Performance Based Association Rule-Mining Technique Using Genetic Algorithm

Anubha Sharma[1], Ritu Patidar[2,] Rupali Dave[3]

[1, 2 ,3] *Assistant Professor SVITS ,RGPV University, Indore*

Abstract:-Data Mining is the process of discovering knowledge by automatically in the form of rules and patterns. Data mining mainly focuses on to discover knowledge that is accurate and interesting. Association rule mining is a well established method for obtaining correlations between data and in transactional datasets. Measures like support count, confidence, Leverage and interestingness, used for evaluating a rule can be thought of as different objectives of association rule mining problem. In the thesis work we solve the association rule-mining problem using genetic algorithm. In the present work, we use the random sampling method. A perfect sample will improve the correctness of the rules generated by the algorithm. We will test in the thesis with binary values and then compared it with categorical attributes.

Keywords: Data Mining, Association Rule Mining, Genetic Algorithm.

I INTRODUCTION

Data mining are useful in business environment for catalog design ,cross marketing where the very huge amount of data generated from cash registers ,and from specific topic based database, throughout the company analysed, explored, reduced and reused .Association, Classification, Clustering, Predictions, Sequential Patterns, and Similar Time Sequences are the most popular data mining techniques. Association rule mining , is used to find the interrelationship of a particular item in a data transaction on other items in the same transaction.

In Classification, the techniques are centers for learning diverse capacities that guide every thing of the chose information into one of a predefined set of classes. Given the arrangement of predefined classes, various traits, and a "learning (or preparing) set," the grouping routines can naturally anticipate the class of other unclassified information of the learning set.

Group examination takes unmatched information and utilizes programmed procedures to put this information into comparative or coordinated information . Bunching is unsupervised system to create the information, and does not require a learning set. It imparts a typical methodological ground to Classification. Forecast investigation is identified with relapse methods. The key thought of forecast examination is to find the relationship between the reliant and autonomous variables, the relationship between the free variables. Successive Pattern investigation used to discover comparative examples in information exchange over a business period.

II PROBLEM STATEMENT

Existing algorithms for mining association rules are mainly worked on a binary database, termed as market basket database. On preparing the market basket database, every record of the original database is represented as a binary record where the fields are defined by a unique value of each attribute in the original database. The fields of this binary database are often termed as an item. For a database having a huge number of attributes and each attribute containing a lot of distinct values, the total number of items will be huge. Storing of this binary database, to be used by the rule mining algorithms, is one of the limitations of the existing algorithms.

Another aspect of these algorithms is that they work in two phases. The first phase is for frequent item-set generation. Frequent item-sets are detected from all-possible item-sets by using a measure called support count (SUP) and a user-defined parameter called minimum support. Support count of an item set is defined by the number of records in the database that contains all the items of that set. If the value of minimum support is too high, number of frequent item sets generated will be less, and thereby resulting in generation of few rules. Again, if the value is too small, then almost all possible item sets will become frequent and thus a huge number of rules may be generated. Selecting better rules from them may be another problem. After detecting the frequent item-sets in the first phase, the second phase generates the rules using another user-defined parameter called minimum confidence [1] and [2] and [3].

III BACKGROUND

Association Rules

Association rules are if and then statements that help uncover relationships between seemingly unrelated data in a relational

database or other information repository. An association rule has two parts, an antecedent (if) and a consequent (then). Association rule is expressed as X=>Y, where X is the antecedent and Y is the consequent. Each association rule has two quality measurements, support and confidence. Support implies frequency of occurring patterns, and confidence means the strength of implication [1-3] and [9].

Genetic Algorithm

Genetic Algorithm (GA) was developed by Holland in 1970. This incorporates Darwinian evolutionary theory with sexual reproduction. GA is stochastic search algorithm modeled on the process of natural selection, which underlines biological evolution. GA has been successfully applied in many search, optimization, and machine learning problems. GA process is an iteration manner by generating new populations of strings from old ones. Every string is the encoded binary, real etc., version of a candidate solution. An evaluation function associates a fitness measure to every string indicating its fitness for the problem. Standard GA apply genetic operators such selection, crossover and mutation on an initially random population in order to compute a whole generation of new strings.

