I.J. Modern Education and Computer Science, **2016, 3, 54-63**
Published Online March 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2016.03.07

Rough Set and Genetic Based Approach for Maximization of Weighted Association Rules

Shrikant Brajesh Sagar
Department of CSE & IT, Madhav Institute of Technology and Science, Gwalior (M.P), 474005, India
E-mail: shrikantsagar19@gmail.com

Akhilesh Tiwari
Department of CSE & IT, Madhav Institute of Technology and Science, Gwalior (M.P), 474005, India
E-mail: atiwari.mits@gmail.com

Abstract—The present paper proposes a new approach for the effective weighted association rule mining. The proposed approach utilizes the power of Rough Set Theory for obtaining reduct of the targeted dataset. Additionally, approach takes the benefit for weighted measures and the Genetic Algorithm for the generation of the desired set of rules. Enough analysis of proposed approach has been done and observed that the approach works as per the expectation and will be beneficial in situation when there is a requirement for the consideration of hidden rules(maximizing generated rules) in decision-making process.

Index Terms—Weighted items**,** Rough Set Theory, Apriori Algorithm, min. w-support, min. w-confidence, weighted association rule mining, the Genetic Algorithm.

I. Introduction

Many organizations have collected large amounts of transactional data stored in databases continues to grow up fast. Instinctively, this large amount of stored transactional data contains valuable hidden knowledge, which could be used to improve the decision-making process of an organization. Data Mining is a method for searching the knowledge and implicit rules in large amounts of data. Association rules mining is to mine the association relationship from data recorded, that is, the appearance of some data items means some other data may come out. The purpose of the association rule mining is to discover association relationships during the projects from large history transaction database. Apriori Algorithm which is designed by Agrawal is the most classical frequent itemsets mining algorithm [1]. In recent years, many improvements of Apriori Algorithm have been put forward. 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 [2], [3].

Rough Set Theory is put forward in 1982 by professor Pawlak, which is a mathematical tool analyzing quantitatively to deal with imprecision, inconsistent, incomplete information and knowledge. At present, Rough Set Theory is widely used in the field of machine learning, data mining and pattern recognition [4]. Under the condition of maintaining the same classification of the knowledge base, it has removed the unnecessary knowledge. Deleting redundant attributes of the relational database can improve the clarity of potential knowledge of the information system deeply. Association rules mining based on Rough Set Theory becomes a necessary way to process problems of mining massive data. Rough Set Theory can find structured relationship of inaccurate data or noisy data, which is based on an equivalence class establishment with the given training data.

Association rule mining has been an active research area during current years. However, the conventional association rules mining algorithm works on binary attributes. This model only reflects on whether an item is present and absence in the transaction of the database, but does not consider the weight/quantity of an item within a transaction. e.g., a customer may buy 4 packets of bread and 2 packets of butter and another may buy 6 packets of milk and 3 packets of butter at an instant. These two transactions of the database will be considered the same in the usual association rule mining approach. This might bring about the loss of some vital information. Assume, e.g., that if a customer buys more than 7 packets of bread; he is likely to purchase 3 or more packets of butter. Otherwise, the purchase tendency of butter is not strong. The conventional association rule does not convey this type of association. Association rules are familiar with discovering relationships among a set of weighted items in a database.

Most of the algorithms generally mine positive association rules without paying particular consideration to negative rules. However, rules such as (A→¬B) may be importance taking into account, as they relate the presence of A to the absence of B. Negative association rules consider the same sets of items as positive association rules but, in addition, may also contain negated items within the antecedent (¬A→B) or the consequent (A→¬B) or both of them (¬A→¬B). In recent years, some researchers have proposed methods for mining positive and negative association rules from quantitative data. For extracting negative association rules, the researchers deal with two types of problems: how to efficiently search for

Copyright © 2016 MECS

interesting itemsets and how to identify interesting negative association rules.

Many evolutionary algorithms have been proposed for extracting a set of positive and negative association rules from datasets. Mostly Genetic Algorithm is considered to be one of the most successful search techniques for complex problems and have proved to be an important technique for learning and knowledge extraction. This algorithm generally considers only one evaluation condition in evaluating the quality of the generated rules. The complexity and large size of rules generated after mining have motivated researchers and practitioners to optimize the rule, for the analysis purpose. All the conventional association rule mining algorithms were developed to discover positive associations among itemsets.

