International Journal of Hybrid Information Technology
Vol.8, No.11 (2015), pp.121-138
http://dx.doi.org/10.14257/ijhit.2015.8.11.10

Rough Set and Genetic based Model for Extracting Weighted Association Rules

Shrikant Brajesh Sagar[1] and Akhilesh Tiwari[2]

Department of CSE & IT, Madhav Institute of Technology and Science, Gwalior (M.P), India
[1]shrikantsagar19@gmail.com, [2]atiwari.mits@gmail.com

Abstract

A novel approach for the efficient weighted association rule mining proposed in this present paper. The proposed approach reducts the transactional dataset (weighted) by utilizing the power of Rough Set theory. Furthermore, proposed approach acquires the benefit for weighted measures (w-support, w-confidence) for obtaining the most profitable weighted frequent itemsets and the Genetic Algorithm for the extracting the desired set of optimized weighted association rules. Experimental analysis of proposed approach has been done and observed that the approach works well and will be helpful in situation when there is a requirement for the consideration of extracting the best weighted association rules in decision-making process.

Keywords: *Weighted items, Rough Set Theory, Apriori Algorithm, min. w-support, min. w-confidence, weighted association rules, Genetic Algorithm*

1. Introduction

Data Mining is a process of extracting hidden, reasonable knowledge from a collection of data, which can be used by users. It is the key step in the knowledge discovery process. The main tasks of Data mining are generally divided in two categories: Predictive and Descriptive. The intention of the predictive tasks is to predict the value of a particular attribute based on the values of other attributes, while for the descriptive ones, is to extract previously unknown and useful information such as patterns, associations, changes, anomalies and significant structures, from large databases. There are several techniques satisfying these objectives of data mining. Some of these can be classified into the following categories: clustering, classification, association rule mining, sequential pattern discovery and analysis. In this study we considered Association Rule Mining for knowledge discovery and generate the rules by applying our developed approach on real and synthetic databases.

Association Rule is an important type of knowledge representation associations with the items present in large number of transactions. Let {I=I_1, I_2…,I_n} be a set of n distinct attributes, T be transaction that contains a set of items such that T⊆I, D be a database with different transaction records Ts. An association rule is an implication in the form of [X=>Y, sup, conf], where X, Y⊂I are sets of items called itemsets, and X∩Y=∅. X is called antecedent while Y is called consequent, the rule means X implies Y. An itemset is said to be frequent if its support (supp) value is greater than a user defined minimum support value (minsupp). The support of X∩Y (supp) in the transactions is greater than minsupp, moreover while X appears in a transaction, Y is about to appear in the similar transaction with a possibility confidence (conf). Given a threshold of minsupp and conf, methods of mining association rule have become active research areas since the publication of Agrawal, Imielinski and Swami [1] and Agrawal and Srikant papers [2].

In recent years many improvement have been suggested for generating frequent itemsets and association rules. Although these improved algorithms can reduce the number of candidate itemsets or improve the mining efficiency by pruning methods, but still can't completely solve the problem of which candidate itemsets appear no longer. And, what's more, facing masses of data, to adopt single association rules mining algorithm is not the solution to problems above, so it is an important challenge to design a highly efficient algorithm in the field of association mining. Association rules mining based on Rough Set theory becomes a basic way to process problems of mining massive data.

Rough set theory is put forward in 1982 by professor Pawlak, which is a mathematical tool analyze quantitatively to deal with imprecision, inconsistent, incomplete information and knowledge [3]. Recently, Rough Set theory is widely used in the field of machine learning, data mining, and pattern recognition. Deleting redundant attributes of the relational database can improve the clarity of potential knowledge of the information system significantly. In the current literature, several research works have been combined the rough set theory with other artificial intelligence methods such as neural networks, fuzzy logic, additionally to other methods resulting in some good results. Among other computational problems, rough set deal with problems such as data significance evaluation, hidden pattern discovery from data, decision rule generation, data reduction and data-driven inference interpretation (Pawlak, 2004).

Because some attributes in database belong to quantitative type, so these types of attributes considered as weight of items. In real-life database different transactions have different weights. The significance of the attributes in a transaction within the whole item space is considered to be same without its significance in traditional association rule mining. If the association rules are generated in this fashion, some interesting rules are missed. *e.g.*, [wine → salmon, 2%, 60%] may be more essential than [bread → milk, 4%, 60%] even if the earlier holds a lower supp. This is because those items in the earliest association rule generally appear with more profit per unit transaction, but the traditional association rules mining (ARM) simply ignores this variation. The model in also considers only whether an item is present in a transaction, but does not take into account the weight/intensity of an item within a transaction. For example, a customer may purchase 13 bottles of coke and 6 bags of snacks and another may purchase 4 bottles of coke and 1 bag of snacks at a time. The conventional association rules approach treats the above two transactions in the same manner, which could lead to the loss of some vital information. Assume, for example, that if a customer buys more than 7 bottles of coke, he is likely to purchase 3 or more bags of snacks. Otherwise, the purchase tendency of coke is not strong. The traditional association rule cannot express this type of relationship. With this knowledge, the supermarket manager may set a promotion such as if a customer buys 8 bottles of coke; he can get two free bags of snacks. So, weight is associated for the items through which rules are found.

Weighted Association Rules cannot only improve the confidence in the association rules, but also give a method to do more efficient objective marketing by recognizing or segmenting consumers based on their potential degree of reliability or quantity of purchases. The major challenge of adapt traditional ARM model in a weighted setting is the invalidation of the “downward closure property”, which is used to justify the efficient iterative process of discoveing and pruning large itemsets from its subsets. In order to attempt this challenge, an daptation made on the traditional association rule mining model under the “significant – weighted support” metric framework instead of the “support” framework with only quantitative attributes. In new proposed method, the iterative generation and pruning of considerable itemsets is acceptable by a “weighted downward closure property”.

