International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
Since 2012

http://www.ijcsjournal.com Volume 12, Issue 1, No 2, 2024. ISSN: 2348-6600
Reference ID: IJCS-493 PAGE NO: 3389-3397

AN OVERVIEW OF GENETIC ALGORITHMS FOR ASSOCIATION RULE DISCOVERY

SHEELA DEVI. R
M.Sc Student,
Department of Computer Science,
Rajeswari College of Arts and Science for Women,
Bommayapalayam, Tamil Nadu, India.
Email ID: rsheeladevi02@gmail.com

Dr.A.SARANYA
Assistant Professor,
Department of Computer Application,
Rajeswari College of Arts and Science for Women,
Bommayapalayam, Tamil Nadu, India.
Email ID: drsaranyarcw@gmail.com

Abstract

Data mining has been a crucial component in the creation of association rules among the numerous item sets in recent years. The process of finding interesting relationships between variables in big databases is called association rule mining. It is considered as one of the key data mining jobs meant to support decision-making. Based on the principles of evolution, genetic algorithms (GAs) have found a solid foundation in the mining of association rules. Optimizing and searching problems can be effectively solved with the use of genetic algorithms, a type of search heuristic. Compared to other formal methods, genetic algorithms have demonstrated the ability to produce findings that are more accurate. Every rule's quality is assessed by the genetic algorithm's fitness function. A genetic algorithm has been developed by numerous academics to extract interesting rules from datasets. For association rule mining, this study includes an overview of genetic algorithms.

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

I. Introduction

The technique of extracting meaningful information from massive amounts of data is known as data mining. A significant amount of data is produced as a result of the advancement of information technology in many areas of human activity. Many formats, including records, documents, pictures, and more, can be used to store the data. For improved decision-making, the data gathered from various applications requires an appropriate technique for knowledge extraction from sizable archives. Finding valuable information in huge amounts of data is the goal of data mining [1]. Data mining, information extraction, data/pattern analysis, and data archaeology are some other names for it. Data mining programs look for patterns in databases that are hidden and uncover information that humans may overlook. While data mining is a part of knowledge discovery, the phrases "knowledge discovery from data" and "data mining" are often used in the same sentence. The following steps make up the knowledge discovery process [2, 3]:

- **Data Selection:** The user-relevant data is obtained from the database during this phase.
- **Data Cleaning:** This step involves removing noise and unwanted data.
- **Data Integration:** Data gathered from various data sources may be integrated at this point.
- **Data Transformation:** Selected data is converted into the proper formats at this phase.
- **Data Mining:** Utilizing specialized methods to identify patterns in the data is the last phase.
- Pattern Evaluation: In this step, supplied measures are used to identify interesting trends that represent knowledge.
- **Knowledge Representation:** It is the conclusion that the user is represented by mined knowledge using knowledge representation techniques.

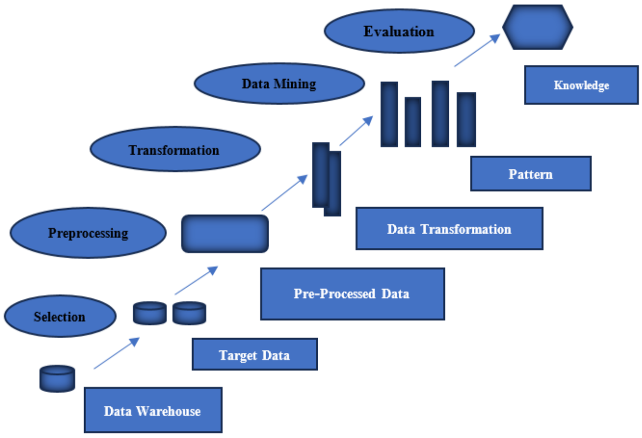


Fig 1: KDD Process

2. Association Rule Mining

Association Rule Mining is a well-liked and extensively studied technique in data mining for identifying intriguing relationships between variables in big databases. When Agrawal et al. created the Apriori algorithm in 1993 to solve ARM-based problems, they first introduced the Association Rule Mining approach. Association rules are applied in a number of fields, including inventory control, market and risk management, and telecommunication networks. An association rule is represented as an implication in the format X⇒Y, where X and Y are collections of items and XnY=Ø. The terms "antecedent" (X) and "consequent" (Y) refer to the left-hand side (LHS) and right-hand side (RHS) in this instance [4]. In both the antecedent and consequent parts, a rule may have more than one item.