Several algorithms have been developed to deal with the popular and computationally exclusive task of association rule mining. With the improvement of data mining techniques and tools, a lot of work has recently focused on the discovery of negative patterns, which can provide valuable information. Although, negative association rules mining is a difficult task, due to the fact that there is necessary dissimilarity between positive and negative association rule mining.

In this paper, proposed algorithm simplifies a large database and then generates optimized weighted association rules with multiple consequents by applying the Genetic Algorithm on the frequent itemsets generated by Apriori Algorithm.

II. Related Works

There is large amount of database increasing in the warehouses gradually. For discovering the interesting association of items from the databases, association rules mining are the dominating research area. There are various algorithms developed in previous years for discovering the frequent itemsets and desired association rules. 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 and respectively.

A. Basic Rough Set Approach

Poland mathematician Pawlak proposed the Rough Set Theory in 1982. Rough Set Theory (RST), first described by [5], is a formal approximation of a crisp set in terms of a pair of sets giving the lower and upper approximation of the original set, respectively. Recently, it has been applied in artificial intelligence [6], knowledge discovery [7], data mining, pattern recognition and machine learning [8]. The Rough Set Theory characterizes an objective approach to a deficiency in data. All calculations are performed directly on data sets. Thus, there is no need for any additional information about the data such as a probability distribution function from statistics and a degree of membership from fuzzy set theory.

Knowledge observed in RST as the partition of the universe, is prescribed as an equivalence relation in algebra. RST may be applied to consistent data to study relations between attributes. Inconsistent data sets are handled by Rough Set Theory using lower and upper approximations for each perception. These approximations are defined using accessible attributes. Moreover, definite and possible rule sets are induced from the lower and upper approximations of the concept.

In general, lots of knowledge are based on information form, Information systems generally deal with the following main steps: First, data preparation, including data discretization, data cleaning, depending on the issue of the form given information table knowledge representation system, incompatible with the object and remove redundant objects, the decision to establish the compatibility table is prepare for the data reduction. And then examine whether the conditional attribute can be omitted, get the simple attribute set, a multi-lateral compression of information table, if the information table reflects the control rules, then the equivalent of all the control rules to reduce the antecedent conditions. On this basis, on the basis of value reduction to reduce the number of properties and individuals, the final extraction rules is to access information systems inherent laws. Using Rough Set Theory for data mining, extraction of knowledge rules, the most important thing is based on rough set attribute reduction and rule redundancy value reduction.

Through some simple operations, the dimension attribute reduction, summed up the knowledge for decision support in the rules in Rough Set Theory is one of the most important applications.

In this paper, rough set and association rule mining techniques used for reduction of a decision table. In recent years, many researchers proposed efficient algorithms for reducing the data set and generation of association rules. Related to this concept, Chen Chu-xiang, et al. [9] proposed Rough Set Theory based improved Apriori Arithmetic. R_Apriori Algorithm solves the problems of Apriori Algorithm to improve the efficiency of the algorithm. XUN Jiao, XULian-cheng, QILin [10] proposed Rough Set based association rules mining algorithm. The benefit of this algorithm lies in three phases, including the removal of redundancy attributes, reducing the number of attributes, while scanning Decision Table just once can produce decision attribute sets. Aritra Roy and Rajdeep Chatterjee [11] proposed a new hybrid Rough and Fuzzy based association rule mining algorithm for generating desired association rules.

Related Concepts

Rough set theory in decision-making systems and decision rules applied to the concept of mining association rules, attribute rules can also limit; before proceeding to association rules mining, to improve the efficiency of mining association rules.Reducing the Decision table before the association rules in the database, should be handled in accordance with the following general concepts.

1. Suppose U be the universe that represents the non-empty set of all cases. If R is an equivalence relation on U, then U/R is a partition set of U. Let $[x]_R$

denote the equivalence class of R including x, or that subset x belongs to a ‘‘category’’ of set R.