The fundamental idea behind w-support is that a frequent item set may not be as essential as it emerges, because the weights of different transactions are different. These weights are totally derived from the internal formation of the dataset based on the statement that good transactions consist of good items. Based on w-support measure, frequent itemsets generates after applying traditional Apriori algorithm. Generated frequent itemsets may be in large quantity so these frequent items may be optimized with some type of evolutionary algorithm such as, Genetic algorithm.

In this paper, Rough set theory and the Genetic algorithm based model proposed for finding the best weighted association rules. Both techniques applied on the database of weighted items. Rough set theory used for reducing the redundant attributes. Genetic algorithm is efficient for global search work, especially when the search space is too large to use a deterministic search method. It imitates the mechanics of natural species evolution with genetics principles, such as natural selection, crossover, and mutation.

2. Related Works

Many researchers have been proposed interesting algorithms for either frequent itemset generation or discovering the association rules from the frequent itemsets. There are different methods for the feature selection and reduction and the optimization of association rules given by different researchers. Related to this present paper work, Rough Set Theory, Weighted Association Rules Mining and the Genetic Algorithm are presented in section A, section B and section C respectively.

2.1. Rough Set Theory

The Rough set theory is a mathematical structure for analyzing decision Table. Rough set theory is given by professor Pawlak in 1982, which is a mathematical tools investigating quantitatively to treat imprecision, inconsistent, incomplete information and knowledge. Currently, Rough Set theory is expansively used in the field of machine learning, data mining such as feature selection, reduction and pattern recognition [4]. Removing redundant attributes of the relational database can improve the accuracy of knowledge of the information system significantly [5].

Rough set theory and the association rules algorithm both are mining methods to discover contained rules model from huge amounts of database. Rough Set based Association rules mining becomes a necessary way to process problems of mining large data. In this related work, Chen Chu-xiang, *et al.,* [6] proposed Rough set Theory based an improvement Apriori Arithmetic. This R_Apriori algorithm resolves the problems of Apriori algorithm to improve the effectiveness of the algorithm. XUN Jiao, XULian-cheng, QILin [7] proposed ARM Algorithm based on Rough Set. The benefit of this technique lies in three phases, including the removal of redundancy attributes; reduct the number of attributes, while scanning decision Table just once can generate decision attribute sets. Aritra Roy and Rajdeep Chatterjee [8] proposed a new hybrid Rough and Fuzzy based ARM algorithm.

2.1.1. Related Concept and Properties: Rough Set Theory is a tool for studying imprecision, vagueness and uncertainty in data analysis. It focuses on delivery patterns, rules and knowledge in data.

There are some basic concepts and properties given below:

2.1.1.1. Information System: A database is representing as a Table, of which every row is an object and every column is an attribute that can be considered for every object or simply provided by the user. Such type of Table can be described as an information system.

Information system S = (U, A)

Where U is a set of finite number of object, A is the finite set of the attributes which is non-empty.

2.1.1.2. Decision System: The information system with the decision attribute represents the set of class and so-called decision system.

Decision System S = <U, R, V, f>

Where U is the sets of non-empty finite object, It is denoted by U={X_1, X_2,…., X_n}.

A is the attribute sets of non-empty finite object, A = CUD. C and D are called conditional attributes and decision attributes respectively, $C \cap D = \emptyset$. V is sets of attribute values.

f: U×A→V is an information function.

2.1.1.3. Indiscernibility: Given a decision system S = <U, R, V, f>, for each subset B, $B \subseteq R$ define an indiscernibility relation IND(B),

$$IND(B)= \{(x, y) \in U \times U. \forall b \in B(V_b(x) = V_b(y))\} \quad (1)$$

Obviously, indiscernibility relation is an equivalence relation.

2.1.1.4. Upper and Lower Approximation Sets: Rough set concept can be defined by means of topological operations, interior and closure, called approximations.

Given an information system S = <U, R, V, f> for each subset X ∈ U and indiscernibility relation R, upper and lower approximation sets of X can be defined by the basic sets of A respectively as follows:

• The set of all objects which can be with certainty classified as members of X with respect to R is called the R-lower approximation of a set X with respect to R, and denoted by

$$R_*(X) = \{x \in U | [X]_R \subseteq X\} \quad (2)$$

• The set of all objects which can be only classified as possible members of X with respect to R is called the R-upper approximation of a set X with respect to R, and denoted by

$$R^*(X) = \{x \in U | [X]_R \cap X \neq \emptyset\} \quad (3)$$

• The set of all objects which can be uncertainly classified neither as members of X nor as members of - X with respect to R called the boundary region of a set X with respect to R, and denoted by $RN_R(X)$, i.e.

$$RN_R(X) = R^*X – R_*X \quad (4)$$

2.1.1.5. Simplified Decision Table: In the decision Table S = <U, C, D, V, f>, if U/C = {C_1, C_2,...,C_m} = {[$x_{1'}$]C, [$x_{2'}$]C,... ..,[$x_{m'}$]C} and S = (U'={$x_{1'}$, $x_{2'}$,...,$x_{m'}$}, C, D, V, f) is called the Simplified Decision Table.

The accuracy of approximation of a Rough set X can be numerically illustrated as:

$$\alpha_R(X) = \frac{|R_*(X)|}{|R^*(X)|} \quad (5)$$

2.1.1.6. Rough set Based Transformation of Transactional Database: Transaction database can be transformed to a decision system. Decision System can be described using data Tables, which takes the lines of transaction database D as the object I[j] of the decision system, and take itemsets of transaction as attribute sets of the decision system.

$$R_{ij} = \begin{cases} 1, & I[j] \in T[i] \\ 0, & I[j] \notin T[i] \end{cases} \qquad 0< i \leq |I|, 0 < j \leq |T| \quad (6)$$

$$Supp(X=>Y) = \frac{|[X]_R \cap [Y]_R|}{|T|} \quad (7)$$

Where $[X]_R$ is indistinguishability class of X whose attribute sets are R, $[Y]_R$ is indistinguishability class of Y whose attribute sets are R.