The most important biological terminology used in a genetic algorithm is:

- The chromosomes are elements on which the solutions are built.
- Population is made of chromosomes.
- Reproduction is the chromosome combination stage.
- Mutation and crossover are reproduction methods.
- Quality factor (fitness) is also known as performance index, it is an abstract measure to classify chromosomes.
- The evaluation function is the theoretical formula to calculate a chromosome's quality factor [4] and [6] and [10].

Genetic Operators:

The GA maintains a population of n chromosomes (solutions) with associated fitness values. Parents are selected to mate, on the basis of their fitness, producing offspring via a reproductive plan (mutation and crossover). Consequently highly fit solutions are given more opportunities to reproduce (selected for next generation), so that offspring inherit characteristics from each parent. As parents mate and produce offspring, room must be made for the new arrivals since the population is kept at a static size (population size). In this way it is hoped that over successive generations better solutions will thrive while the least fit solutions die out. The representation scheme, Population Size, Crossover rate, Mutation rate, and fitness function and selection operator are the GA operators [4-6].

Genetic Algorithm for Association Rule Mining

Genetic Algorithm (GA) is an adaptive heuristic search algorithm based on the evolutionary ideas of natural selection and genetics. The evolutionary process of a GA is a highly simplified and stylized simulation of the biological version. It starts from a population of individuals randomly generated according to some probability distribution, usually uniform and updates this population in steps called generations. In each generation, multiple individuals are randomly selected from the current population based upon some application of fitness, bred using crossover, and modified through mutation to form a new population [7] and [8].

IV RELATED WORKS

Manish Saggar, Ashish Kumar Agarwal and Abhimunya Lad et. al. proposed to optimize the rules generated by Association Rule Mining (apriori method), using Genetic Algorithms. In general the rule generated by Association Rule Mining technique do not consider the negative occurrences of attributes in them, but by using Genetic Algorithms (GAs) over these rules the system can predict the rules which contains negative attributes. The main motivation for using GAs in the discovery of high-level prediction rules is that they perform a global search and cope better with attribute interaction than the greedy rule induction algorithms often used in data mining. The improvements applied in GAs are definitely going to help the rule based systems used for classification.

In this paper the authors have tried to use the enormous robustness of GAs in mining the Association Rules. The results generated when the technique applied on the synthetic database, includes the desired rules, i.e. rules containing the negation of the attributes as well as the general rules evolved from the Association Rule Mining. [2].

Anandhavalli M, Suraj Kumar Sudhanshu, Ayush Kumar and Ghose M.K et. al. is to find all the possible optimized rules from given data set using genetic algorithm. The rule generated by association rule mining algorithms like priori, partition, pincer-search, incremental, border algorithm etc, does not consider negation occurrence of the attribute in them and also these rules have only one attribute in the consequent part. By using Genetic Algorithm (GAs) the system can

predict the rules which contain negative attributes in the generated rules along with more than one attribute in consequent part. The major advantage of using GAs in the discovery of prediction rules is that they perform global search and its complexity is less compared to other algorithms as the genetic algorithm is based on the greedy approach.

They have dealt with a challenging association rule mining problem of finding optimized association rules. The frequent itemsets are generated using the Apriori association rule mining algorithm. The genetic algorithm has been applied on the generated frequent itemsets to generate the rules containing positive attributes, the negation of the attributes with the consequent part consists of single attribute and more than one attribute.

The results reported in this paper are very promising since the discovered rules are of optimized rules [3].

Farah Hanna AL-Zawaidah, Yosef Hasan Jbara and Marwan AL-Abed Abu-Zanona et. al. presented a novel association rule mining approach that can efficiently discovered the association rules in large databases. The proposed approach is derived from the conventional Apriori approach with features added to improve data mining performance. They had performed extensive experiments and compared the performance of the algorithm with existing algorithms found in the literature. Experimental results show that the approach outperforms other approaches and show that approach can quickly discover frequent item sets and effectively mine potential association rules.

The developed approach adopts the philosophy of Apriori approach with some modifications in order to reduce the time execution of the algorithm. First, the idea of generating the feature of items is used and; second, the weight for each candidate item set is calculated to be used during processing. The feature array data structure is built by storing the decimal equivalent of the location of the item in the transaction. Transforming here means reorganizing and transforming a large database into manageable structure to fulfill two objectives: (a) reducing the number of I/O accesses in data mining, and (b) speeding up the mining process. There is one mandatory requirements for the transforming technique, that the transaction database should be read only once within the whole life cycle of data mining. By storing the appearing feature of each interested item as a compressed vector separately, the size of the database to be accessed can be reduced greatly.