2. If set R is a partition on U, equivalence relation R = $\{X_1, X_2, \ldots, X_n\}$, denoted by (U, R) is defined as an approximation space.
3. If P⊂R, then ∩P (intersections of all equivalence relation S in P) is an equivalence relation and is also an indiscernibility relation on P, denoted by ind(P).
4. Let X⊆U, and R be an equivalence relation. When X is composed of some basic category on R, we say X is R-definable, or otherwise X is R-indefinable. An R definable set is a subset of U, and can be defined exactly in the repository, called an R exact set, or contrarily, called an R rough set.
5. Suppose that repository K=(U, R) denotes all subsets X∈U and an equivalence relation R∈ind(U), then it can make a set partition on X according to the elementary sets on R.
6. Lower approximation of X is the maximal definable set of X in R:

$$R_*(X) = \cup \{Y \in \frac{U}{R} : Y \subseteq X\} \tag{1}$$

7. Upper approximation of X is the minimal definable set of X in R:

$$R^*(X) = \cup \{Y \in \frac{U}{R} : Y \cap X \neq \emptyset\} \tag{2}$$

8. R boundary set of X is defined to be $BN_R(X) = R^*(X)-R_*(X)$, while $Pos_R(X) = R_*(X)$ denotes the R positive region of X. Let $Neg_R(X) = \cup - Pos_R(X)$ be the R negative region of X and $BN_R(X)$ be the boundary region of X. We know that X is an R definable set, if the boundary region is an empty set.
9. Accuracy is defined by:

$$Dr(X) = \frac{Card(R_*(X))}{Card(R^*(X))} \tag{3}$$

Where Card(X) denotes the cardinality of set X and $X \neq \emptyset$.

10. If R is an equivalence relation and r ∈ R, when ind(R) = ind(R-r), we say r is R-dispensable, or else r is R-indispensable. When ∀ r ∈ R is R-indispensable, then R is independent.
11. Let Q be independent and ind(Q) = ind(P), then Q⊂P is a reduction of P. The core of P, denoted by Core(P), is composed of all indispensable sets in P, i.e., if Red(P) is a reduct of P, then Core(P)=∩Red(P).
12. A knowledge representation system is defined to be <U, C, D, V, f>, where U is the universal set, C∪D = A is the set of attributes, and C and D are condition and decision sets, respectively.

$$V = \cup\ a \epsilon A V_a \tag{4}$$

Where V_a denotes the domain of attribute a∈A and f: U_a × A→V is an information function which denotes the attribute value of each x in U.

Data Reduction By Rough Set Theory

In Decision Table (DT), c ∈ C denotes a condition attribute; d ∈ D denotes a decision attribute; c => d denotes the relation between c and d; this shows that c’s occurrence leads to d’s occurrence.

In this proposed work dataset reducing bases on the support of itemsets. The frequency of some item-sets in the database expressed as support(X), where X is an item-set. A big support value indicates more popularity in the database. In this paper, Transaction database can be transformed to a decision system. Decision System can be described using data tables, which takes the lines of transaction database as the object I[j] of the decision system, and take itemset of transaction as attribute sets of the decision system.

$$R_{ij} = \begin{cases} 1, & I[j] \in T[i] \\ 0, & I[j] \notin T[i] \end{cases} 0 < i \leq |I|, 0 < j \leq |T| \tag{5}$$

$$Sup(X \Rightarrow Y) = \frac{|[X]_R \cap [Y]_R|}{|T|} \tag{6}$$

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 considered 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 $[X]_R \cap [Y]_R = [X \cup Y]_R$.

B. Weighted Association Rules Mining

In this paper, proposed approach extends the traditional association rules mining problem by assigning a weight (amount) to be connected with each item in a transaction of database, to replicate the amount of every item within the transaction. Consecutively, this provides us an opportunity to associate a weight constraint with every item in a resultant association rules and called weighted association rules. For example, A[3; 1]=>B[2; 1] is a weighted association rule representing that if a customer buys item “A” in the amount between 3 and 1, he is probably to buy item “B” in the amount 2 and 1. Thus, weighted association rules improve the confidence in the association rules and give a method to do more efficient target marketing by identifying customers based on their potential degree of reliability of purchases.

The traditional model of association rule mining utilizes the support measure, which treats each transaction in the same way. In contrast, dissimilar items in dissimilar transactions have dissimilar weights in the real-life database. Hence, in this paper, formulated “weighted support” measurement in place of the “support” framework 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 new proposed model, the iterative generation and pruning of significant itemsets is justified by a "weighted downward closure property".

There are many effective algorithms for finding frequent itemsets using user-defined minimum support based weighted theory. Some of weighted based algorithms are [12], [13], [14], [15]. In most of the related work, the weighted support is computed by multiplying a support with a known weight of items. Cai, C.H., et al. [12] proposed in the year 1998 two algorithm mining association rules with weighted items. The first step of algorithms is to search for the maximum size of the large itemsets. This requires a scan of the database. Further, these algorithms are based on candidate generation and pruning techniques, in addition to the application of k-support bound property. Therefore, multiple scans of the database are required to find all weighted frequent itemsets.