Suppose Decision system S = <U, R>, where U is consider the non-empty finite sets of the object and R is the non-empty finite set of all the attributes.

Assume X and Y are two subsets of R, $X \cap Y \neq \emptyset$, and $[X]_R \cap [Y]_R = [X \cup Y]_R$.

2.2. Weighted Association Rules Mining

The weighted items transaction database is defined as follows: Given a database D with a set of transactions $T = \{t_1,t_2,...,t_m\}$, a set of items $I = \{i_1,i_2,...,i_n\}$ and a set of positive weights $W = \{w_1,w_2,...,w_n\}$ corresponds with each item in I.

Suppose $I = \{i_1, i_2, i_3,....i_n\}$ be a set of distinct items and W be a set of non-negative real numbers. A pair (x, w) is called a weighted item, where $x \in I$ and $w \in W$ is the weight associated with x. A transaction is a set of weighted items, each of which may appear in multiple transactions with different weights. In this paper, weight of an item considering bases on the quantity of items or number of items and weighted support used for generating frequent itemsets.

For example: Consider the database in Table 1 and Table 2. Table 1 has six transactions $T = \{t_1,…,t_6\}$, and five items I = {A, B, C, D, E}. As discussed above, the weights of these items are in quantity based so W = {2, 1, 3, 4, 2} stored in Table 2.

Table 1. Transactional Database

Transactions	Items
1	A, B, D, E
2	B, C, E
3	A, B, D, E
4	A, B, C, E
5	A, B, C, D, E, F
6	B, C, D

Table 2. Weighted Items

Items	Weights
A	2
B	1
C	3
D	4
E	2

According to the definition of weighted support in this paper, the weighted support of the items as follows:
A = 0.16 or 16%, B = 0.08 or 8%, C = 0.25 or 25%, D = 0.33 or 33%, E = 0.16 or 16%.

2.2.1. Related Definitions: There are some basic definitions related to weight association rules given below:

2.2.1.1. Weighted Attributes: Weighting attributes $A = \{a_1, a_2, a_3,.....a_n\}$ are variables selected to calculate weights. There are two types of weights – the item weight and the itemset weight.

2.2.1.2. Item Weight: Item weight is a value attached to an item representing its meaning.

2.2.1.3. Itemset Weight: Based on the item weight, the weight of an itemset can be derived from the weights of its enclosing items.

2.2.1.4. Transactional Weight: Transaction weight is a kind of itemset weight. It is a value attached to every of the transactions. Generally the higher a transaction weight, the more it gives to the mining result. In a supermarket scenario, the weight can be the "significance" of a customer who made a assured transaction.

2.2.1.5. Weighting space: Weighting space (WS) is the context within which the weights are estimated.

(i) Inner-transaction space: This space refers to the host transaction that an item is weighted in.

(ii) Item space: This space refers to the space of the item collection that covers all the items appears in the transactions.

(iii) Transaction space: This space is defined for transactions rather than for items.

2.2.1.6. Weighted Support: Weighted support of an itemset is a set of transaction T respects a rule R in the form A=>B, where A and B are non-empty sub-itemsets of the item space I and they distribute no item in common. Its w-support is the fraction of the weight of the transactions that contains both A and B relative to the weight of all transactions.

W-Support for frequent items formulated as:

$$\frac{\sum_j \sum_i q_{ij}}{\sum N} \tag{8}$$

Where i = 1, 2,….k and j = 1, 2,....n and q_{ij} represents a quantity of an item i ∈ I , in a j^{th} transaction and ∑ N is the sum of all transactions of all items.

W-Support for frequent itemsets formulated as:

$$\frac{\sum_j \sum_i (Xq_{ij} \cap Yq_{ij})}{\sum N} \tag{9}$$

Where i = 1, 2,….k and j = 1,2,....n and q_{ij} represents a quantity of an item i ∈ I , in a j^{th} transaction and ∑ N is the sum of all transactions of all items.

2.2.1.7. Weighted Confidence: Weighted confidence of an itemset is a set of transaction T respects a rule R in the form A=>B, where A and B are non-empty sub-itemsets of the item space I and they distribute no item in common. Its w-confidence is the fraction of the weight of the transactions that contains both A and B relative to the weight of transactions A or B.

Weighted Confidence formulated of a rule as:

$$\frac{\sum_j \sum_i (Xq_{ij} \cap Yq_{ij})}{\sum X} \tag{10}$$

Where i = 1, 2,….k and j = 1, 2,....n and q_{ij} represents a quantity of an item i ∈ I , in a j^{th} transaction and ∑ X is the sum of transaction of sub-item.

2.2.1.8. Weighted Downward Closure Property: It means that any subset of a weighted frequent itemset is a frequent itemset. In this paper, the idea of replacing the support with significance is proposed for the first time and we argue that a "weighted downward closure property" can be retained by using weighted support.

In this paper, quantity based association rules consider as weighted association rule mining, in which different type of amount or weight associated with each item in a transaction. Consecutively, this provides us a prospect to associate a weight parameter with each item in a resulting association rules and called weighted association rules. For example, A[2,4]=>B[1,3] is a weighted association rule representing that if a customer buys item "A" in the amount 2 and 4 then he is possibly to buy item "B" in the amount 1 and 3. Thus weighted association rules improve the confidence in the rules and offer a method to do more efficient target marketing by recognizing or dividing customers based on their possible degree of reliability of purchases.

The traditional association rule mining framework occupies the support measure, which treats each transaction uniformly. But in the real-life database different items in different transactions have different weights. Thus, in this paper, "weighted support" measurement formulated in place of the "support" measurement for generating the frequent itemsets, and then the weighted association rules for each frequent item set are generated. Our objective is to fragment the weight domain of each item in the item set so that rules with higher confidence can be discovered. In this novel proposed model, the iterative generation and pruning of considerable itemsets is justified by a "weighted downward closure property".

