rule is sorted in descending order in terms of higher value to lower value. Compare the unsorted F-measure of rule with the sorted F-measure. Sorted F-measure gives the optimal association rules.

6.5 Calculating Correlation by using the Lift and Cosine measure

Lift and cosine are the most important correlation measures among all the correlation measures (including): All_confidence, Max_confidence, and Kulczynski. After the strong association rules generated, calculate the correlation between associated items, by using one of the correlation measures such as Lift. The Lift between the occurrence of A and B can be computed by: *Lift (X, Y) = P*

(X ∪ Y) / P (X).P (Y) which is equivalent to P (X|Y)/P(Y).Lift in terms of support and confidence:

$$\text{Lift (X, Y)} = \frac{\text{Confidence}(X \rightarrow Y)}{\text{Support}(Y)} \quad (6.2)$$

It is also referred as the lift of association rule, as it tend to lift the occurrence of an item with other items.

The cosine measure is similar to lift except that in cosine; square root is used upon the product of probabilities X and Y. The cosine measure is computed by

$$\text{Cosine (X, Y)} = \frac{\text{Support}(X \cup Y)}{\sqrt{\text{Support}(X)\text{Support}(Y)}} \quad (6.3)$$

Lift and Cosine, calculated value is shown in the result.

6.6 Genetic Algorithm Specification

By applying the Genetic Algorithm (GA), the optimum values of α, β and γ are obtained, which is used to calculate the F-measure.

6.7 Heuristic Employed

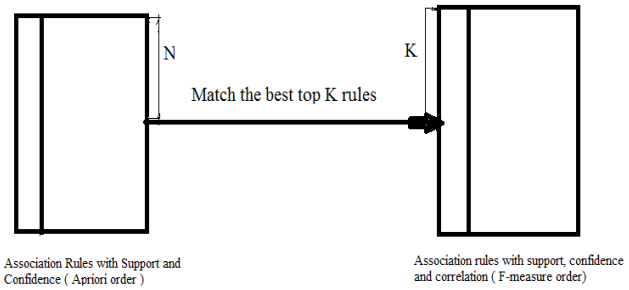

Fig. 3 Match top 'N' association rule with top 'K' Association rules in order to maximize the match number

In Fig. 3, Left side shows: Association Rules with support and confidence (Apriori order, Old Sequence). Right side shows: Association Rules with support, confidence and correlation (F-measure order, New Sequence).

To find the appropriate value of α, β and γ (the coefficients), the genetic-algorithm's optimizing heuristic is as follows:

"*Generate a new sequence of rules based on the F-measure values with the aim of maximizing the total number of matches out of top K rules between the old and new sequence, subject to the constraints mentioned before".*

The corresponding values of α, β and γ are the ones used in our further steps, i.e. the 'appropriate values'.

7. Results and Discussion

The objective of this paper is to discover the interesting association rules by considering all the three parameters:- support, confidence and correlation. Thus, all three of them contribute to the result F-measure. We attach three coefficients α, β, and γ to these three parameters, which as a weight for the individual parameters contribution to the value of F-measure. Thus, it prunes out the week association rule which tends to creep into the top n association rules

Apriori algorithm is used to generate association rules on the basis of support-confidence framework. Support = 20% and confidence = 30% is used, thus all the association rules which qualify these two threshold, will be generated.

Table 3 shows the top 20 Association rules in confidence order (Apriori order) for Supermarket example dataset from WEKA 3.6

1	milk-cream=t fruit=t 2038 ==> bread and cake=t 1684 conf:(0.83)
2	milk-cream=t vegetables=t 2025 ==> bread and cake=t 1658 conf:(0.82)
3	fruit=t vegetables=t 2207 ==> bread and cake=t 1791 conf:(0.81)
4	margarine=t 2288 ==> bread and cake=t 1831 conf:(0.8)
5	biscuits=t 2605 ==> bread and cake=t 2083 conf:(0.8)
6	milk-cream=t 2939 ==> bread and cake=t 2337 conf:(0.8)
7	tissues-paper prd=t 2247 ==> bread and cake=t 1776 conf:(0.79)
8	fruit=t 2962 ==> bread and cake=t 2325 conf:(0.78)
9	baking needs=t 2795 ==> bread and cake=t 2191 conf:(0.78)
10	frozen foods=t 2717 ==> bread and cake=t 2129 conf:(0.78)
11	bread and cake=t vegetables=t 2298 ==> fruit=t 1791 conf:(0.78
12	sauces-gravy-pkle=t 2201 ==> bread and cake=t 1710 conf:(0.78)
13	vegetables=t 2961 ==> bread and cake=t 2298 conf:(0.78)
14	party snack foods=t 2330 ==> bread and cake=t 1808 conf:(0.78)
15	bread and cake=t fruit=t 2325 ==> vegetables=t 1791 conf:(0.77
16	juice-sat-cord-ms=t 2463 ==> bread and cake=t 1869 conf:(0.76)
17	vegetables=t 2961 ==> fruit=t 2207 conf:(0.75)
18	fruit=t 2962 ==> vegetables=t 2207 conf:(0.75)
19	bread and cake=t fruit=t 2325 ==> milk-cream=t 1684 conf:(0.72)
20	bread and cake=t vegetables=t 2298 ==> milk-cream=t 1658 conf:(0.72)