Lu, S., Hu, H., and Li, F. [13] proposed an algorithm called mining weighted association rules with weighted support. In this algorithm, it is possible to generate vertical and horizontal association rules. Feng Tao, Fionn Murtagh, Mohsen Farid [14] proposed weighted association rule mining algorithm using w-support and significant framework. In this algorithm, both scalable and efficient in determining important relationships in weighted setting performed on simulated datasets. Luca Cagliero and Paolo Garza [15] proposed frequent pattern growth based infrequent weighted itemset mining algorithm, this work deal with the problem of finding irregular and weighted itemsets, i.e., the infrequent weighted itemset (IWI) mining problem. M. Sulaiman Khan, Maybin Muyeba, and Frans Coenen [16] proposed an algorithm for mining association rule from binary and fuzzy data with weighted support. Fu Jinghong, et al [17] proposed a weighted relational classification algorithm based on Rough Set. In which the relations of tables are classified in database, relational graph is converted into 0-1 matrix, the weight is calculated using UCINET; at the same time, different condition attributes are weighted differently by using attribute frequency of Rough Set.

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. Preetham Kumar, Ananthanarayana V. S. [18] 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 a weighted tree. In this algorithm "weighted support" measured as

$$\frac{\sum_j \sum_i q_{ij}}{n} \tag{7}$$

Where i = 1, 2,….k and j = 1,2,....n and q_{ij} represents a quantity of an item $i \in I$, in a j^{th} transaction and n is the number of transaction which contains weighted items.

From the literature review for computing the weighted support of weighted items, there is one possible problem is that if we are calculating the weighted support using the above-mentioned method than one item appears in a small quantity in every transaction is not a frequent item because of less than minimum w-support and another item which has large quantity in some transaction is a frequent item.

Suppose a sample database of weighted items presented in the transactional form in Table 1, which contain some conditional and decisional attributes.

The Attributes contain the weight in the form of quantity of the attributes. Let A, B, C, D, E and F are attributes of the database as shown in the Table 1:

Table 1. Database of Weighted Items

ITEM/TID	T1	T2	T3	T4	T5
A	2	3	1	0	4
B	4	0	3	3	2
C	2	3	2	2	2
D	1	1	2	1	3
E	0	0	0	2	6

From the above-mentioned method-

W-support (A) = 10/4 = 2.5
W-support (B) = 12/4 = 3
W-support (C) = 11/5 = 2.2
W-support (D) = 8/5 = 1.6
W-support (E) = 8/2 = 4

From the database observe that the total amount of items in item D and item E are same and weighted support of both items is different. Suppose user defined minimum weighted support is 2, then item D will be pruned from the frequent items but item E will remain which is not more profitable than D. It means item D is not a frequent item and item E is a frequent item with same weight. In this case, downward closure property also does not hold good.

Sun, K., Fengshan Bai [19] proposed measurement of weighted support which does not require preassigned weights. In this work link based model used for computing weighted support. Generally, our proposed method for computing weighted support based on this method.

In this paper, we are calculating weighted support for weighted items as following:

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

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

From this method-

W-support (A) = 10/49 = 0.20
W-support (B) = 12/49 = 0.24
W-support (C) = 11/49 = 0.22
W-support (D) = 8/49 = 0.16
W-support (E) = 8/49 = 0.16

Suppose user defined minimum weighted support 10%, therefore items D and E will be considered in the frequent items and if the user-defined minimum support is 20%, then both items D and E will be eliminated from the frequent items.

In this paper, the first goal is to retain the downward closure property in case of weighted items too and if weights of each individual item are same than weighted support should be same. So that frequent itemsets generates by Apriori algorithm will be more accurate and the mining efficiency will be good.

C. Genetic Algorithm

In general the major motivation for using Genetic Algorithm in the discovery of high-level prediction rules is that they perform a global search and deal with enhanced attribute relations than the greedy rule induction algorithms frequently used in data mining.

Genetic Algorithm for optimization of association rules is divided into three parts:

Section [1] discusses how to represent frequent items in the binary representation bases on the prediction (IF-THEN) rules.

Section [2] discusses how genetic operators can be modified to hold individuals representing rules.

Section [3] discusses some matter involved in the design of fitness functions for rule finding.