Till now, based on the weighted concepts many efficient algorithms have been proposed by the researchers for finding frequent itemsets using user defined minimum weighted support. In most of the related work the weighted support is measured by multiplying a support with a known weight of items. Feng Tao, Fionn Murtagh, Mohsen Farid [9] proposed WARM using w-support and significant framework, In this algorithm both scalable and efficient in finding important relationships in weighted setting performed on simulated datasets. Ke Sun and Fengshan Bai [10] proposed mining weighted association rules without preassigned weights. In this algorithm a new measure w-support introduced which does not need preassigned weights. It takes the feature of transactions into deliberation using link-based models. Weimin Ouyang, Qinhua Huang [11] proposed an algorithm for mining both direct and indirect weighted association rules. Preetham Kumar, Ananthanarayana V S [12] proposed two algorithms for discovery of weighted association rule mining from large volumes of data in a single scan of database structured in the form of weighted tree. Bac Le, Huy Nguyen [13] proposed an efficient algorithm for mining frequent weighted itemsets from weighted items databases. Guo-Cheng Lan, Tzung-Pei Hong, Vincent S. Tseng [14] proposed an algorithm for mining high transaction-weighted utility itemsets, which considers not only individual profits and quantities of the items in a transaction, but also the contribution of each transaction in a database. Yun and U. [15] proposed an efficient mining of weighted interesting patterns with a strong weight and/or support affinity. M. Sulaiman Khan, Maybin Muyeba, and Frans Coenen [16] proposed weighted association rule mining from binary and fuzzy data. In this related work on mining association rules of weighted items, "weighted support" plays a major role. In this way researchers proposed many formulae for measuring "weighted support" framework.

2.3. Genetic Algorithm

Optimization of association rules by the GA is the important task in the data mining. As discussed earlier that optimization works for maximization and minimization of objective function. Minimization means to extract only the best association rules from the large amount of database. In this thesis implementation work for extracting the association rules have been completed.

The GA for optimization of association rules is divided into three parts: representation of frequent items in the binary representation, genetic operators customized to hold individuals representing rules, design of fitness functions for optimization of association rules [17].

2.3.1. Individual Representation

Genetic algorithms (GA) for optimization of association rules first encode an initial population is created with Frequent Itemsets into binary encoding. A Population is a group of individuals (Chromosomes) and represents a candidate solution. A Chromosome is a string of genes.

For example A, B, C etc. alphabets used in hexadecimal therefore Frequent Itemsets first converted to integer then binary. Each frequent item set is represented by one bit having two possible values:

For item is in solution, and for item is not in solution.

e.g., if population is 11100000 then 1st, 2nd and 3rd frequent item set is selected, rest not selected.

2.3.2. Genetic Operators

The GA applies genetic operators such as selection, crossover and mutation on a primarily random population so as to calculate an entire generation of new strings. The GA applies to generate solutions for succeeding generations. The probability of an individual reproducing is proportional to the goodness of the solution it represents. Hence the quality of the solutions in successive generations improves. The process is terminated when an acceptable or optimum solution is found. The GA is suitable for problems which need optimization, with respect to some computable condition.

Genetic algorithms operators work as follows:

(i). Selection: The selection of the individual member from the chromosome can be done using the Tournament selection method. Tournament selection is a process of select elements from the population of chromosomes in a method that is proportional to their fitness.

(ii). Crossover: Crossover is simply a matter of replacing some of the genes in one parent by the corresponding genes of the other.

Suppose there are two strings :

000101011|0011100
000001011|0001101

Choose a random bit along the length, let at position 9, and exchange all the bits after that point. The resulting chromosomes become

0001010110001101
0000010110011100

(iii). Mutation: While crossover operates on two or more parental chromosomes, mutation locally but randomly modifies a solution. There are many variations of mutation, but it usually involves one or more changes being made to an individual's trait or traits. In other words, mutation performs a random walk in the vicinity of a candidate solution.

2.3.3. Fitness Functions

The GA applied on the selected population from the database and computes the fitness function after each step until the GA is terminated. Fitness function is an objective function used to summarize as how close a given suggest solution is to achieving the required solution. The fitness function is always problem dependent. In this present work, to maximize our involvement fitness value is an inverse of average of selected itemsets in the given population.

Therefore Fitness value of selected itemsets calculated as

$$F = \frac{1}{\sum(w-support)/n} \qquad (11)$$

Where n = number of selected frequent itemsets.

Recently many optimization techniques have been proposed basis on the fitness value. Genetic algorithm is the one of them to optimization of association rules. In this paper, proposed work is for finding the best association rules basis on the fitness value. Many effective algorithms have been proposed for optimization of association rules. X. Yan, C. Zhang, and S. Zhang [18] proposed the GA based approach for recognizing association rules without identifying actual minsupp. S. Ghosh, S. Biswas, D. Sarkar and P. P. Sarkar [19] proposed mining frequent itemsets using the GA. Manish Saggar, A. K. Agrawal and A. Lad [20] proposed optimization of ARM using improved genetic algorithms. In this algorithm, the rule generated by ARM technique do not consider the negative incidents of attributes in them, but by using Genetic Algorithms (GAs) over these rules the method can predict the association rules which holds negative attributes. Diana Martin and, *et al.*,

[21] proposed a new multiobjective evolutionary algorithm for finding a reduced set of interesting positive and negative quantitative association rules.

3. Proposed Method

The proposed approach in this paper comprises a novel algorithm based on Rough Set Theory and the Genetic Algorithm applied on weighted items with the Rough Set based association rule mining for generation of desired association rules. Experimental results analyze the power of proposed algorithm, in which profitable desired weighted association rules extract with the utilization of genetic algorithm.