Table 4 shows the top 20 Association rules in F-measure order (considering support, confidence and correlation) for supermarket dataset from WEKA 3.6

1	milk-cream=t fruit=t 2038-->bread and cake=t 1684 Corr: 0.21
2	milk-cream=t vegetables=t 2025-->bread and cake=t 1658 `Corr:0.23
3	fruit=t vegetables=t 2207-->bread and cake=t 1791 Corr: 0.0
4	margarine=t 2288-->bread and cake=t 1831 Corr: -0.1
5	biscuits=t 2605-->bread and cake=t 2083 Corr: -0.4
6	milk-cream=t 2939-->bread and cake=t 2337 Corr: -0.66
7	tissues-paper prd=t 2247-->bread and cake=t 1776 Corr: -0.05
8	fruit=t 2962-->bread and cake=t 2325 Corr: -0.67
9	baking needs=t 2795-->bread and cake=t 2191 Corr: -0.56
12	sauces-gravy-pkle=t 2201-->bread and cake=t 1710 Corr: 0.0
10	frozen foods=t 2717-->bread and cake=t 2129 Corr: -0.5
11	bread and cake=t vegetables=t 2298-->fruit=t 1791 Corr: -0.11
13	vegetables=t 2961-->bread and cake=t 2298 Corr: -0.67
14	party snack foods=t 2330-->bread and cake=t 1808 Corr: -0.14
15	bread and cake=t fruit=t 2325-->vegetables=t 1791 Corr: -

	0.14
16	juice-sat-cord-ms=t 2463-->bread and cake=t 1869 Corr: -0.28
17	vegetables=t 2961-->fruit=t 2207 Corr: -0.67
18	fruit=t 2962-->vegetables=t 2207 Corr: -0.67
19	bread and cake=t fruit=t 2325-->milk-cream=t 1684 Corr: -0.14
20	bread and cake=t vegetables=t 2298-->milk-cream=t 1658 Corr:-0.11

From table 3 and table 4, it clearly seen that change occurs at rule 10, 11, and 12.

Support, confidence and correlation of each association rules calculated. Support, confidence and correlation value of top 10 rules is shown in the table 5. The example dataset is taken from the well known data mining tool, WEKA 3.6. Lift correlation measure is selected for all the dataset.

Table 5 Support Confidence Correlation value of top 10 rules among n rules (supermarket dataset)

	Support	Confidence	Correlation
1	0.363950724011	0.826300294406	0.212953876349
2	0.358331532313	0.818765432098	0.231172839506
3	0.387075859087	0.811508835523	-0.004361123697
4	0.395720769397	0.800262237762	-0.097137237762
5	0.450183704344	0.799616122840	-0.404750479846
6	0.505078884806	0.795168424634	-0.657068730860
7	0.383834017722	0.790387182910	-0.051012461059
8	0.502485411713	0.784942606347	-0.672349763673
9	0.473524962178	0.783899821109	-0.555679785330
10	0.460125351199	0.783584836216	-0.496273463378

All three values with the unknown coefficients α, β and γ in linear equation, gives the F-measure value (Equation 6.1). JGAP, a popular GA package in JAVA is used to implement the genetic algorithm to get the value of three unknown coefficients

Lift and cosine correlation value from equation (6.2) and (6.3) can be seen in the table 6:

Table 6 Lift and cosine value for top 10 rules (supermarket dataset)

Index	Lift	Cosine
1	0.21295387634936214	0.909010612922798
2	0.2311728395061723	0.9048565809556592
3	-0.004361123697326774	0.9008378519596825
4	-0.0971372377622376	0.8945737743541545
5	-0.4047504798464491	0.8942125713949066
6	-0.6570687308608372	0.8917221678495099
7	-0.051012461059189995	0.889037222455026
8	-0.6723497636731937	0.8859698676292908
9	-0.5556797853309481	0.885381172777648
10	-0.4962734633787266	0.8852032739526076

From table 6, we can view the order of lift and cosine value. Cosine value is in descending order from higher to lower which is similar to a confidence order. Thus, it gives 100% match when we compare with the confidence order (association rule from Apriori algorithm). So here in our algorithm, we have used Lift measure which lifts the occurrence of one itemset with the other itemset by their values.