1. Individual Representation

Genetic Algorithms (GA) for rule discovery can be divided into two approaches, based on how association rules are encoded in the population of individuals ("chromosomes"). In the Michigan approach, each individual encodes a single prediction rule, whereas in the Pittsburgh approach each individual encodes a set of prediction rules [20].

Generally, in the Michigan approach the individuals are simpler and syntactically shorter. This has a tendency to decrease the time taken to calculate the fitness function and to simplify the design of genetic operators.

The encoding can be done in a number of ways like, binary encoding or expression encoding etc [21]. For example let's consider a rule "If a customer buys product A and B then he will also buy C, not D", which can be simply written as

If A and B then C, not D

Now, following Michigan's approach and binary encoding, for simplicity sake, this rule can be represented as

00 01 **01** 01 **10** 01**11** 00

Where, the bold digits are used as item id, like **00** for A, **01** for B, **10** for C and **11** for D and the normal digits are 00 or 01 which shows absence or presence respectively. Now this rule is ready for further computations.

2. Genetic Operators for Rule Discovery

Genetic Algorithms apply genetic operators such as selection, crossover and mutation on initial random population so as to calculate total generation of new strings. GA applies to generate solutions for succeeding generations. The possibility of an individual reproducing is proportional to the integrity of the solution it signifies. Therefore, the quality of the solutions in succeeding generations develops. The process is terminated when a suitable or optimum solution is found. The Genetic Algorithm is suitable for problems which need optimization, with respect to some computable condition. Genetic Algorithms operators as follows:

2.1 Selection

The selection of the individual member from the chromosome can be done using the Roulette Wheel selection method. Roulette Wheel selection is a process of select members from the population of chromosomes that is proportional to their fitness value.

2.2 Crossover

Crossover is performed by selecting gene randomly besides the length of the chromosomes and swapping all the genes after that point. e.g., given two chromosomes

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

2.3 Mutation

Mutation modifies the new solutions so as to add stochastic in the search for improved solutions. In this case a bit inside a chromosome will be flipped (0 becomes 1, 1 becomes 0). Whenever chromosomes are selected from the population the algorithm first verify to observe if crossover should be applied and then the algorithm iterates along the length of each chromosome mutating the bits if valid.

3. Fitness Functions for Rule Discovery

Generally the generated rules should be high predictive accuracy, comprehensible and interesting. The population is categorized by the fitness function. The Genetic Algorithm applied on the selected population from the database and computes the fitness function after every step until the Genetic Algorithm is terminated.

Let a rule be of the form: "IF A THEN B", where A is the antecedent and B is the consequent (predicted class), where A and B contain some item present or its negation.

The Fitness function is an objective function used to summarize as how close a given suggest solution is to achieving the required solution. It is very important to define a good fitness function that rewards the right kinds of individuals. The fitness function is always problem dependent. In this present work, four considerable measures of the rules such that weighted support, confidence, simplicity, and interestingness are considered. These measurements used for computing an objective fitness function with user-defined weights. By using the four measures Association Rule Mining problems can be consideration of as a Multi-objective problem instead of as a single objective one [22].

The weighted support Σ(X), of an itemset X, is defined as the amount of transaction in the dataset which contain the weighted item set.

The weighted support can be formulated as:

$$\text{W-Support} = \frac{\Sigma(\mathrm{X}\cap\mathrm{Y})}{\Sigma\mathrm{N}} \tag{9}$$

Where Σ(N) is the total number of transactions and Σ(X ∩ Y) is the numbers of transactions containing both items X and Y. Weighted Support is usually used to remove non-interesting rules.

A measure to predict the association rule accuracy is the weighted confidence or predictive accuracy. It measures the conditional probability of the consequent given the antecedent and formulated as:

$$\text{W-Confidence} = \frac{\Sigma(\mathrm{X}\cap\mathrm{Y})}{\Sigma(\mathrm{X})} \tag{10}$$

Where Σ(X) is the number of transactions of item X. A higher weighted confidence recommends a strong association between X and Y.

The discovered rule may have a large number of attributes involved in the rule makes it difficult to comprehend.

If the discovered rules are not easy and comprehensible to the user, the user will never use them. So the Comprehensibility (comp.) measure is desirable to make the discovered rules easy to understand. The comprehensibility attempts to compute appreciating of the rule. Comprehensibility of an association rule can be defined by the following expression:

$$\text{Comp.} = \frac{\log(1+|\mathrm{Y}|)}{\log(1+|\mathrm{X}\cap\mathrm{Y}|)} \tag{11}$$

Where |Y| and |X ∩ Y| are the number of attributes involved in the consequent body and the total rule respectively.