2.1 Measures of Association Rules

Confidence and support are two essential foundational factors to take into account while selecting very interesting rules from the set of all possible rules. Users typically set their own support and confidence criteria to eliminate less interesting or useful rules. Minimum support and minimum confidence are the names given to the two levels, respectively. Strong rules fulfill both the minimum confidence and minimum support thresholds [4].

2.1.1 Support

In the specified transactional database, it is the likelihood that an item or item sets:

Where n is the total number of transactions in the database and n(X) is the number of transactions that contain the itemset X, Support(X)=n(X)/n calculates. If an item has a 0.1% support, then just 0.1% of transactions contain it [5].

2.1.2 Confidence

The percentage of database transactions that contain both X and Y is known as the confidence of a rule (X⇒Y). It is calculated in this way:

- Confidence (X⇒Y) is equal to support (X and Y)/support(X).
- 70% of the transactions containing X also contain Y if the association rule X⇒Y has a 70% confidence level [5].

2.2 Apriori Algorithm

Several association rule mining algorithms with varying mining efficiency have been created. All frequently used sets are collected and association rules are learned using the classical algorithm Apriori. Rules of association can be derived from the frequent item sets produced by Apriori. Databases with transactions are the target application for Apriori (e.g., collections of products that customers have purchased). An itemset is a collection of items found in every transaction. When exploring (k+1)-item sets, Apriori employs level-wise search using k-item sets, which are item sets that contain k-items. The collection of often occurring 1-itemsets is located at the start. L1 designates this set, which consists of objects that meet minimum support requirements. We start each new pass with a group of item sets that were discovered to be common in the prior session. This set is used to create new item sets, also called candidate item sets, so that the true support for these candidate item sets can be counted during the data pass. We identify the candidate item sets that are truly frequent at the end of the pass, and those are employed in the following pass. As a result, L1 is utilized to locate L2, which in turn locates L3, the set of frequent 2-itemsets, and so on, until no more frequent k-item sets are discovered. The Apriori property, which is defined as follows: "All nonempty subsets of a frequent itemset must also be frequent," is a significant feature that is used to minimize the search space [6, 7].

3. Genetic Algorithm

John Holland first proposed genetic algorithms, or GAs, in 1970. Natural selection and natural genetics are the foundational concepts of GA, a random search method. It has shown ability in generating significant solutions for an array of machine learning and optimization problems. The Genetic Algorithm creates new populations of strings from older ones through iteration. Darwin's theory of evolution serves as an inspiration for genetic algorithms [8]. The method runs in generations, beginning with a population of randomly produced people.

Each generation decides the fitness of the population by evaluating each member; stronger members are selected from the current population, and genetic operators are employed to generate a new population. The new population is then exploited in the algorithmic iteration that follows. The algorithm ends after a maximum number of generations has been produced or the population gets a desirable fitness level. Reproduction (selection), crossover, and mutation are the three genetic operators that a genetic algorithm typically uses. The following are the roles that genetic operators play:

- **Selection:** Fit chromosomes are chosen to survive in selection according to the theory of "survival of the fittest."

Numerous methods, including ranking, tournaments, roulette wheels, and more, can be used to select the best chromosomes [9].

- **Crossover:** It creates new chromosomes by combining fragments of the parents' chromosomes. There are various types of crossovers, including single-point and multipoint ones.
- **Mutation:** Mutation changes the new solutions to produce superior ones. It completely alters the individual with a fixed probability of mutation (0 turns into 1, 1 transforms into 0).

