An Optimized Association Rule Mining using Genetic Algorithm

Dimple S. Kanani
ME Scholar
Computer science and engineering
Parul Institute of technology, India

Shailendra K. Mishra
Asst Professor
Computer science and engineering
Parul Institute of technology, India

ABSTRACT

Association Rule Mining technique that attempt to unearthing interesting pattern or relationship between data in large Database. Genetic Algorithm is a search heuristic which is used to generate useful solution for optimization and search problems. Genetic Algorithm based evaluation in Mining Technique is backbone for mining interesting Rule based on GA parameters like fitness function, Crossover Rate, Mutation Rate. The key focus of this synthesize approach is to optimize the rule that generated by mining methodology and to provide more accurate results. The Proposed Approach is to generate rules based on Quantitative dataset, using the concept of threshold - frequent item sets are define as initial population which the first step of Genetic algorithm. Crossover & mutation is applied to generate more combination of rule & can identify Co-occurrence of item sets.

Keywords

Association rule Mining, Genetic Algorithm, Data mining, and Optimization.

1. INTRODUCTION

Data Mining

Data Mining is the process of extracting information from large data sets through the use of algorithms and techniques drawn from the field of Statistics, Machine Learning and Data Base Management. Data mining is the process of discovering meaningful knowledge from massy data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques. Data mining techniques can be classified into the following categories: classification, clustering, association rule mining, sequential pattern analysis, prediction, data visualization etc. Data mining, also Define as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. Data mining Tools Search databases for hidden patterns, find information that users may miss. Data mining is the part of knowledge discovery process: Data Cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Pattern Evaluation, and Knowledge Representation.

Technologies in Data mining

Rapid advance in data capture, transmission and storage, Data Mining is Associated with many other technologies that will help to implement new and innovative ways to mine the after-Market value of their vast stores of detailed & massy data. The most commonly used techniques in data mining are:

- Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

- Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset.

- Genetic algorithms: Optimization techniques that proceed as genetic combination, mutation, and natural selection. Mostly used in classification & association rule extraction.

- Nearest neighbor method: A technique of classifying each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset. It also called as the k-nearest neighbor technique.

- Rule induction: The if-then rules from data based on statistical significance.

Association Rule Mining

Association rule mining is one of the important tasks of data mining intended towards decision support. Basically it is the process of finding some relations among the attributes' values of a huge database. Finding Association Rule is to find Co-occurrence Relationships is called associations. It was introduced in 1993 by Agrawal.

Discovery of Togetherness or connection among objects help in taking some decisions. Let D be the database of transactions and J = {J1... Jn} be the set of items. A transaction T includes one or more items. An association rule has the form $X \Rightarrow Y$, where X and Y are non-empty sets of items (i.e. $X \subseteq J$, $Y \subseteq J$) such that $X \cap Y = \emptyset$[6]. These relationships can be represented as an IF–THEN statement. IF <some conditions are satisfied> THEN <predict some values of other attribute(s)>. The conditions associated in the IF part is termed as Antecedent and those with the THEN part is called the Consequent. Here item sets refer as X and Y, respectively. So, symbolically we can represent this relation as $X \rightarrow Y$ and each such relationship that holds between the attributes of records in a database fulfilling some criteria are termed as an association rule.

To extract interesting rules from all possible rules, there are two basic measures: support & confidence. The threshold of support & confidence are predefined by user to drop those rules that are not so interesting or useful. The rules that satisfy both min_support threshold & min_confidance threshold are strong rule.

Support: The rule X ⇒ Y holds with support s if s% of transactions in D contains X ∪ Y. Rules that have a 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.

Genetic Algorithm:

Genetic Algorithm is a search heuristic that mimic the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic Algorithm is based on ideas of evolution theory (Holland, 1975) as key principle is that only the fittest entities survive. The genetic algorithms are important when discovering association rules because they work with global search to discover the set of items frequency and they are less complex than other algorithms often used in data mining. The genetic algorithms for discovery of association rules have been put into practice in real problems such as commercial databases, biology and fraud detection event sequential analysis.

Genetic Algorithm work in this flow:

The functions of genetic operators are as follows:-

1) Selection: Selection deals with the probabilistic survival of the fittest, in that, more fit chromosomes are chosen to survive. Where fitness Function is a comparable measure of how well a chromosome solves the problem at hand.