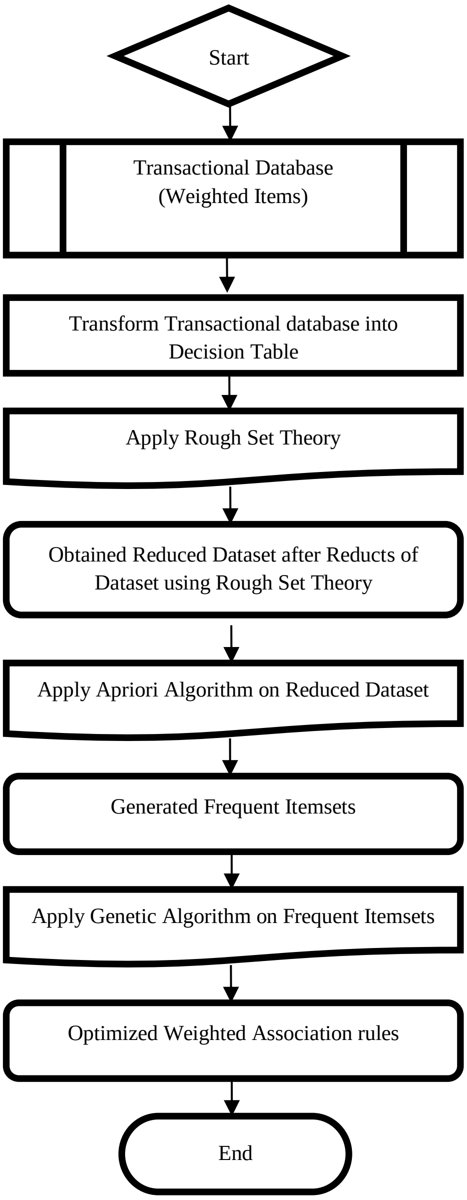

Start
Transactional Database
(Weighted Items)
Transform Transactional database into
Decision Table
Apply Rough Set Theory
Obtained Reduced Dataset after Reducts of
Dataset using Rough Set Theory
Apply Apriori Algorithm on Reduced Dataset
Generated Frequent Itemsets
Apply Genetic Algorithm on Frequent Itemsets
Optimized Weighted Association rules
End

Figure 1. Block Diagram of Proposed Algorithm

In this proposed algorithm, initially Rough Set Theory is applied on the dataset which eliminates uncertain and incomplete data from the dataset. Association rules mining based on Rough Set Theory becomes a necessary way to process problems of mining massive data. Rough Set Theory applied in a preprocessing step of data mining for reducing the number of attributes from the database. In the next step, Apriori Algorithm applied on the reduced dataset (weighted) for finding the frequent itemsets and desired association rules bases on the weighted support and weighted confidence measurement. In the last step, the Genetic Algorithm applied for extracting profitable weighted association rules.

Generally, optimization techniques work for maximization or minimization for an objective function. In this proposed work, the Genetic Algorithm used for the minimization of association rules or extracting the best association rules from the database.

3.1. Proposed Algorithm:

Input: Transactional database of weighted items, min. w-support, min. w-confidence.
Output: Optimized weighted association rules.

1. Start
2. Load transactional database (weighted).
3. Transform transactional database (weighted) into decision Table system, which contain the conditional and decisional attributes
4. Apply Rough Set Theory.
5. Assume minimum support.
6. Choose two items from decision Table system, which contains the least attributes $\{C_1, C_m \in L_{k-1}\}$, $[C_1 \cup C_m]_1$, $[C_1]_1 \cap [C_m]_1$ $(1 \leq m \leq T)$
7. If $[C_1]_1 \cap [C_m]_1 / [C_1 \cup C_m] <$ min. support
8. Delete this item from decision Table system, if not, then reserve, and continue to study the next pairs attribute.
9. If the number of items “weight” in a list of attributes < min. support, then delete this list of attributes.
10. Obtained simplified decision Table.
11. Apply Apriori Algorithm on remaining attributes of a decision Table.
12. Assume min. w-support and min. w-confidence.
13. Calculate w-support and w-confidence.
    The w-support formulated as:
    W-Support = $\frac{\Sigma(X \cap Y)}{\Sigma N}$
    Where $\Sigma(N)$ is the total number of transactions and $\Sigma(X \cap Y)$ is the numbers of transactions containing both items X and Y. W-Support is usually used to remove non-interesting rules.
    The w-confidence formulated as:
    W-Confidence = $\frac{\Sigma(X \cap Y)}{\Sigma(X)}$
    Where $\Sigma(X)$ is the number of transactions of item X. A higher w-confidence recommends a strong association between X and Y.
14. Generated frequent itemsets
15. Apply Genetic Algorithm on Frequent Itemsets, Genetic optimize these Frequent Itemsets
16. Input the termination condition of genetic algorithm.
17. An initial population is created with Frequent Itemsets. A Population is a group of individuals (Chromosomes) and represents a candidate solution. A Chromosome is a string of genes.
18. Calculate the fitness value of selected frequent itemset:

$$f = \frac{1}{\sum (w - support)/n}$$

19. Select chromosomes with higher fitness from the frequent item set using Tournament selection method.
20. Crossover between the selected chromosomes to produce new offspring with better higher fitness.
21. Mutate the new chromosomes if needed.
22. Terminate when an optimum solution is found.
23. This generational process is repeated until a termination condition has been achieved.
24. Output of Genetic Algorithm feed to Apriori.
25. Apriori start making rules with Genetic output and generate Strong rules.
26. End

Illustration

1. Load transactional database (weighted items).

Table 3. Transactional Database

T-ID	ITEMS
T1	A(3), B(2), E(4)
T2	B(1), D(1)
T3	B(2), C(1)
T4	A(1), B(1), D(1)
T5	A(2), C(2)
T6	B(1), C(1)
T7	A(1), C(1)
T8	A(1), B(1), C(2), E(3)
T9	A(2), B(1), C(2), F(1)
T10	E(1), F(1)

2. Transform transactional database (weighted items) into decision Table system, which contain the conditional attribute and decisional attributes.