If the number of conditions in the antecedent body is less, the rule is considered as more simple. Interestingness of a rule, denoted by Interestingness X→Y is used to quantify how much the rule is surprising for the users. As the most important point of rule mining is to find some hidden information, it should discover those rules that have comparatively less happening in the database. Interestingness can be defined as

$$\text{Interestingness }(\mathrm{X}\rightarrow\mathrm{Y}) = \frac{\mathrm{W-Sup}(\mathrm{X}\cap\mathrm{Y})}{\mathrm{W-Sup}(\mathrm{X})} \times \frac{\mathrm{W-Sup}(\mathrm{X}\cap\mathrm{Y})}{\mathrm{W-Sup}(\mathrm{Y})}\left(1-\frac{\mathrm{W-Sup}(\mathrm{X}\cap\mathrm{Y})}{\Sigma(\mathrm{N})}\right) \tag{12}$$

Where Σ(N) indicate the total number of transactions in the database. As illustrated above, Association Rule Mining is considered as a Multi-objective problem rather than Single Objective one. So, the fitness function is defined as:

$$\mathrm{F} = \frac{((\mathrm{W}_1\times\mathrm{W-Sup.}) + (\mathrm{W}_2\times\mathrm{W-Conf.}) + (\mathrm{W}_3\times\mathrm{Comp.}) + (\mathrm{W}_4\times\mathrm{Interest.}))}{\mathrm{W}_1+\mathrm{W}_2+\mathrm{W}_3+\mathrm{W}_4} \tag{13}$$

As finding the frequent itemsets for any given transactional database, is of huge computational complexity, the problem of discovering association rules can be reduced to the problem of finding frequent itemsets. Therefore, in this paper the weight values of $W_1 = 4$, $W_2 = 3$, $W_3 = 2$ and $W_4 = 1$ were taken according to the relative importance of the worth measures weighted support, weighted confidence, comprehensibility and interestingness.

III. Proposed Approach

Reduction of the dataset using RST and optimization of association rules using the Genetic Algorithm, both are different kind of approach in data mining. In this proposed algorithm, Rough Set Theory (RST) and the Genetic Algorithm applied on weighted items for generation of positive and negative association rules.

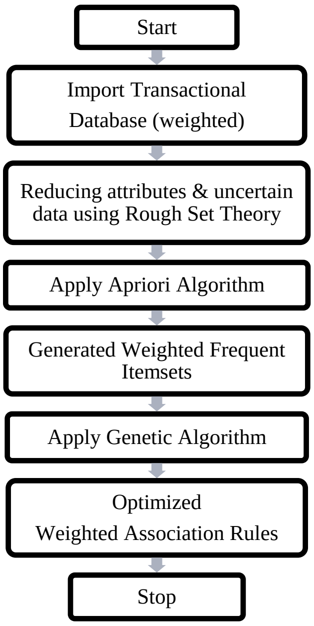
Start
Import Transactional Database (weighted)
Reducing attributes & uncertain data using Rough Set Theory
Apply Apriori Algorithm
Generated Weighted Frequent Itemsets
Apply Genetic Algorithm
Optimized Weighted Association Rules
Stop

Fig.1. Block Diagram of 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 positive and negative association rules (hidden).

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

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

$$\{C_1, C_m \epsilon L_{k-1}\}, [C_1 \cup C_m]_1, [C_1]_1 \cap [C_m]_1 \ (1 \leq m \leq T)$$

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 weighted support and weighted confidence.

The weighted support formulated as:

$$\text{W-Support} = \frac{\Sigma(X \cap Y)}{\Sigma N}$$

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. Weighted Support is usually used to remove non-interesting rules.

The weighted confidence formulated as:

$$\text{W-Confidence} = \frac{\Sigma(X \cap Y)}{\Sigma(X)}$$

Where $\Sigma(X)$ is the number of transactions of item X.A higher weighted confidence recom mends a strong association between X and Y.