2) Crossover: This operation is performed by selecting a random gene along the length of the chromosomes and swapping all the genes after that point[9].

3) Mutation: Mutation changes randomly the new offspring. For binary encoding we can switch a few randomly chosen bits from 1 to 0 or from 0 to 1[9].

2. RELATED WORK

Data is defined as collection of facts such as number, words, measurements, Observation or description of things. Base on Format of data two categories are there, Qualitative data which display descriptive information and Quantitative data which deal with Numerical data. Quantitative data can be represented as in form of discrete (group of value, take certain values, ex: “he has 4 legs”) or Continuous (take any value within a range, ex: “he weights 25.5 kg”) Association rules can deal with categorical and numerical attributes [3]. If attributes are categorical generated rules are define as Boolean rules – shows presence “buys A” or absence of items “does not buy A”. If attributes are numerical generated rules are define as quantitative rules. It is easy to deal with categorical attributes because its rule only work with yes/ no, 0/1, present/absent while numerical attribute having many values so it need to categorise via discretization or distribution. It also denoted as Quantitative Rule mining Methodology [1].

1) Discretization-Based Approach: Reduction in the number of values of continuous attribute by dividing the range of the attribute into intervals, Replacement of actual values in to Interval labels will make process more simpler, As for Instance: Age [12, 20] ⇒ Weight [18, 40].

2) Distribution-Based Approach [1]: Right-hand side of a rule express the distribution of the values of numeric attributes such as the mean or variance, like Sex=male ⇒ Height: mean = 179 ∧ Weight: mean = 75.

3) Optimization-Based Approach: In this approach, numeric attributes are optimized during the mining process. The term optimization was first used by Fukuda [1]. Genetic algorithm & particle swam optimization technique used to achieve Optimum Rules.

Numerous works have been carried out using genetic Algorithm for Association Rule mining. This section Describe the Work dine in Related Field.

Rupali Haldulakar and Prof. Jitendra Agrawal[15] Proposed a novel method for generation of strong rule. In which a general Apriori algorithm is used to generate the rules after that they used the optimization techniques. Genetic algorithm is one of the best ways to optimize the rules .In this direction for the optimization of the rule set the design a new fitness function that uses the concept of supervised learning then the GA will be able to generate the stronger rule set.

Anandhavalli M, Suraj Kumar Sudhanshu, Ayush Kumar and Ghose M.K [12] 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 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.

M. Ramesh Kumar and Dr. K. Iyakutti [9] a novel genetic algorithm based association rule mining algorithm is discussed in this paper. Prioritization of the rules has been discussed with the help of genetic algorithm. Fitness function is designed based on the two measures like all confidence and the collective strength of the rules, other than the classical support and the confidence of the rules generated. The algorithm is been tested for the four data sets like Adult, Chess, Wine, Zoo. They presented a novel algorithm for the rule prioritizing, generated by the apriori algorithm through the application of genetic algorithm. The fitness function is designed based on the user’s interesting measure and M is the threshold value of the interesting measure considered.

R. O. Oladele, J. S. Sadiku [18] Selection is one of the key operations of genetic algorithm (GA). This paper presents a comparative analysis of GA performance in solving multi-objective network design problem (MONDP) using different parent selection methods. Three problem instances were tested and results show that on the average tournament selection is the most effective and most efficient for 10-node network design problem, while Ranking & Scaling is the least effective and least efficient. For 21-node and 36-node network problems, Roulette Wheel is the least effective but most efficient while Ranking & Scaling equals and outperformed tournament in effectiveness and efficiency respectively.

Noraini Mohd Razali, John Geraghty [19] A genetic algorithm (GA) has several genetic operators that can be modified to improve the performance of particular implementations. These operators include parent selection, crossover and mutation. Selection is one of the important operations in the GA process. There are several ways for selection. This paper presents the comparison of GA performance in solving travelling salesman problem (TSP) using different parent selection strategy. Several TSP instances were tested and the results show that tournament selection strategy outperformed proportional roulette wheel and rank-based roulette wheel selections, achieving best solution quality with low computing times. Results also reveal that tournament and proportional roulette wheel can be superior to the rank-based roulette wheel selection for smaller problems only and become susceptible to premature convergence as problem size increases.