Table 4. Decision Table System

Items/TID	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
A	3	0	0	1	2	0	1	1	2	0
B	2	1	2	1	0	1	0	1	1	0
C	0	0	1	0	2	1	1	2	2	0
D	0	1	0	1	0	0	0	0	0	0
E	4	0	0	0	0	0	0	3	0	1
F	0	0	0	0	0	0	0	0	1	1

3. Assume minimum support = 20%
4. Choose two items from decision Table system, which contains the least attributes
C_1, $C_m \in L_{k-1}$, $[C_1 \cup C_m]$, $[C_1] \cap [C_m]$, $(1 \leq m \leq T)$
A∪B = {T1, T2, T3, T4, T5, T6, T7, T8, T9)
A∩B = {T1, T4, T8, T9)
Support = 4/9 *100= 44%
A∪D = {T1, T2, T4, T5, T7, T8, T9}
A∩D = (T4)

Support = 1/7*100 = 14.28%
A∪E = {T1, T4, T5, T7, T8, T9, T10}
A∩E = (T1, T8)
Support = 2/7*100 = 28.57%
A∪F = {T1, T4, T5, T7, T8, T9, T10}
A∩F = (T9)
Support = 1/7*100 = 14.28%

From the above definition [A∩D] = {T4} < min. support, [A∩F] = {T9} < min. support, Therefore delete D, F from Table III.

5. If the number of items "weight" in a list of attributes < min. support, then delete this list of attributes.

Therefore simplified decision Table as follows:

Table 5. Simplified Decision Table

Items /TID	T1	T3	T4	T5	T6	T7	T8	T9
A	3	0	1	2	0	1	1	2
B	2	2	1	0	1	0	1	1
C	0	1	0	2	1	1	2	2
E	4	0	0	0	0	0	3	0

6. Apply Apriori Algorithm

Assume min. weighted support = 20%, min. weighted confidence= 60%

Generated frequent itemsets with weighted support:

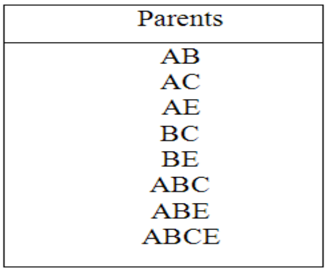
Table 6. Frequent Itemsets

Frequency sets	Itemsets	W-Support
1-itemset	A B C E	29.41% 23.52% 26.47% 20.58%
2-itemset	AB AC AE BC BE	35.29% 38.23% 32.35% 32.35% 29.41%
3-itemset	ABC ABE	26.47% 41.17%
4-itemset	ABCE	26.47%

Apply Genetic Algorithm on Frequent Itemsets, to optimize these Frequent Itemsets.

The GA works as follows: An initial population is created with Frequent Itemsets. A Population is a group of individuals (Chromosomes) and represents a candidate solution. A Chromosome is a string of genes.

Table 7. Chromosomes



Parents
AB
AC
AE
BC
BE
ABC
ABE
ABCE

Convert it into binary. Each frequent item set is represented by one bit having two possible values:

First for item is in solution and second for item is not in solution.

e.g., - if population is 11100000 then 1st 2nd 3rd frequent item set is selected rest not selected

Fitness function: To maximize our involvement fitness value is average of selected itemsets in the given population.

Generating initial population given in the Table 8:

Table 8. Selected chromosomes

Initial population	Solution	Fitness Value
00001010	{BE, ABE}	3.70841
01101101	{AC,AE,BE,ABC,ABCE}	3.26947
00110110	{AE, BC, ABC, ABE}	3.02252
00011100	{BC, BE, ABC}	3.40020
00010110	{BC, ABC, ABE}	3.00030
01001000	{AC, BE}	2.95683
00000010	{ABE}	2.42895
00001101	{BE, ABC, ABCE}	3.64298

Select chromosomes with higher fitness (default fitness value of chromosomes are their support for selection, selection tournament required).

Tournament size= 2.

In above process chromosomes 2, 3, 6 & 7 selected twice, but 1, 4, 5 & 8 were not selected in therefore they are eliminated.

Apply Crossover between the selected chromosomes for producing the new offspring with better higher fitness value.

Table 9. Crossover

Parent1	Parent2	Offspring 1	Offspring 2
0110|1101	0100|1000	01101000	01001101
0011|0110	0000|0010	00110010	00000110
0100|1000	0110|1101	01001101	01101000
0000|0010	0011|0110	00000110	00110010

Mutate the new offspring if needed.

Table 10. Mutation

Offspring 1	Fitness value	Offspring 2	Fitness value
01101000	2.6177	01001101	2.1707
00110010	2.4723	00000110	3.8697
01001101	2.1707	01101000	2.6177
00000110	3.8697	00110010	2.4723

Select those offspring who can survive in the environment. Only fitter offspring go to next generation

01101000
00110010
01001101
00000110
01101000
00110010
01001101
00000110

This is first generation population

.

.

Similarly after 80 generation population will be

1 1 1 0 0 0 1 0
1 1 1 0 0 0 1 0

are selected as fitter parents they are able to reproduce new generation.

This generational process is repeated until a termination condition has been reached.

Genetic algorithm output logical value ‘0’ for terminate and ‘1’ for select.

Table 11. Genetic Output

Chromosome s	Logical O/p	Selected Chromosomes
1	1	AB
2	1	AC
3	1	AE
4	0	_
5	0	_
6	0	_
7	1	ABE
8	0	_

Terminate when an optimum solution is found. BC, BE, ABC, & ABCE are optimum solutions for Genetic Algorithm therefore terminated.

We have not need any new generation but our need to select the fit parents for making strong rules.

Table 12. Optimized Itemsets

Selected Chromosomes	w-support
AB	35.29%
AC	38.23%
AE	32.35%
ABE	41.17%

Generated output of Genetic Algorithm feed to Apriori.
Apriori start making rules with Genetic output and generate strong rules.