14. Suppose P is set of generated frequent itemsets based on the Apriori Algorithm.
15. Set Q = θ where Q is the desired output set, which includes the association Rules.
16. Input the termination condition of the Genetic Algorithm.
17. Represent each frequent itemset of P as binary string.
18. Select the two members from the frequent itemsets using Roulette wheel sampling method.
19. Apply the crossover and mutation (if needed) on the selected members to discover the association rules.
20. Find the fitness function of each association rule (x=>y) using from the formula-

$$F= \frac{((W_1 \times \text{W}-\text{Sup.}) + (W_2 \times \text{W}-\text{Conf.}) + (W_3 \times \text{Comp.}) + (W_4 \times \text{Interest.}))}{W_1+W_2+W_3+W_4}$$

Where W_1, W_2, W_3 and W_4 are the weight values according to the relative importance of the worth measures weighted support, weighted confidence, comprehensibility and interestingness.

21. Check the following condition: If (fitness function> min. w-confidence).
22. Set Q=Q∪(x=>y)
23. If the desired number of generation is not completed then go to step 16.
24. End

Example:

1. Load transactional database (weighted items).

Table 2. 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 3. Decision Table System

Items/ TD	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

Copyright © 2016 MECS

C1, Cm∈Lk-1, [C1∪Cm], [C1] ∩ [Cm], (1≤m≤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 4. Simplified Decision Table

Items/ T-ID	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. W-Support = 20%,
Min. W-Confidence= 60%
Generated frequent itemsets with W-Support:

Table 5. 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%

7. Apply Genetic Algorithm-

7.1. Input the termination condition of Genetic Algorithm

7.2. Representation of each frequent itemset in binary string as discussed above.

Binary representation of each itemset:

AB= if A and not C then B, not E
0001**01**01**10**00**11**00

Bold digits represent item id and normal digits represent absence, presence and not consider of an item.
Therefore,

AC = **00**01**10**01**01**00**11**00
AE = **00**01**11**01**01**00**10**00
BC = **01**01**10**01**00**00**11**00
BE = **01**01**11**01**00**00**11**00
ABC = **00**01**01**01**10**01**11**00
ABE = **00**01**01**01**11**01**10**00
ABCE = **00**01**01**01**10**01**11**01

7.3. Find the fitness value of each item set using the formulae discussed above in fitness function

For AB, A=>B
W-support = 0.3529
Confidence = 1
Comprehensibility = 0.4306
Interestingness (A→B) = 1.7209
Therefore fitness = 0.6993 or 69%
Similarly
For AC, A=>C
Fitness = 0.8077 or 80%
For AE, A=>E
Fitness= 0.7514 or 75%
For BC, B=>C
Fitness= 0.9837 or 98%
For BE, B=>E
Fitness = 0.7909 or 79%
For ABC, A=>BC
Fitness= 0.6470 or 64%
For ABE, A=>BE
Fitness=0.9705 or 97%
For ABCE, A=>BCE
Fitness=0.8464 or 84%

7.4. Selection based on the roulette wheel sampling method. Select two item set which has maximum Fitness value.

7.5. Apply crossover and mutation-

Crossover:

Parent BC	**01**01**10**01|**00**00**11**00
Parent ABE	**00**01**01**01|**11**01**10**00
Offspring 1	**01**01**10**01**11**01**10**00
Offspring 2	**00**01**01**01**00**00**11**00

Mutation:

Offspring 1	**01**01**10**01**11**01**10**00
Offspring 2	**00**01**01**01**00**00**11**00
Offspring 1	**01**01**10**01**11**01**10**01
Offspring 2	**00**01**01**00**00**00**11**00

Generated itemset BCE

7.6. Fitness value of B=>CE 0.7005 or 70%.
7.7. Fitness value of BCE > min. w-confidence.

Results:

The final results generated by both the Apriori algorithm and the Genetic algorithm are shown in the Table 6 and Table 7:

Table 6. Association Rules by Apriori Algorithm

Association rules generated by Apriori Algorithm
A=>B, B=>A, A=>C, C=>A, A=>D, D=>A, B=>C, C=>B, B=>D, D=>B, A=>BC, B=>AC, C=>AB, AB=>C, BC=>A, AC=>B, A=>BD, B=>AD, D=>AB, AB=>D, AD=>B, BD=>A, A=>BCD, B=>ACD, C=>ABD, D=>ABC, AB=>CD, BC=>AD, CD=>AB, AC=>BD, ABC=>D, BCD=>A, ABD=>C, ACD=>B

Table 7. Association Rules by Genetic Algorithm

Association rules generated by Genetic Algorithm	Fitness value	w-Conf.
A=>B¬C	66%	70%
A=>CE	63%	60%
A=>¬CE	72%	70%
AC=>¬E	71%	100%
B=>CE	70%	75%
B=>¬CE	72%	75%
B=>C¬E	66%	100%
A=>B¬CE	90%	90%

IV. Conclusion

This paper proposes a new Rough Set Theory and genetic based approach for weighted association rule mining. Due to the use of rough set based concept, proposed approach considers only the reduct of the initial database. Hence, it is clear that the proposed approach works on the reduced dataset which leads to the enhancement in the performance. Furthermore, proposed approach makes use of the Genetic Algorithm, which helps in exposing invisible rules that may also be considered for a fruitful decision-making process.