F-measure unsorted and sorted values can be seen in the table 7 which shows the F-measure values of each association rules. Unsorted values is in the order of association rule (the confidence order), which we have sorted, to rank the association rule with their F-measure values.

Table7 F-measure unsorted and sorted value of top 20 association rule (supermarket dataset)

Index	Unsorted	Sorted
1	0.5776133248680343	0.5776133248680343
2	0.573079791397666	0.573079791397666
3	0.5617495404376078	0.5617495404376078
4	0.5520390481703926	0.5520390481703926
5	0.548763379095723	0.548763379095723
6	0.5485042997570542	0.5485042997570542
7	0.5460633898773722	0.5460633898773722
8	0.541094818520943	0.541094818520943
9	0.5381751537457886	0.5381751537457886
10	0.5373669143595539	0.537968259781111
11	0.5371964038684165	0.5373669143595539
12	0.537968259781111	0.5371964038684165
13	0.5343325212232992	0.5343325212232992
14	0.5341545239330652	0.5341545239330652
15	0.5302811404629029	0.5302811404629029
16	0.5199857583053356	0.5199857583053356
17	0.510914690299641	0.510914690299641
18	0.5107354305554193	0.5107354305554193
19	0.4979036566894753	0.4979036566894753
20	0.4966222572412611	0.4966222572412611

As we can see, that out of 20 association rules, total number of matches is 17, thus percentage match in top 20 rules is 85%. 3 of the association rules (rule 10, 11, 12) are not in the order as it was in previous because rules having F-measure value higher is shifted up (ranked). Higher the F-measure value, better the association rule will be

The analysis is done for each dataset and the results can be seen in the graphs. The proposed algorithm efficiently gives the actual best association rule of the dataset, depending upon F-measure value. The same process was employed for five different examples giving the following results:

Case1: supermarket
Case 2: weather.nominal
Case 3: vote
Case 4: breast-cancer

Case 5: contact-lenses

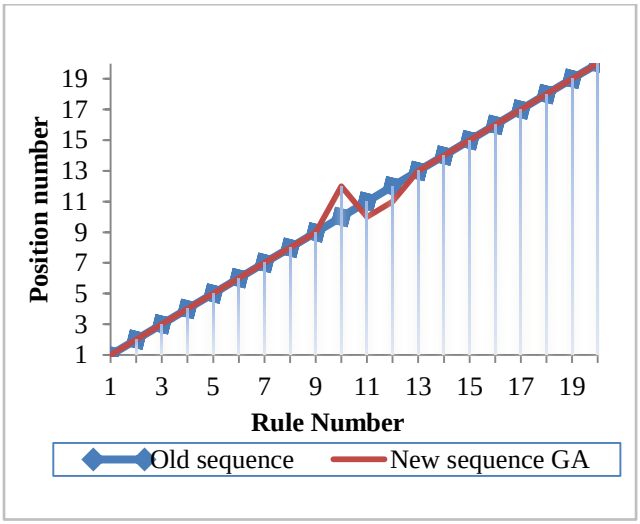

Fig. 4(a) Case 1: supermarket dataset

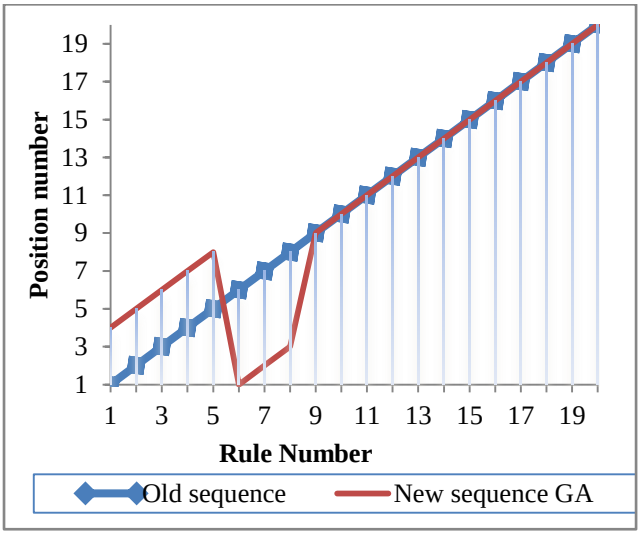

Fig. 4(b) Case 2: weather.nominal dataset

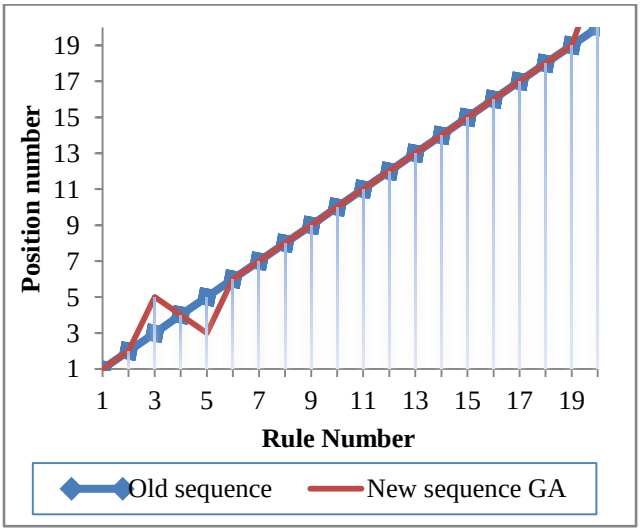

Fig. 4(c) Case 3: vote dataset

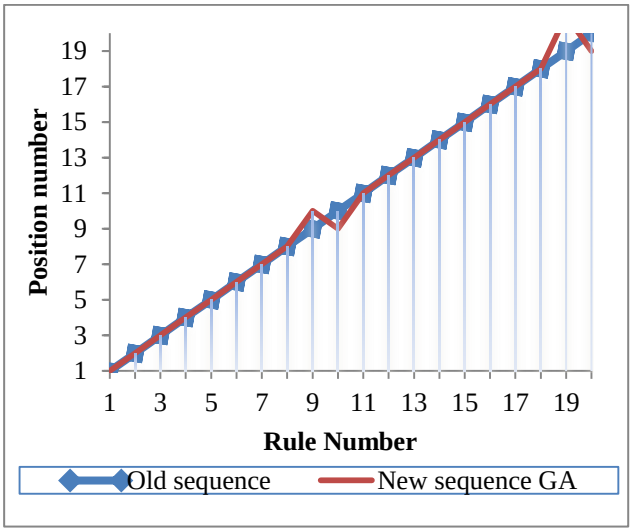

Fig. 4(d) Case 4: breast-cancer datasets.

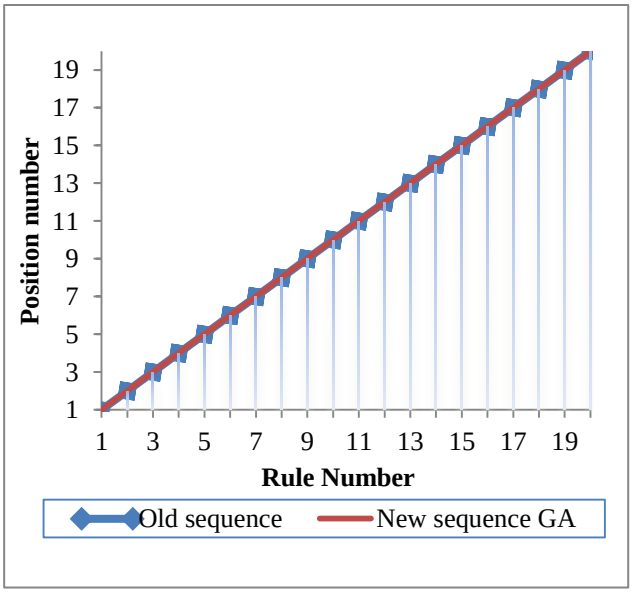

Fig 4(e) Case 5: contact-lenses dataset

Fig. 4(a) to 4(e) shows the comparison of order (rank) of top 20 association rules, according to their F-measure values (new sequence) vs confidence (Apriori or old sequence). As is evident from the figure 4(a), the change in order from old to new sequence occurs at association rule number 10, 11 and 12. The 10^{th} rule shifts to 11^{th} position, pushing the 11^{th} rule to 12^{th} position while the 12^{th} rule is promoted up to 10^{th} position.