4. Related Works

Mining association rules has made extensive use of genetic algorithms. The work done in the area of association rule mining with genetic algorithms is covered in this part. The Apriori technique was used by Anandhavalli M. et al. [8] to generate frequent item sets, and the evolutionary method was then applied to generate rules with positive attributes, the negation of attributes with one attribute, or more than one attribute in the next section. The fitness function in this study is comprised of the confidence factor and completeness, which are a combination of true positives, true negatives, false positives, and false negatives. Most of the created rules had lower support levels because it concentrated on negative qualities. To demonstrate the efficacy of the suggested approach, they make use of the Lens database from UCI. Since the identified rules are optimized rules, the results presented in this research are extremely promising. By applying Baye's theorem to the generated rules, the complexity of database scanning and the Genetic Algorithm can be reduced.

Rupali Haldulakar et al. [10] say that powerful rule creation is an important component of data mining. The authors of this paper established a brand-new method for generating strong rules. The rules are generated by the Apriori algorithm. They then apply optimization strategies. The Genetic Algorithm is among the greatest methods for improving the rules. Using the idea of supervised learning, they create a novel fitness function for the rule set optimization. When the fitness function is divided into two classes—one for discrete rules and the other for continuous rules—c1 and c2. For their experiments, they use the Abalone dataset, which they got from the UCI machine learning repository. There are 4177 samples in the data set. The new fitness function, which makes use of the idea of supervised learning, is what the authors used to get better results. The genetic algorithm can be used with other methods to maximize its effectiveness and efficiency and yield the best outcome.

A unique association rule mining approach based on genetic algorithms was proposed by M. Ramesh Kumar et al. [11]. The fitness function is created using two metrics, namely the confidence in all rules and their aggregate strength, as opposed to the traditional support and the confidence in the rules that are developed. The approach has been tested on the following four data sets: Adult, Chess, Wine, and Zoo. The UCI data repository provided the sample data sets used in the algorithms testing. They used 200 people as the population size for GA testing, a 10% selection rate, a 6% crossover rate, and a fixed 1% mutation rate as their environmental measure. The rules are prioritized by the fitness function according to user preference. Using four data sets, their method significantly reduces the amount of rules generated. Future work can expand the technique by adding interesting measures.

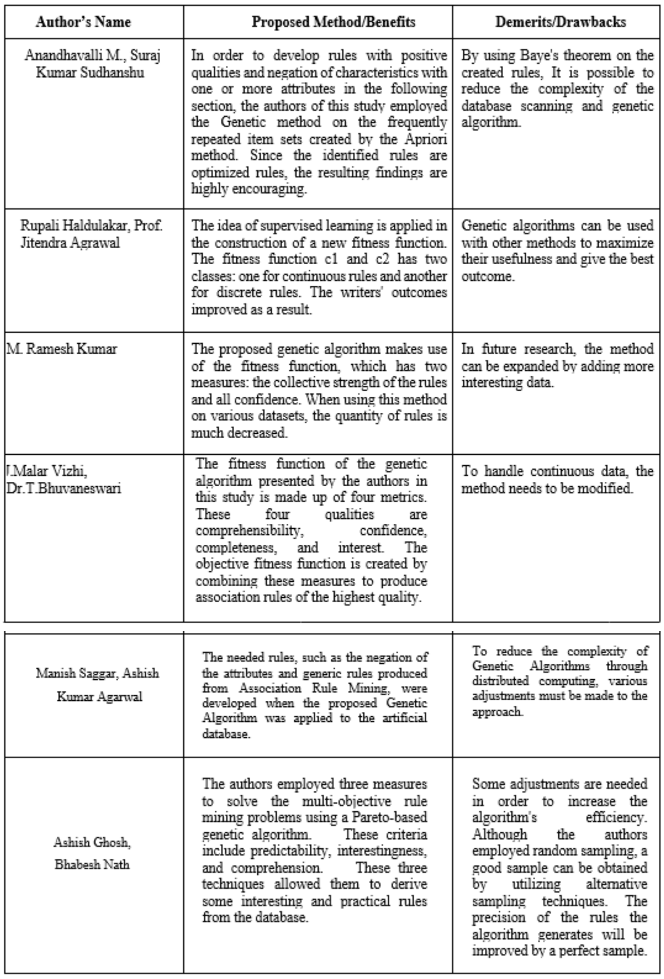
A genetic approach for producing high-quality association rules was proposed by J. Malar Vizhi et al. [9]. Problems involving association rule mining might be viewed as multi-objective as opposed to single-objective. When mining association rules, they employed a multi-objective evolutionary framework. They use four metrics in their fitness function. These are clarity, confidence, intrigue, and completion. The objective fitness function is created by combining these metrics. The fitness function is the arithmetic weighted average of the confidence, completeness, attraction, and comprehensibility. They used the primary tumor dataset from the real world for their experiments. 339 occurrences and 18 attributes make up the Primary-tumor database. The tournament selection and multipoint crossover techniques were employed, as they offer flexibility for a multitude of database attributes. Only the numerical and categorical valued elements were used to test the methodology. To handle continuous data, the method must be modified.

