ISSN: 2321-7782 (Online)
Volume 3, Issue 2, February 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

A Survey on Genetic Algorithm for Association Rule mining

Dimple S.Kanani[1]
Research Scholar, CSE Department
Parul Institute Of Engineering
Vadodara, India

Shailendra Mishra[2]
Assistant Professor, CSE Department
Parul Institute Of Engineering
Vadodara, India

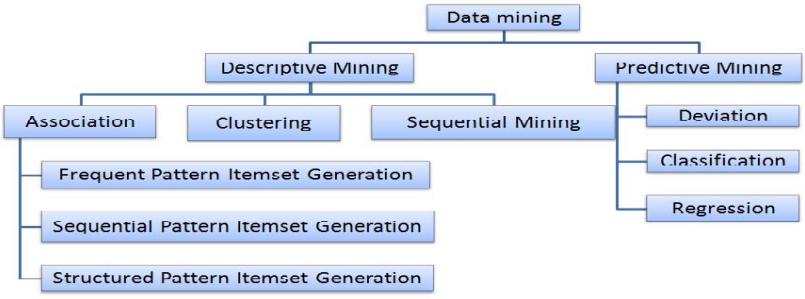
Abstract: ***Association Rule Mining technique that attempt to unearthing interesting pattern or relationship between data in large Database. Genetic Algorithm is a search heuristic which is used to generate useful solution for optimization and search problems. Genetic Algorithm based evaluation in Mining 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. Many researches has been carried out in this area, this paper is to represents the survey on genetic algorithm & fundamentals.***

Keywords: Association rule Mining, Genetic Algorithm, Data mining, Optimization, selection mechanisms.

I. INTRODUCTION

Data Mining

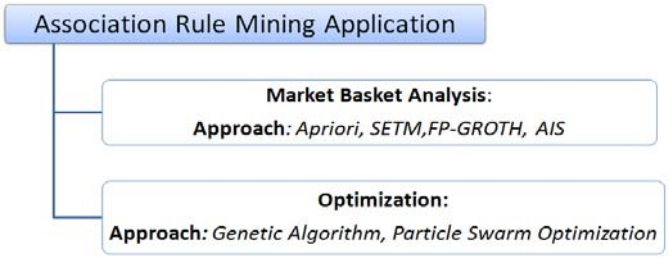
Data Mining is the process of extracting information from large data sets through the use of algorithms and techniques drawn from the field of Statistics, Machine Learning and Data Base Management. Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of 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 popularly known 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.

II. 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 in J (i.e., T ⊆ J). An association rule has the form X ⇒ Y , where X and Y are non-empty sets of items (i.e. X ⊆ J, Y ⊆ J) such that X ∩ Y = Ø. 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 → Y and each such relationship that holds between the attributes of records in a database fulfilling some criteria are termed as an association rule.



MEASURES OF ASSOCIATION RULES:

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.

An association rule is an implication relation in the form X ⇒ Y between two disjunctive sets of items X and Y. An example of an association rule on "market basket data" is that "70% of customers who purchase bread also purchase butter ". Each rule has two quality measurements, support and confidence. The rule X ⇒Y has confidence c if c% of transactions in the set of transactions D that contains X also contains Y. The rule has a support S in the transaction set D if S% of transactions in D contains XUY. The problem of mining association rules is to find all association rules that have a support and a confidence exceeding the user-specified threshold of minimum support and threshold of minimum confidence respectively.

III. Genetic Algorithm

Genetic algorithm (GA) is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms are 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.

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. But using genetic approach has certain limitations also like the genetic algorithm cannot assure constant optimisation response times. There is no absolute assurance that a genetic algorithm will find a global optimum solution.

Genetic algorithm:

Initialization: randomly generate population of N chromosomes

Fitness: Calculate the fitness of all chromosomes.

Create a **new population**:

a. **Selection**: select 2 chromosomes from the population

b. **Crossover**: (Recombination) produce 2 offspring from the 2 Selected Chromosome.

c**. Mutation**: perform mutation on each offspring (bit inversion).

Replace: replace the current population with the new population

Evaluation: evaluation of objective function

Termination: Test if the termination condition is satisfied. If so stop. If not, go to step 2.

A. Representation:

The classical genetic algorithm paradigm deals with the solutions encoded as a literal string, called chromosomes. A chromosome is the representation of a single solution of the problem.