This paper is to improve the performance of the conventional Apriori algorithm that mines association rules by presenting fast and scalable algorithm for discovering association rules in large databases. The approach to attain the desired improvement is to create a more efficient new algorithm out of the conventional one by adding new features to the Apriori approach. The proposed mining algorithm can efficiently discover the association rules between the data items in large databases. In particular, at most one scan of the whole database is needed during the run of the algorithm. Hence, the high repeated disk overhead incurred in other mining algorithms can be reduced significantly. They compared our algorithm to the previously proposed algorithms found in literature. The findings from different experiments have confirmed that the proposed approach is the most efficient among the others. It can speed up the data mining process significantly as demonstrated in the performance comparison. Furthermore, gives long maximal large itemsets, which are better, suited to the requirements of practical applications. They demonstrated the effectiveness of the algorithm using real and synthetic datasets. They developed a visualization module to provide users the useful information regarding the database to be mined and to help the user manage and understand the association rules [4].

Peter P. Wakabi-Waiswa and Dr. Venansius Baryamureeba et. al. present a Pareto based multi objective evolutionary algorithm rule mining method based on genetic algorithms. They used confidence, comprehensibility, interestingness, surprise as objectives of the association rule mining problem. Specific mechanisms for mutations and crossover operators together with elitism have been designed to extract interesting rules from a transaction database. Empirical results of experiments carried out indicate high predictive accuracy of the rules generated.

In this paper deal with the ARM problem as a multi-objective problem rather than as a single one and try to solve it using multi-objective evolutionary algorithms with emphasis on genetic algorithms (GA). The main motivation for using GAs is that they perform a global search and cope better with attribute interaction than the greedy rule induction algorithms often used in data mining tasks. Multi-objective optimization with evolutionary algorithms is well discussed.

The scheme applied here for encoding/decoding the rules to/from binary chromosomes is that the different values of the attributes were encoded and the attribute names are not. For encoding a categorical valued attribute, the market basket encoding scheme is used. For a real valued attribute their binary representation can be used as the encoded value. The range of values of that attribute will control the number of bits used for it. The archive size is fixed, i.e., whenever the number of non-dominated individuals is less than the

predefined archive size, the archive is filled up by dominated individuals. Additionally, the clustering technique used does not loose boundary points.

The proposed algorithm was tested on a dataset drawn from the UCI repository of machine learning databases. For brevity, the data used is of a categorical nature.

In this paper they had dealt with a challenging NP-Hard association rule mining problem of finding interesting association rules. The results of this paper were good since the discovered rules are of a high predictive accuracy and of a high interesting value [5].

V PROPOSED METHODOLOGY

Association Rule Mining is computationally and Input/Output serious. The quantity of tenets becomes exponentially with the quantity of things. Since information is expanding as far as the measurements (number of things) and size (number of exchanges), one of the primary characteristics required in an Association Rule Mining calculation is adaptability: the capacity to handle gigantic information stores. Successive calculations can't give adaptability, as far as the information measurement, size, or runtime execution, for such substantial databases.

In the present work we will take care of the association rule mining issue with hereditary calculation. The principal errand for this is to speak to the conceivable principles as chromosomes, for which a suitable encoding/interpreting plan is required. For this, two methodologies can be received. In the past methodology every chromosome speaks to an arrangement of tenets, and this methodology is more suitable for characterization principle mining; as we don't need to interpret the resulting part and the length of as far as possible the quantity of guidelines created. Alternate methodologies where every chromosome speaks to a different standard. We need to encode the forerunner and resulting parts independently; and hence this perhaps a productive route from the purpose of space use since we need to store the void conditions as we don't known from the earlier which qualities will show up in which part. So we took after another methodology that is superior to this methodology from the purpose of capacity necessity. With every property we relate two additional label bits. In the event that these two bits are 00 then the ascribe beside these two bits shows up in the precursor part and on the off chance that it is 11 then the characteristic shows up in the ensuing part. What's more, the other two mixes, 01 and 10 will show the nonattendance of the characteristic in both of these parts. Along these lines we can deal with variable length rules with more stockpiling productivity, including just an overhead of bits.