Table 13. Optimized Rules

Association Rules
A=>B, B=>A
A=>C, C=>A
A=>E, E=>A
A=>BE, AB=>E, AE=>B

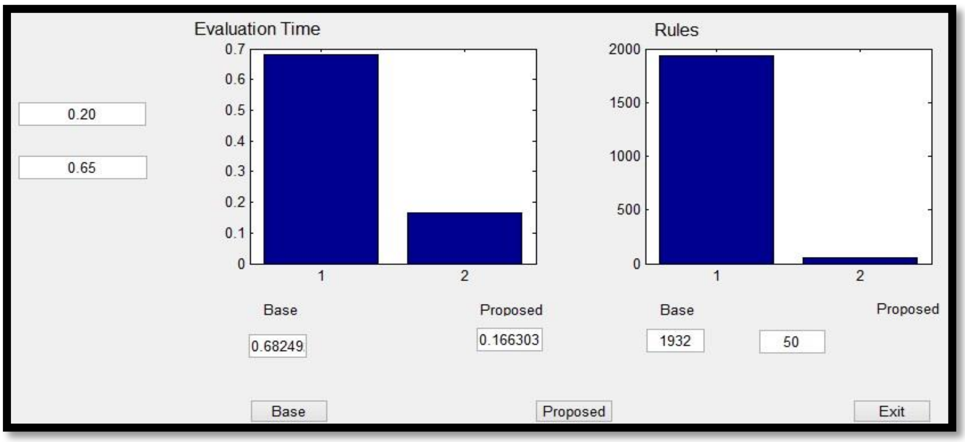
4. Experimental Analysis

All the experiments have been performed on a personal computer with following configuration: Pentium core 2 duo 2.2 GHz processors, 2 GB Ram and windows 7 (64-bit) operating system. Proposed Algorithm has been implemented using MATLAB 2013a.There is mainly two parameters on which implementation work have been evaluated:

1. Execution Time and
2. Number of Association Rules

This dataset is the Heart Disease Data Set taken from taken from Machine Learning Repository of UCI, which has 76 attributes, but all available experimentation refer to using a subset of 14 of them.

Particularly, the database of Cleveland is the merely one that has been used by ML researchers to this date. The "objective" field refers to the occurrence of heart disease in the patient. This is an integer value from 0 (no presence) to 4. Researches with the database of Cleveland have determined on merely attempting to discriminate presence (types 1, 2, 3, 4) from absence (value 0).



0.20

0.65

Base 0.68249

Proposed 0.166303

Base 1932

Proposed 50

Figure 2. Experimental Results

The comparative analysis shown in the Figure 2, in which time complexity and number of association rules improve the efficiency of the proposed algorithm.

4.1. Comparative Study and Results

In this paper, comparative analysis have done on the basis of Rough Set based association rule mining (base work) [7]and Rough set and Genetic based weighted association rule mining (proposed work). In the base work, Rough set theory and Apriori algorithm used for the generation of association rules and in the proposed work, additionally weighted concept with the Genetic algorithm introduced.

When researchers have executed the application, they found that association rules results remain same in the proposed work at each execution because of the Genetic algorithm. The execution time required to execute the application for the Proposed (RS_Apriori_GA) and Rough set based association rules mining (RS_Apriori) [7] changed in each execution but in proposed method execution time always appear less than the existing work. This is the reason to execute the application 8 repetitive times and to find the results. Later on, they are compared in Tables with their respective graphs.

4.2. Execution Time

Table 14 and Figure 3 show the time requirement of the proposed work over base work. The Execution time estimated on different support and confidence.

Table 14. Execution Time

S. No.	Supp (%)	Conf (%)	Base RS_Apriori (Time in Seconds)	Proposed RS_Apriori_GA (Time in Seconds)
1	15	60	0.7967	0.2022
2	20	65	0.6824	0.1663
3	25	70	0.7498	0.1877
4	30	75	0.2293	0.1658
5	35	80	0.2289	0.1762
6	40	85	0.2262	0.1799
7	45	90	0.2287	0.1559
8	50	95	0.2280	0.1677

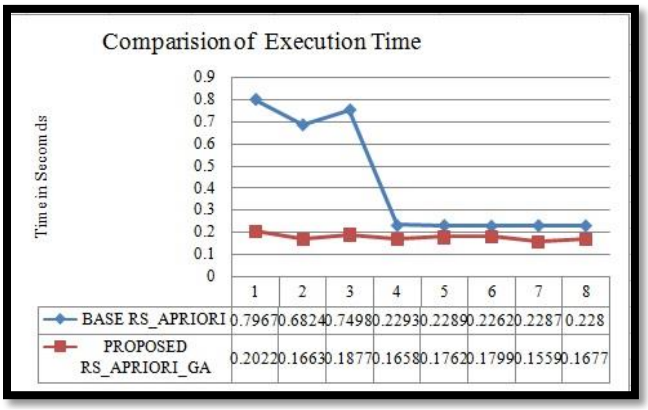


Figure 3. Execution Time

4.3. Number of Association Rules

Table 15 and Figure 4 show the number of weighted association rules generated by base and proposed work. The number of weighted association rules generation estimated on different support and confidence.

Table 15. Number of Association Rules

S. No.	Supp. (%)	Conf. (%)	Base Algorithm	Proposed Algorithm
1	15	60	2000	50
2	20	65	1932	50
3	30	75	602	50

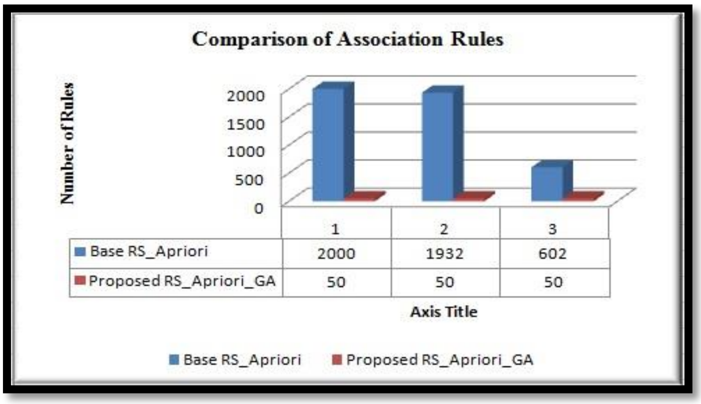


Figure 4. Number of Rules Generated by Both Methods

5. Conclusion

This paper proposes a novel approach for weighted association rules mining based on Rough Set and Genetic Algorithm. Rough set based concept used only for the reduct of the initial database. Therefore it is clear that the proposed approach works on the reduced dataset which leads to the improvement in the performance. Additionally, proposed approach makes use of the Genetic Algorithm, which helps in extracting profitable weighted association rules that may also be considered for a profitable decision-making process. Experimental results show the effectiveness of the proposed algorithm by comparative analysis to the base work.