B. Population and initialization:

In initialization, the initial set of chromosomes, also called as the initial population, is created. The size of initial population is important for the overall genetic algorithm. A small size of the initial population can lead to finding of a local optimum only, while a larger initial population gives a higher probability that the global optimum will be found.

C. Selection for reproduction:

The selection operator is used to identify chromosomes which will be used in reproduction and will survive in the next generation. Selection deals with the probabilistic survival of the fittest, in that, more fit chromosomes are chosen to survive. Where fitness is a comparable measure of how well a chromosome solves the problem. Different techniques can be used in selection operators:

A. Tournament Selection mechanism:

Tournament selection is having good efficiency and simple implementation. Here, *n* individuals are selected randomly from the larger population, and the selected individuals will compete with each other. The individual with the highest fitness wins and will be included as one of the next generation population. Tournament size is defined as the number of individuals competing in

each tournament, commonly set to 2 (also called binary tournament). Tournament selection also gives a chance to all individuals to be selected and thus it preserves diversity, although keeping diversity may degrade the convergence speed. The tournament selection has several advantages which include efficient time complexity, especially if implemented in parallel, and no requirement for fitness scaling or sorting [2, 4]. Larger the values of tournament size, higher expected loss of diversity [5]. The larger tournament size means that a smaller portion of the population actually contributes to genetic diversity, making the search increasingly greedy in nature.

B. Roulette Wheel Selection mechanism:

In proportional roulette wheel, individual's selection probability is directly proportional to their fitness values (portion of a roulette wheel.). The probabilities of selecting a parent can be seen as spinning a roulette wheel with the size of the segment for each parent being proportional to its fitness. So, largest fitness (largest segment sizes) have more probability of being chosen. The fittest individual occupies the largest segment, whereas the least fit have correspondingly smaller segment within the roulette wheel. The circumference of the roulette wheel is the sum of all fitness values of the individuals. When the wheel is spun, the wheel will finally stop and the pointer attached to it will point on one of the segment, most probably on one of the widest ones. However, all segments have a chance, with a probability that is proportional to its width. By repeating this each time an individual needs to be chosen, the better individuals will be chosen more often than the poorer ones, thus fulfilling the requirements of survival of the fittest. Let $f1, f2,\ldots, fn$ be fitness values of individual 1, 2,…, n. Then the selection probability, Pi for individual i is define as, $Pi = \text{fi/Pfj}$.

Roulette wheel selection gives a chance to all of individuals to be selected. Therefore, diversity in the population is preserved. However, proportional roulette wheel selection has few major deficiencies. Outstanding individuals will introduce a bias in the beginning of the search that may cause a premature convergence and a loss of diversity. For example, if an initial population contains one or two very fit but not the best individuals and the rest of the population are not good, then these fit individuals will quickly dominate the whole population and prevent the population from exploring other potentially better individuals. Such a strong domination causes a very high loss of genetic diversity which is definitely not advantageous for the optimization process. On the other hand, if individuals in a population have very similar fitness values, it will be very difficult for the population to move towards a better one since selection probabilities for fit and unfit individuals are very similar.[3]

C. Rank-Based Selection mechanism:

Rank-based selection is the selection strategy where the probability of a chromosome being selected is based on its fitness rank relative to the entire population. Rank-based selection schemes first sort individuals in the population according to their fitness and then computes selection probabilities according to their ranks rather than fitness values. Hence rank-based selection can maintain a constant pressure in the evolutionary search where it introduces a uniform scaling across the population and is not influenced by super-individuals or the spreading of fitness values at all as in proportional selection. Rank-based selection uses a function to map the indices of individuals in the sorted list to their selection probabilities. Although this mapping function can be linear (linear ranking) or non-linear (non-linear ranking), the idea of rank-based selection remains unchanged. The performance of the selection scheme depends greatly on this mapping function. Rank-based selection schemes can avoid premature convergence and eliminate the need to scale fitness values, but can be computationally expensive because of the need to sort populations. Rank-based selection scheme helps prevent premature convergence due to "super" individuals, since the best individual is always assigned the same selection probability, regardless of its objective value. However this method can lead to slower convergence, because the best chromosomes do not differ so much from other ones.

Comparison based on iteration time & computation time: rank based selection, Tournament selection, Roulette wheel selection Method from higher to lower respectively. Rank-based selection on the other hand continues to explore the search space and reaching the lowest traveling distance in the tour. Therefore it can be conclude that tournament selection is more

appropriate for small size problem while rank-based can be used to solve larger size problem. If solution quality is the main concern and computation time is still negotiable, then rank-based selection strategy is the best choice.