3. PROBLEM STATEMENT

Association rules are represent as X→Y, where antecedent part X (left side) of the rule and consequent part Y (right side). So it represents that if X is present, then Y will be present. Co-occurrence of Item sets are define based on main 2 measures ‘Support’ and ‘Confidence’ to Extracting rule. This Mining technique contains two phase: First one includes extraction of total Frequent Item Sets. And second includes rule generation that follows minimum confidence.

- This can not guarantee to obtain unsuitable Association rules and obtain rules should be interesting and Comprehensible too, for having appropriate coverage and reliability of rule. As number of objective increase will be easy to find optimal solution in larger search space. So, its triggers numerical association rule mining towards Multi-Objective rather than single objective.

-Rules that generated by simple Association rule mining algorithm not considering negation of Item set like “If Customer will choose A but not B then will choose C” and also not able to handle more than one attribute in consequent part. But this can be handling very well by **Genetic Algorithm** [7].

-Numeric valued attribute have many values so it’s harder to deal with this compare to Boolean valued attribute so for **Numeric Valued Attribute** need to approach Distribution of value to fetch Association Rules.

4. PROPOSED STATEMENT

“Multi-Objective Genetic Algorithm approach for mining Quantitative Association rules based on Distribution Approach”. Several measures are defined inorder to determine more efficient rules. Three measures: confidence, interestingness, and comprehensibility have been used as different objectives for multi objective optimization which is amplified with genetic algorithms approach. Finally, the best rules are obtained Numerical dataset.

As Working with Quantitative dataset it is more complex then the categorical dataset so at first need to generate threshold (based on maximum value) for each Attribute and which records are fulfil the minimum support are define as initial frequent item set – this is initial population to work with genetic Algorithm. Here three objectives are considered Comprehensibility, interestingness, Confidence so generated rules are Define as Multi-Objective Association rules. These Objective help to minimize Search Space for Fitness function & some previously unknown, easily understandable and compact rules can be generated. Non-dominance property of the chromosomes is determined by three measures confidence, interestingness and comprehensibility. A highest value of these three measures is the fitness function. So, mathematically fitness function is Average of three objectives. Highest the fitness function will denoted by first rank and then incrementally. First ranked record from dataset will define as Non- dominated solution and others are as dominated solution. Single point Crossover & mutation is applied on dominated solution to generate more suitable combinations & finally rules are generated.

Here, *Confidence*: is define as the rule $X \Rightarrow Y$ having confidence $C= \sigma\ (X\cup Y)/\sigma(X)$. Define that number of transactions that contains item set X &Y both with respect to total number of transactions X.

Comprehensibility shows that generated rules are easy to understand means it is the simplicity measure. Comp $=\log(1+ Y\ )/\log(1+\ X\cup Y\ )$.The rule may have a large number of attributes involved in the rule having single attribute at antecedent part is generally consider as simpler rule [1].

Interestingness states that knowledge discovery should be subjective, all rules are not important but only those that are ‘interesting’ based on quantifiable objective like validity of Pattern are important, Means Process of finding rule that occurs less but are useful for user. $X\rightarrow Y=$ [Sup $(X\cup Y)/\text{Sup}(X)] \times [(\text{Sup}\ (X\cup Y)/\text{Sup}(Y))\ (1-\text{Sup}\ (X\cup Y)/\sigma\ (N))]$. Here, [Sup $(X\cup Y)/\text{Sup}(X)$] defines Probability of generating rules based on antecedent part. (Sup $(X\cup Y)/\text{Sup}(Y)$) defines Probability of generating rules based on consequent part. ($1-\text{Sup}\ (X\cup Y)/\sigma\ (N)$) defines Probability of not generating rules based on whole dataset [1].

Algorithm:

Step 1: Load Database in .xlsx or .csv type.

Step 2: Apply Distribution Approach for numeric valued Attribute.

Step 3: Calculate:

Comprehensibility = log (1+ |C|) / log(1+|AUC|) [1].

Interestingness = [SUP(AUC) / SUP(A)] x [SUP(AUC) / SUP(C)] x[1- SUP(AUC) / SUP(D)][1].