In fig 4(b), rules 1, 2 and 3 are collectively demoted to a position after rule 8, automatically shifting the collection of rules 4, 5, 6, 7 and 8 to the positions 1 to 5. In the set of rules {1, 2, 3} and {4, 5, 6, 7, 8}, the relative positions remain intact. In Fig. 4(c), the change in order from old to new sequence occurs at association rule number 3 and 5. Association rule 5^{th} moves upward to 3^{rd} position and 3^{rd} rule is shifted to 5^{th} position. In Fig 4(d) example the change occurs at positions 9, 10, 19 and 20. 10^{th} rule is promoted to 9^{th} position and 21^{st} rule is promoted to 19^{th} position. While 9^{th} rule is pushed to 10^{th}

position and 19th rule is shifted to 20th position. In Fig 4(e) example, it is evident from the figure that no change occurs in order from old to new sequence. So, in this particular example, confidence and support measures were enough to generate the right order of rules. All the examples when matched with their respective support counts and confidence values, clearly show that for some rules the high support count unnecessarily makes them more interesting rules as per support and confidence measures. The effect of correlation in the F-measure thus tends to change that order accordingly.

Figure 5(a) and 5(b) shows the comparison of the percentage matches of new sequence with old sequence in case of lift and cosine measures for different dataset cases. It clearly depicts the advantage of lift measure over cosine measure, as the cosine measure doesn't change the old sequence at all.

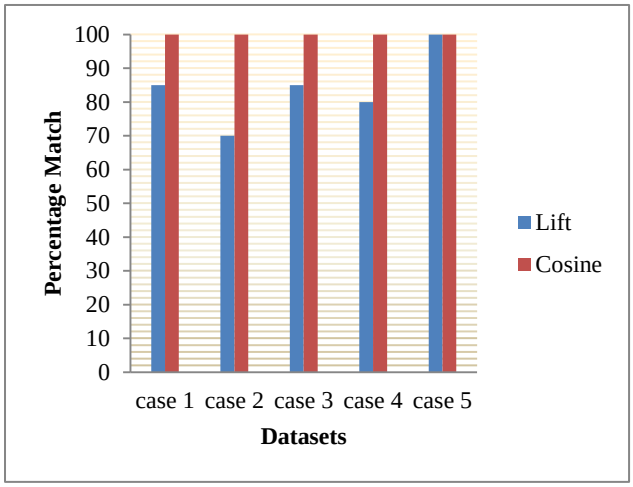

Fig. 5(a): Percentage matches of F-measure order with confidence order (For top 20 rules.)

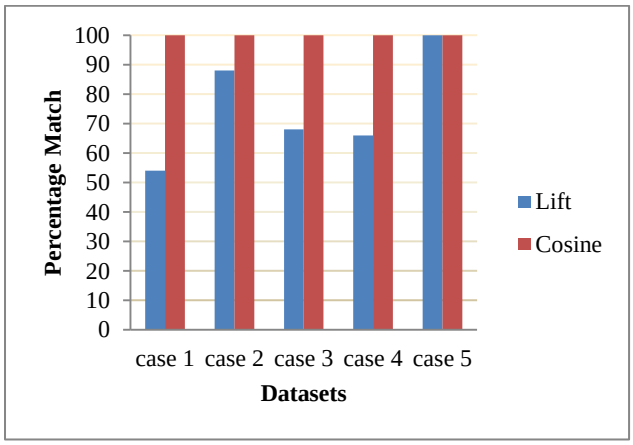

Fig. 5(b): Percentage matches of F-measure order with confidence order (For top 50 rules.)

Thus, it is evident that the effect of correlation changes the order of the association rules obtained from Apriori algorithm, giving us an order where an actually interesting rule gets a better rank than the one padded up by support count.

8. Conclusion and Future Scope

This paper presented the heuristics to rank the association rules by considering three parameters: support, confidence and correlation. This proposed method will generate a best association rules as it can weed out the relatively weaker association rules and the actual best association rules will be easily noticed and identified in the original dataset. Therefore, for those databases which contain large numbers of transactions, our algorithm can efficiently give the actual best association rule of the database. It is very useful for the market strategies, such as in the supermarket example the sales manager can recommend the relevant related products to the customers.

Any field in which association rules are required will benefit from this methodology. Like: - Business Solutions, Industrial Solutions, and in any other case where we want to make a better choice.

References

A. Sharma, N. Tivari, (Aug-2012), A Survey of Association Rule Mining Using Genetic Algorithm, *International Journal of Computer Applications and Information Technology*, Vol. 1, Issue-2, ISSN: 2278-7720, pp.1-8.