References

[1] R. Agrawal, T. Imielinski and A. Swami., “Mining association rules between sets of items in large databases”, In the Proc. of the ACM SIGMOD Int'l Cod, on Management of Data (ACM SIGMOD '93), Washington, USA, **(1993) May**.

[2] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules”, In Proceedings of the 20th VLDB Conference, **(1994)**, pp. 487-499.

[3] Z. Pawlak, “Rough Set Theory and its application”, Journal of Telecommunication and Information Technology, **(2012).**

[4] W. Guo-Yin, Y. Yi-Yu and Y. Hong, “A Survey on Rough Set Theory and Applications”, Chinese Journal of Computers, **(2009)**, pp. 1230-1246.

[5] Y. Yao, “Notes on Rough Set Approximations and Associated Measures”, Journal of Zhejiang Ocean University (Natural Science), vol. 29, no. 5, **(2010)**, pp. 399-410.

[6] C. Chu-Xiang, S. Jian-jing, C. Bing, S. Chang-Xing and W. Yun-Cheng, “An Improvement Apriori Arithmetic based on Rough set Theory”, IEEE, **(2011)**.

[7] X. Jiao, X. Lian-Cheng and Q. Lin, "Association Rules Mining Algorithm Based on Rough Set", International symposium on information technology in medicine and education, IEEE, **(2012)**.

[8] A. Roy and R. Chatterjee, "Introducing New Hybrid Rough Fuzzy Association Rule Mining Algorithm", ACEEE, Proc. of Int. Conf. on Recent Trends in Information, Telecommunication and Computing, ITC, **(2014)**.

[9] F. Tao, F. Murtagh and M. Farid, "Weighted Association Rule Mining using Weighted Support and Significance Framework", SIGKDD, **(2003)**.

[10] K. Sun and F. Bai, "Mining Weighted Association Rules without Preassigned Weights", Knowledge and Data Engineering, IEEE Transactions on, vol. 20, Issue 4, **(2008)**, pp. 489-495.

[11] W. Ouyang and Q. Huang, "Discovery Algorithm for Mining both Direct and Indirect Weighted Association Rules", International Conference on Artificial Intelligence and Computational Intelligence, IEEE, **(2009)**.

[12] P. Kumar and V. S. Ananthanarayana, "Discovery of Weighted Association Rules Mining", IEEE, **(2010)**.

[13] B. Le and H. Nguyen, "Efficient Algorithms for Mining Frequent Weighted Itemsets from Weighted Items Databases", IEEE, **(2010)**.

[14] G.-C. Lan, T.-P. Hong and V. S. Tseng, "Mining High Transaction-Weighted Utility Itemsets", Second International Conference on Computer Engineering and Applications, IEEE, **(2010)**.

[15] U. Yun, "Efficient mining of weighted interesting patterns with a strong weight and/or support affinity", Information Science, vol. 177, **(2007)**, pp. 3477-3499.

[16] M. S. Khan, M. Muyeba and F. Coenen, "Weighted Association Rule Mining from Binary and Fuzzy Data", pringer-Verlag Berlin Heidelberg, LNAI, vol. 5077, **(2008)**, pp. 200–212.

[17] D. E. Goldberg, Editor, "Genetic Algorithms in Search Optimization and Machine Learning", Addison Wesley, **(1989)**, pp. 41.

[18] X. Yan, C. Zhang and S. Zhang, "Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support", **(2008)**, Elsevier.

[19] S. Ghosh, S. Biswas, D. Sarkar and P. P. Sarkar, "Mining Frequent Itemsets Using Genetic Algorithm", International Journal of Artificial Intelligence & Applications (IJAIA), vol. 1, no. 4, **(2010)**.

[20] M. Saggar, A. K. Agrawal and A. Lad, "Optimization of Association Rule Mining using Improved Genetic Algorithms", IEEE International Conference on Systems, Man and Cybernetics, **(2004)**.

[21] D. Martin, A. Rosete and J. Alcala-Fdez, Member, IEEE, and Francisco Herrera, Member, IEEE, "A New Multiobjective Evolutionary Algorithm for Mining a Reduced Set of Interesting Positive and Negative Quantitative Association Rules", IEEE Transactions on Evolutionary Computation, vol. 18, no. 1, **(2014)** February.

Authors

Shrikant Brajesh Sagar, is pursuing M.Tech degree in CSE from the department of CSE & IT Madhav Institute of Technology & Science (MITS), Gwalior (M.P.), India. He received his B.E. degree from Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal (M.P.), India. His area of current work is the discovery of association rules in data mining and their applications.

Akhilesh Tiwari, has received the Ph.D. degree in Information Technology from Rajiv Gandhi Technological University, Bhopal (M.P.), India. He is currently working as Associate Professor in the Department of CSE & IT, Madhav Institute of Technology & Science (MITS), Gwalior (M.P.), India. He has guided several theses at Master and Under Graduate level. His area of current research includes Knowledge Discovery in Databases and Data Mining, Wireless Networks. He has published more than 20 research papers in the journals and conferences of international repute. He is also acting as a reviewer & member in the editorial board of various international journals. He is having the memberships of various Academic/ Scientific societies including IETE, CSI, GAMS, IACSIT, and IAENG.