The following step is to locate a suitable plan for encoding/translating the tenets to/from twofold chromosomes. Subsequent to the positions of properties are altered, we require not store the name of the qualities. We need to encode the estimations of various property in the chromosome just. For encoding a straight out esteemed property, the business sector wicker container encoding plan is utilized. As talked about before this plan is not suitable for numeric esteemed qualities. For a genuine esteemed quality their parallel representation can be utilized as the encoded esteem. The scope of estimation of that property will control the quantity of bits utilized for it. Disentangling will be essentially the opposite of it. The length of the string will rely on upon the required precision of the quality to be encoded. Interpreting can be executed as:

Value = Minimum value - (most extreme value - least value) X ((∑ (2i-1 X ith bit value))/(2n - 1))

Where 1<= i <= n and n is the quantity of bits utilized for encoding; and least and most extreme are least and greatest estimations of the property.

Utilizing these encoding plans estimations of various properties can be encoded into the chromosomes. Subsequent to in the affiliation leads a quality might be included with various social administrators, it is ideal to encode them additionally inside of the guideline itself.

To handle this circumstance we utilized another piece to show the administrators included with the characteristic. Equity and not balance are not considered with the numerical property. Along these lines the entire guideline can be spoken to as a twofold string, and this parallel string will speak to one chromosome or a conceivable principle.

Subsequent to getting the chromosomes, different hereditary administrators can be connected on it. Vicinity of vast number of qualities in the records will brings about huge chromosomes, accordingly requiring multi-point hybrid. There are a few troubles to utilize the standard multi-target GAs for affiliation principle mining issues. If there should be an occurrence of tenet mining issues, we have to store an arrangement of better guidelines found from the database. In the event that we take after the standard hereditary operations just, then the last populace may not contain a few decides that are better and were produced at some middle eras. It is ideal to keep these tenets. For this errand, a different populace is utilized. In this populace no hereditary operation is performed. It will essentially contain just the non-overwhelmed chromosomes of the past era. The client can settle the measure of this populace. Toward the end of original, it will contain the non-commanded chromosomes of

the original. It will simply contain only the non- dominated chromosomes of the previous generation. The user can fix the size of this population. At the end of first generation, it will contain the non-dominated chromosomes of the first generation. After the next generation, it will contain those chromosomes, which are non-dominated among the current population as well as among the non-dominated solutions till the previous generation.

Definitions:

Support:-

The rule X ⇒ Y holds with support s if s% of transactions in D contains X ∪ Y. Rules that have a s greater than a user-specified support is said to have minimum support.

Confidence:-

The rule X ⇒ Y holds with confidence c if c% of the transactions in D that contain X also contain Y. Rules that have a c greater than a user-specified confidence is said to have minimum confidence.

Itemset:-

An item set is a set of items. A k-itemset is an itemset that contains k number of items.

Frequent item set:- This is an itemset that has minimum support.

Candidate set:- This is the name given to a set of item sets that require testing to see if they fit a certain requirement.

Chromosome: - A chromosome (also sometimes called a genome) is a set of parameters which define a proposed solution to the problem that the genetic algorithm is trying to solve. The chromosome is often represented as a simple string, although a wide variety of other data structures are also used. We have to redefine the Chromosome representation for each particular problem, along with its fitness, mutate and reproduce methods.

Fitness:- Fitness (often denoted ω in population genetics models) is a central idea in evolutionary theory. It can be defined either with respect to a genotype or to a phenotype in a given environment. In either case, 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; the alleles with higher fitness become more common.

The most important part of Genetic Algorithm is a design of Fitness Function:

f (x) = M/

Where M= Support (x) with condition is support (x) > Minsupport

And N= support (x) with condition is support < minsupport

Support is the Support of New rule generated through genetic operation. Normal case the value of (Support(x) < minsupport) is rejected for the better performance of genetic algorithm. We have used class-learned classifier for the prediction for rejected those value near to the Maximum value.

VI PROPOSED APPROACH

Step 1: Load a specimen of records from the database that fits in the memory.