R. Agrawal, R. Srikant, (Sep-1994), Fast algorithms for mining association rules. In Proceedings of the *20th International Conference on Very Large Data Bases*, Santiago, Chile, pp.487-499.

R. Agrawal, T. Imielinski, A. N. Swami, (May-1993), Mining association rules between sets of items in large databases. In Proceeding of the ACM SIGMOD *International Conference on Management of Data*, Washington D.C, pp. 207-216.

B. Ramasubbareddy, A.Govardhan, A. Ramamohanreddy, (Nov-2010), Mining Positive and Negative Association Rules, *International Journal of Recent Trends in Engineering and Technology*, Vol.4, pp.151-155.

P. N. Tan, V. Kumar, and J. Srivastava (2002), Selecting the right interestingness measure for association patterns, *Information Systems*, Vol.29, pp.293-313.

L. Fang, Q. Qizhi, (2012), The Studying on the Application of Data Mining based on Association Rules, *International Conference on Communication Systems and Network Technologies*, Rajkot, India pp.477-480.

Jun-Sese, S. Morishita, (Aug-2002), Answering the Most Correlated N Association Rules Efficiently. In Proceedings of the *6th European Conference on Principles of Data Mining and Knowledge Discovery*, Helsinki, Finland, pp.410-422.

Yong Xu, Sen-Xin Zhou, Jin-Hua Gong, (Aug-2005), Mining Association Rules with new Measure Criteria, Proceedings of *the Fourth International Conference on Machine Learning and Cybernetics*, Guangzhou, Vol.4, pp.2257-2260.

H.S. Anand , S.S. Vinodchandra , (Mar-2013), Applying Correlation Threshold on Apriori Algorithm, *IEEE International Conference on Emerging trends in Computing, Communication and Nanotechnology*, Tirunelveli, India, pp. 432-435.

M. Renuka Devi, A. Babysarojini, (Aug-2012), Applications of Association Rule Mining in Different Databases, *Journal of Global Research in Computer Science*, Vol.3, pp. 30-34.

J. Han, M. Kamber, J. Pei, (2012), Data Mining: *Data Mining concepts and techniques Morgan Kaufman*, Third Edition, Elsevier, India.

E. Garcia, C. Romero, S. Ventra, T. Calders, (2007), Drawbacks and Solutions of applying association rule mining in learning management systems, In Proceedings of *International*

Workshop on Applying Data Mining in e-learning, Crete, Greece, pp.-15-25.

S. Ghosh, S. Biswas, D. Sarkan, P.P. Sarkar, (Oct-2010), Mining Frequent Itemsets Using Genetic Algorithm, *International Journal of Artificial Intelligence and Applications*, Vol.1 , No.4, pp. 133-143.

B. Rani, S. Aggarwal, (Dec-2013), Optimization of Association Rule Mining Techniques using Ant Colony Optimization, *International Journal of Current Engineering and Technology*, Vol.3,No.-5, pp.1804-1808.

P. Mandrai, R. Barskar, (July-2013), A Novel Algorithm for Optimization of Association Rule with Karnagh Map and Genetic Algorithm, *4th International Conference on Computing, Communications and Network Technologies*, Tiruchengode, India, pp.1-7.

Kannika Nirai Vaani M, E. Ramaraj, (Feb-2013), An Integrated Approach to derive effective rules from Association Rule Mining using Genetic Algorithm, In Proceedings of the *International Conference on Pattern Recognition, Informatics and Mobile Engineering*, Salem, pp. 90-95.

Shweta, K. Garg, (June-2013), Mining Efficient Association Rules Through Apriori Algorithm Using Attributes and Comparative Analysis of Various Association Rule Algorithms, *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol.3, Issue-6, pp. 306-312.

R. Santhi, K. Vanitha, (April-2012), An Effective Association Rule Mining in Large Database, International Journal of Computer Application and Engineering Technology, Vol.1(2), ISSN: 2277-7962,pp.72-76.

M. Dhanda, S. Guglani, G. Gupta, (Sep-2011), Mining Efficient Association Rules Through Apriori Algorithm Using Attributes, *International Journal of Computer Science and Technologies*, Vol.2 ,Issue 3, pp.342-344.

J. Singh, H. Ram, J. Sodhi, (Jan-2013), Improving Efficiency of Apriori Algorithm Using Transaction Reduction, *International Journal of Scientific and Research Publications*, Vol.3 (1), ISSN: 2250-3153, pp.1-4.