Confidence = sup (XUY) / sup (X)[1].

Step 4: fitness: if (Com(x) > Comprehensibility, Int(x) > Interestingness, C(x) > min confidence).

Step 5: Set rank to the based on highest fitness.

Step 6: Rank with highest fitness → Define as non-dominated solutions.

Step 7: Perform Crossover on current Population.

Step 8: Perform Mutation on new Population.

Proposed Work Flow:

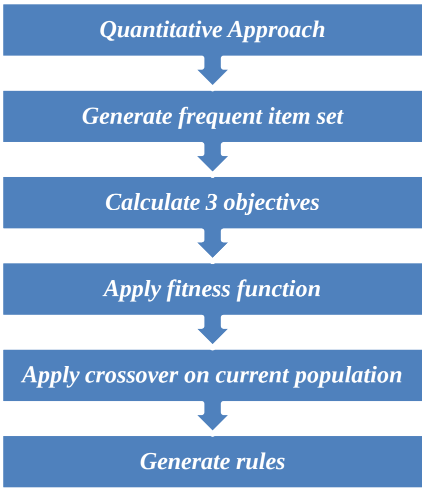

Fig1: Proposed work flow

Parameter	Values
Crossover Probability	0.8
Mutation Probability	0.01
No. of Iteration	400

Table 1: parameters setting for Genetic Algorithm

Outcome of Proposed Work, which represents distribution-based approach for numeric valued attributes using Multi objective genetic algorithm for Association rule mining. Generated rule are like:
Percent body fat using siri's equation=2.37 → density=36.08, age=175.91, weight (lbs) = 11.42, height (inches) = 8.89.

Graph that represents the no. of Generated rule based on no. of iterations:
As per genetic Algorithm termination condition is user specified which is number of iteration that will perform crossover and mutation on dominated solution. Here is the Pictorial representation that display generated number of rule according to user specified iterations.

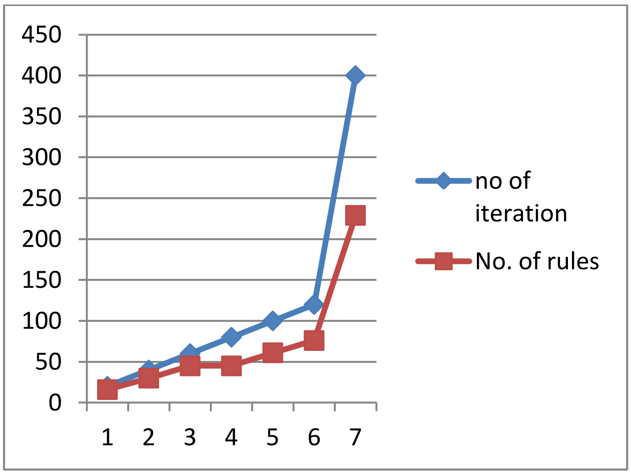

Fig 2: no of generated rules vs. no of Iteration

5. CONCLUSION

Genetic Algorithm solves multi-objective Rule mining problem can Work on Numerical Attribute. A number of works are already published in this field, this is the modified approach that shows enormous robustness of genetic algorithm in mining Association rules –where here fitness function is key to find rank based accurate rule, selection method play major role in reducing execution time by selecting parents for reproduction, Different selection mechanisms work well under different situations. Crossover and Mutation rate avoid premature Convergence of algorithm. Finally, optimal rules are generated that is based on Distribution approach for numeric valued Attribute. In future work can be extended by Incorporation of other interesting measure that will help to discover more interesting rule and applicable to real time application & Business values that can be tested based in generated rules. Distributed computing can be approached for complexity reduction purpose.

6. ACKNOWLEDGEMENT

I'm extremely obliged to my guide Mr. Shailendra Mishra devoid of his guidance the work would not have happened and He supported me to solve my difficulties arise & give Valuable suggestions. I would like to pay my sincere gratitude for his endless motivation & support in progress and success of this Survey work.

7. REFERENCES

[1] B. Minaei-Bidgoli, R. Barmaki, M. Nasiri, "Mining numerical association rules via multi-objective genetic algorithms", 2013 Elsevier, 15–24.

[2] 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, October 2013, ISSN: 2078-0958, VOL II,pp.23-25.