Step 2: Generate N chromosomes haphazardly.

Step 3: Decode them to get the estimations of the distinctive properties.

Step 4: Scan the stacked specimen to discover the backing of predecessor part, resulting part and the guideline.

Step 5: Find the certainty, conceivability and interestingness values.

Step 6: Rank the chromosomes relying upon the non-strength property.

Step 7: Assign wellness to the chromosomes utilizing the positions, as specified prior.

Step 8: Bring a duplicate of the chromosomes positioned as 1 into a different populace, and store them in the event that they are non-commanded in this populace too. In the event that a portion of the current chromosomes of this populace get to be commanded, because of this insertion, then expel the overwhelmed chromosomes from this populace.

Step 9: Select the chromosomes, for cutting edge, by roulette wheel determination plan utilizing the wellness figured as a part of.

Step10: Replace all chromosomes of the old populace by the chromosomes chose in Step 9.

Step11: Perform multi-point hybrid and change on these new people.

Step12. In the event that the fancied number of eras is not finished, then go to Step 3.Otherwise next step.

Step13. Unravel the chromosomes in the last put away populace, and get the created rules.

VII CONCLUSION

To discover association rules is the heart of data mining. Mining for association rules between things in vast database of offers exchanges has been perceived as a critical territory of database research. These rules can be successfully used to reveal obscure connections, creating results that can give a premise to determining and decision making. This research work utilizes an association rule based genetic algorithm to solve the multi-objective rule mining problem using three measures comprehensibility, interestingness and the predictive accuracy. We discuss a way to deal with speak to the guidelines as chromosomes, where every chromosome speaks to a different tenet. To enhance the productivity of this calculation, some refinement might be required. For instance, this calculation takes a shot at an example of the first database, and the specimen may not really mirror the genuine database. In the present work, we utilize the irregular inspecting strategy. An immaculate example will enhance the accuracy of the rules produced by the calculation. In addition, we will test the methodology just with the numerical esteemed traits. It must be tried with the straight out characteristics too.

REFERENCES

[1] Farah Hanna AL-Zawaidah, Yosef Hasan Jbara and Marwan AL-Abed Abu-Zanona, “An Improved Algorithm for Mining Association Rules in Large Databases”, World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 1, No. 7, 2011, pp. 311-316.

[2] Peter P. Wakabi-Waiswa and Dr. Venansius Baryamureeba, “Extraction of Interesting Association Rules Using Genetic Algorithms”, Advances in Systems Modelling and ICT Applications, pp. 101-110.

[3] M. Ramesh Kumar and Dr. K. Iyakutti, “Genetic algorithms for the prioritization of Association Rules”, IJCA Special Issue on “Artificial Intelligence Techniques - Novel Approaches & Practical Applications” AIT, 2011, pp. 35-38.

[4] 1S. Dehuri,A. K. Jagadev, A. Ghosh and R. Mall, “Multi-objective Genetic Algorithm for Association Rule Mining Using a Homogeneous Dedicated Cluster of Workstations”, American Journal of Applied Sciences 3 (11), 2006, pp. 2086-2095.

[5] Rupali Haldulakar and Prof. Jitendra Agrawal, “Optimization of Association Rule Mining through Genetic Algorithm”, International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 3 Mar 2011, pp. 1252-1259.

[6] Indira K1 and Kanmani S, “Performance Analysis of Genetic Algorithm for Mining Association Rules”, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012, pp. 368-376.

[7] Sufal Das and Banani Saha, “Data Quality Mining using Genetic Algorithm”, International Journal of Computer Science and Security, (IJCSS) Volume (3) : Issue (2), pp. 105-112.

[8] F.Q.Shi, S.Q.Sun, and J.Xu, “Association rule mining of Dansei knowledge based on rough set,” Computer Integrated Manufacturing Systems, vol.14, 2008. pp.407-411.

[9] Manish Saggar, Ashish Kumar Agarwal and Abhimunya Lad, “Optimization of Association Rule Mining using Improved Genetic Algorithms”IEEE 2004.

[10] Ansaf Salleb-Aouissi, Christel Vrain andCyril Nortet, “QuantMiner: A Genetic Algorithm for Mining Quantitative Association Rules”, IJCAI-2007, pp. 1035-1040.