Manish Saggar et al. [12] optimized the rules produced by Association Rule Mining (Apriori technique) using Genetic Algorithms. The Association Rule Mining technique typically generates rules that do not take into account negative attribute occurrences. But the system can predict the rules that do have negative qualities by using Genetic Algorithms (GAs) on these rules. The rule-based classification methods benefit from the enhancements made by authors to GAs. Prior to applying GAs to generate rules with negations in attributes and higher quality, the authors first implemented Association Rule mining. They created the database they used artificially. The authors created general rules using association rule mining and rules that negated specific features on an artificial database using their method. To reduce the complexity of Genetic Algorithms through distributed computing, various adjustments must be made to the approach.

The multi-objective rule mining problems were solved by Ashish Ghosh et al. [13] using a genetic algorithm based on Pareto. It is possible to think of association rule mining as a multi-objective task compared to a single-objective one. The authors used three methods to extract some interesting and useful ideas from the information. These three characteristics are clarity, interest level, and prediction accuracy. Some easily understood rules that were not known before can be produced with the use of these three measurements. They represented the rules as chromosomes using a Michigan-style approach, in which each chromosome stands for a distinct rule. A few changes are needed to increase the algorithm's efficiency. Although a good sample can be obtained by using alternative sampling techniques, the authors of this work choose to employ the random sampling method. The algorithm's rules will be more accurate if there is a perfect sample.

The benefits and drawbacks of the several Association Rule Mining with Genetic Algorithm techniques are listed in the following table:

Table 1: A comparative Analysis of Genetic Algorithms



Author's Name	Proposed Method/Benefits	Demerits/Drawbacks
Anandhavalli M., Suraj Kumar Sudhanshu	In order to develop rules with positive qualities and negation of characteristics with one or more attributes in the following section, the authors of this study employed the Genetic method on the frequently repeated item sets created by the Apriori method. Since the identified rules are optimized rules, the resulting findings are highly encouraging.	By using Baye's theorem on the created rules, It is possible to reduce the complexity of the database scanning and genetic algorithm.
Rupali Haldulakar, Prof. Jitendra Agrawal	The idea of supervised learning is applied in the construction of a new fitness function. The fitness function c1 and c2 has two classes: one for continuous rules and another for discrete rules. The writers' outcomes improved as a result.	Genetic algorithms can be used with other methods to maximize their usefulness and give the best outcome.
M. Ramesh Kumar	The proposed genetic algorithm makes use of the fitness function, which has two measures: the collective strength of the rules and all confidence. When using this method on various datasets, the quantity of rules is much decreased.	In future research, the method can be expanded by adding more interesting data.
J.Malar Vizhi, Dr.T.Bhuvaneswari	The fitness function of the genetic algorithm presented by the authors in this study is made up of four metrics. These four qualities are comprehensibility, confidence, completeness, and interest. The objective fitness function is created by combining these measures to produce association rules of the highest quality.	To handle continuous data, the method needs to be modified.
Manish Saggar, Ashish Kumar Agarwal	The needed rules, such as the negation of the attributes and generic rules produced from Association Rule Mining, were developed when the proposed Genetic Algorithm was applied to the artificial database.	To reduce the complexity of Genetic Algorithms through distributed computing, various adjustments must be made to the approach.
Ashish Ghosh, Bhabesh Nath	The authors employed three measures to solve the multi-objective rule mining problems using a Pareto-based genetic algorithm. These criteria include predictability, interestingness, and comprehension. These three techniques allowed them to derive some interesting and practical rules from the database.	Some adjustments are needed in order to increase the algorithm's efficiency. Although the authors employed random sampling, a good sample can be obtained by utilizing alternative sampling techniques. The precision of the rules the algorithm generates will be improved by a perfect sample.