D. Recombination / Crossover:

This operation is performed by selecting a random gene along the length of the chromosomes and swapping all the genes after that point. Single point crossover, Two point crossover, Uniform crossover, Arithmetic crossover are types of crossover.

E. Mutation:

Mutation changes randomly the new offspring. For binary encoding reverse the randomly chosen bits from 1 to 0 or from 0 to 1. A mutation operator is used with intention to prevent getting stuck in the local optimum and increase a probability to find the global optimum (Honget al., 2002). Types of Mutation are: Flip Bit - (Used for binary represented genes), Uniform – (Used for integer and float representation), Boundary - (Used for integer and float represented genes)

F. Replacement:

This process will compare between several chromosomes to choose the best. Types of replacement are: Binary Tournament, Triple Tournament.

G. Termination:

Stopping criteria for genetic Algorithm are the maximum computation time, the maximum iteration number, Allocating budget (ex: time, money) reached, received solution satisfies the minimum criterion, or iterations that are counted from the last successful improvement of the best individual. Etc.

IV. RELATED WORK

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

Rupali Haldulakar and Prof. Jitendra Agrawal [10] 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 [8] 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 [1] 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 [3] 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.

R.Sivaraj, Dr.T.Ravichandran [12] define that Genetic Algorithms are optimization algorithms that maximize or minimize a given function. Selection operator deserves a special position in Genetic algorithm since it is the one which mainly determines the evolutionary search spaces. It is used to improve the chances of the survival of the fittest individuals. There are many traditional selection mechanisms used and many user specified selection mechanisms specific to the problem definition.

V. CONCLUSION

Genetic Algorithm solves multi-objective Rule mining problem can Work on Categorical and Numerical Attribute. A number of works are already published in this field, this is the survey study to use the enormous robustness of genetic algorithm in mining Association rules –where fitness function is key to find accurate rule, selection method play major role in reducing execution time by selecting parents for reproduction, Different selection mechanisms work well under different situations. Appropriate method has to be chosen for the specific problem to increase the optimality of the solution & Crossover and Mutation rate avoid premature Convergence of algorithm.

ACKNOWLEDGEMENT

I'm extremely obliged to my Gguide 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.

References

1. 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
2. Goldberg, D. E. and Deb Kalyanmoy. 1991. A Comparative Analysis of Selection Schemes Used in Genetic Algorithms. In: G.J.E. Rawlins (Ed), Foundations of Genetic Algorithms, Morgan Kaufmann, Los Altos, 69 – 93.
3. 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.
4. Blickle, T, Thiele, L. A. 1995. Comparison of Selection Schemes used in Genetic Algorithms. TIK-Report, Zurich.
5. Whitley, D. 1989. The genitor algorithm and selection pressure: Why rank-based allocation of reproductive trials is the best. In Proceeding of the 3rd International Conference on Genetic Algorithms.
6. B. Minaei-Bidgoli, R. Barmaki, M. Nasiri, Mining numerical association rules via multi-objective genetic algorithms, Information Sciences 233 (2013) 15–24
7. Mehrdad Dianati, Insop Song, and Mark Treiber, An Introduction to Genetic Algorithms and Evolution Strategies, 200 Univ. Ave. West, University of Waterloo, Ontario, N2L 3G1, Canada.
8. Anandhavalli M, Suraj Kumar Sudhanshu, Ayush Kumar AndGhose M.K., "Optimized Association Rule Mining Using Genetic Algorithm", advances in information mining, ISSN: 0975–3265, volume 1, issue 2, 2009, pp-01-04.

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. RupaliHaldulakar 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.

11. Firas Alabsi, Reyadh Naoum, Comparison Of Selection Methods And Crossover Operations Using Steady State Genetic Based Intrusion Detection System, Journal Of Emerging Trends In Computing And Information Sciences, VOL. 3, NO.7, July 2012

12. R.Sivaraj, Dr.T.Ravichandran, A Review Of Selection Methods, In Genetic Algorithm, International Journal Of Engineering Science And Technology (Ijest).

AUTHOR(S) PROFILE

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.

Shailendra Mishra, HOD of Parul Institute of Technology of Gujarat Technological University and her interesting area of research is in Data Mining, Artificial intelligence, Soft Computing.