[3] J. Malar Vizhi And Dr. T. Bhuvaneswari, "Data quality measurement on Categorical Data using Genetic Algorithm" International journal of Data Mining & Knowledge Management process (ijdkp) VOL.2, No.1, January 2012, pp.33-42.

[4] Indira K and Kanmani S, "Performance Analysis of Genetic Algorithm for Mining Association Rules" ijcsi international journal of computer science, ISSN: 1694-0814, vol. 9, issue 2, no 1, march 2012.

[5] Sanat Jain, Swati Kabra, "Mining & Optimization of Association rules using Effective Algorithm" international journal of emerging technology and advanced engineering, ISSN: 2250-2459, volume 2, issue 4, April 2012, pp.281-285.

[6] Deep Hakani1 Harshita Kanani, "Association Rule Optimize by Multi Objective Genetic Algorithm" IJSRD - International Journal for Scientific Research & Development, ISSN 2321-0613, Vol. 2, Issue 01, 2014, pp.385-389.

[7] Manish saggar, Ashish Kumar Agarwal and Abhimunya lad, "Optimization Of Association Rule Mining Using Improved Genetic Algorithms", 2004 IEEE International Conference on Systems, Man and Cybernetics IEEE 2004, pp.3725-3729.

[8] Peter P. Wakabi-Waiswa, Dr. Venansius Baryamureeba, Karunakaran Sarukesi, "Optimized Association Rule Mining with Genetic Algorithms", Seventh International

Conference on Natural Computation- IEEE 2011, pp.1116-1120.

[9] 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.

[10] Jeetesh Kumar Jain, Nirupama Tiwari, Manoj Ramaiya, “Mining Positive and Negative Association Rules from Frequent and Infrequent Pattern Using Improved Genetic Algorithm”, 5th International Conference on Computational Intelligence and Communication Networks-IEEE-2013, pp.516-521.

[11] Sameer Kumar Vishnoi1, Vivek Badhe, “Association Rule Mining for Profit Patterns Using Genetic Algorithm”, International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Volume 4, Issue 5, May 2014,pp.855-858.

[12] Anandhavalli M, Suraj Kumar Sudhanshu, Ayush Kumar and Ghose M.K., “Optimized Association Rule Mining Using Genetic Algorithm”, advances in information mining, ISSN: 0975–3265, volume 1, issue 2, 2009, pp-01-04.

[13] Satya Ranjan Dash, Satchidananda Dehuri, Susil Rayaguru, “Discovering Interesting Rules from Biological Data Using Parallel Genetic Algorithm”, 2012 IEEE.pp.631-636.

[14] J.M. Luna, J.R. Romero, S. Ventura, “Grammar-based multi-objective algorithms for mining association rules”, Data & Knowledge Engineering 2013 Elsevier.pp-19-37.

[15] 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.

[16] F.G. Lobo, D.E. Goldberg, M. Pelikan, “Time complexity of genetic algorithms on exponentially scaled problems”, in: Proceedings of the Genetic and Evolutionary Computation Conference, Morgan-Kaufmann, 2000, pp. 151–158.

[17] Gagandeep Kaur, Shruti aggarwal, “Survey on Genetic Algorithm for Association Rule Mining” IJCA VOLUME-67 April 2013.

[18] R. O. Oladele, J. S. Sadiku,Genetic Algorithm Performance with Different Selection Methods in Solving Multi-Objective Network Design Problem, International Journal of Computer Applications (0975 – 8887) Volume 70– No.12, May 2013

[19] Noraini Mohd Razali, John Geraghty, Genetic Algorithm Performance with Different Selection Strategiesin Solving TSP, Proceedings of the World Congress on Engineering 2011 Vol II WCE 2011, July 6 - 8, 2011, London, U.K

8. AUTHORS PROFILE

Shailendra Mishra, HOD of Computer Department of Parul Institute of Technology college of Gujarat Technological University and his interesting area of research is in Data Mining, Artificial intelligence.

Dimple S. Kanani received the B.E. degree in Information Technology from Om Shani Engineering College, Gujarat Technological University in 2012.Currently she is doing her M.E. Degree from Parul Institute of Technology of Gujarat Technological University and her interesting area of research is in Data Mining.