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, May 1993.

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

[3] Darshan M. Tank, “Improved Apriori Algorithm for Mining Association Rules”, *I.J. Information Technology and Computer Science*, 2014, 07, 15-23, MECS, June 2014.

[4] WANG Guo-Yin, YAO Yi-Yu, YU Hong, “A Survey on Rough Set Theory and Applications” [J]. CHINESE JOURNAL OF COMPUTERS, 2009, 32(7): 1230-1246.

[5] Zdzisław Pawlak, “Rough Set Theory and its application” Journal of Telecommunication and Information Technology, 2012.

[6] T. Nishino, M. Nagamachi, H. Tanaka, “Variable precision Bayesian rough set model and its application to human evaluation data”, in: RSFDGrC, Lecture Notes in Artificial Intelligence, vol. 3641, Springer-Verlag, Berlin, 2005.

[7] Z. Pawlak, S. Andrzej, “Rough sets and Boolean reasoning”, Inform. Sci. 177 (1) (2007) 41–73.

[8] P. Pattaraintakorn, N. Cercone, K. Naruedomkul, “Rule learning: ordinal prediction based on rough sets and soft-computing”, Appl. Math. Lett 19 (12) (2006) 1300–1307.

[9] Chen Chu-xiang, Shen Jian-jing, Chen Bing, Shang Chang-xing, Wang Yun-cheng, “An Improvement Apriori Arithmetic based on Rough Set Theory” IEEE 2011.

[10] XUN Jiao, XULian-cheng, QILin, “Association Rules Mining Algorithm Based on Rough Set” IEEE International symposium on information technology in medicine and education, 2012.

[11] Aritra Roy, Rajdeep 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.

[12] Cai, et al., "Mining Association Rules with Weighted Items", Database Engineering and Applications Symposium, 1998, In Proceedings of IDEAS '98.

[13] Lu, et al., "Mining weighted association rules", Intelligent Data Analysis 5, pp.211-225, 2001.

[14] Feng Tao, Fionn Murtagh, Mohsen Farid, “Weighted Association Rule Mining using Weighted Support and Significance Framework” SIGKDD 2003.

[15] Luca Cagliero and Paolo Garza, "Infrequent Weighted Itemset Mining Using Frequent Pattern Growth", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 4, 2014.

[16] M. Sulaiman Khan, Maybin Muyeba, and Frans Coenen, “Weighted Association Rule Mining from Binary and Fuzzy Data” pringer-Verlag Berlin Heidelberg, LNAI 5077, pp. 200–212, 2008.

[17] Fu Jinghonga1, Zhang Chunyinga, Wang Jinga, Tian Fang, “A Weighted Relational Classification Algorithm Based on Rough Set”, *I.J. Education and Management Engineering*, 2013, 2, 20-19, MECS, February 2013.

[18] Preetham Kumar, Ananthanarayana V S, “Discovery of Weighted Association Rules Mining”, IEEE, 2010.

[19] Ke Sun and Fengshan Bai, “Mining Weighted Association Rules without Preassigned Weights” Knowledge and Data Engineering, IEEE Transactions on (Volume:20 , Issue: 4), Page(s): 489-495, 2008.

[20] Alex A. Freitas, "A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery" Postgraduate Program in Computer Science, Pontilicia Universidade Catolica do Parana Rna Imaculada Conceicao, 1155. Curitiba PR. 80215-901. Brazil.

[21] Manish Saggar, Ashish Kumar, Agrawal Abhimanyu Lad, “Optimization of Association Rule Mining using Improved Genetic Algorithms” International Conference on Systems, Man and Cybernetics, IEEE 2004.

[22] Mohit K. Gupta and Geeta Sikka “Association Rules Extraction using Multi-objective Feature of Genetic Algorithm”, Proceedings of the World Congress on Engineering and Computer Science 2013 Vol. II.

Authors' Profiles

Shrikant B. 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.

Dr. Akhlesh 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.