5. CONCLUSION

Genetic algorithms are less complex than other algorithms and can do global searches, they are utilized in the finding of high-quality rules. Genetic algorithms have been widely used in mining Association Rules in recent years. This research reviewed the literature on genetic algorithms for association rule mining. The use of multi-objective genetic algorithms in association rule mining is less common. It is possible to find more fascinating rules by using more interesting metrics.

6. REFERENCES

[1] Venkatadri.M and Dr. Lokanatha C. Reddy, "A Review on Data mining from Past to the Future", International Journal of Computer Applications, Volume 15– No.7, pp. 19-22, February 2011.

[2] Gurjit Kaur and Lolita Singh, "Data Mining: An Overview", International Journal of Computer Science and Technology, Volume 2, Issue 2, pp. 336-339, June 2011.

[3] Vibha Maduskar and Prof. Yashovardhan Kelkar, "Survey on Data Mining", International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 2, pp. 275-279, February 2012.

[4] Sotiris Kotsiantis and Dimitris Kanellopoulos, "Association Rules Mining: A Recent Overview", International Transactions on Computer Science and Engineering, Volume 32 (1), pp. 71-82, 2006.

[5] Shanta Rangaswamy and Shobha G, ''Optimized Association Rule Mining using Genetic Algorithm'', Journal of Computer Science Engineering and Information Technology Research, Volume 2, Issue 1, pp 1-9, Sep 2012.

[6] Sanjeev Rao and Priyanka Gupta, "Implementing Improved Algorithm over Apriori Data Mining Association Rule Algorithm", International Journal of Computer Science and Technology, Volume 3, Issue 1, pp. 489-493, Jan. - March 2012.

[7] Han J. and M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann publishers, 2nd Edition.

[8] Anandhavalli M., Suraj Kumar Sudhanshu, Ayush Kumar and Ghose M.K., "Optimized Association Rule Mining using Genetic Algorithm", Advances in Information Mining, Volume 1, Issue 2, pp. 1-4, 2009.

[9] J.Malar Vizhi and Dr. T.Bhuvaneswari, "Data Quality Measurement With Threshold Using Genetic Algorithm", International Journal of Engineering Research and Applications, Volume 2, Issue 4, pp. 1197- 120, July-August 2012.

[10] Rupali Haldulakar and Prof. Jitendra Agrawal, "Optimization of Association Rule Mining through Genetic Algorithm", International Journal on Computer Science and Engineering, Volume 3, No. 3, pp. 1252-1259, Mar 2011.

[11] M. Ramesh Kumar and Dr. K. Iyakutti, "Application of Genetic algorithms for the prioritization of Association Rules", International Journal of Computer Applications, pp. 35-38, 2011.

[12] Manish Saggar, Ashish Kumar Agarwal and Abhimanyu Lad, "Optimization of Association Rule Mining using Improved Genetic Algorithms", IEEE International Conference on Systems, Man and Cybernetics, pp. 3725- 3729, 2004.

[13] Ashish Ghosh, Bhabesh Nath, "Multi-objective Rule Mining using Genetic Algorithms", Information Sciences, pp. 123–133, 2004.