International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
e-ISSN: 2455-2585
Volume 8, Issue 3, March-2022

GENETIC ALGORITHM FOR FINDING FREQUENT PATTERNS

CH Mahender Reddy
Asst.Professor, CSE, CMR Institute of Technology, Hyderabad, India
S Sravani,
Asst.Professor, CSE, CMRInstitute of Technology, Hyderabad, India
HaripriyaPatra
Asst.Professor, CSE, CMR Institute of Technology, Hyderabad, India

ABSTRACT: *Frequent pattern mining is a popular problem in data mining, which consists of finding frequent patterns in transaction databases. Many popular methods have been proposed for this problem by applying Association Rule Mining algorithms likeApriori algorithm, Pincer search algorithm, Border algorithm, Partition algorithm etc. In this paper we are proposing Genetic algorithm for mining frequent patterns. This method improves the computational complexity to find frequent patters compared to the association rule mining methods.The main aim of this paper is to find all the frequent patterns from given data sets using genetic algorithm.*

Keywords: *– Genetic Algorithm (GA), Association Rule, Frequent itemset, Support, Confidence,Data Mining***.**

Received: *28/02/2022,****Accepted:****14/03/2022****,Published:****21/03/2022*

INTRODUCTION: With the rapid development of computer technology in different sectors, the data which is generated by different industries are becoming more and more, but extracting valuable information from the big data has become a new problem. Data mining, that is data knowledge discovery, came into being in this backdrop. Data mining is to find out the implied, unknown, interesting knowledge and rules from a large number of data. Association rules is an important part of data mining, it was first put forward by R.Agrawal, to solve the customer transaction association rules between sets of items in the transaction library. In the next year, R.Agrawal proposed the most classical algorithm to calculate association rules, that is Apriori algorithm, which is to construe the (k+1) – item sets by the k- item sets.

Data mining has become a popular and interesting research area in computer science engineering and information industry in recent times, due to the wide availability of huge amounts of data and the immediate need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration.

Frequent patterns play an important role in many Data Mining tasks that try to find interesting patterns from databases, such as association rules, correlations, sequences, episodes, classifiers and clusters. The mining of association rules is one of the most popular problems of all these. The identification of sets of items, products, symptoms and characteristics, which often occur together in the given database, can be seen as one of the most basic tasks in Data Mining.

IMPLEMENTATION:

Association rule mining is a rule-based machine learning method for identifyinginteresting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness. Let us consider some terms here:

- **T**: A set of all transactions that customers make and are recorded in the stores system. Each transaction has a unique transaction ID associated to their shopping list.
- **Basket**: A set of all items bought by a customer.
- **Item-set**: A set of items that we are interested in.

Now, let's assume that we analyzed customer's transactions and realized that "Many of them", if they had *milk* in their baskets, they also have eggs! This gives us a frequent pattern. There has to be a way to evaluate the importance of a discovered rule. Here comes the **support** and **confidence**. Suppose the rule we discovered is as follows:

Milk → Eggs [*Support*: 9%, *Confidence: 65*%]

If you think about it, apparently not many customers have this item-set (*milk and eggs*) in their baskets (only 9%), however, if they go for *milk*, it is somewhat likely (65%) that they also go for *eggs!* This means, if we have any offer for our cheese produces, we should definitely inform *wine*-buyers, since they are the best potential buyers.

Support

The support of an itemset X, $supp(X)$ is the percentage of transactions in T that contain both *milk* and *eggs* together. (9% of all baskets had these 2 items together.)

$$supp(X) = \frac{\text{Number of transaction in which} X \text{appears}}{\text{Total number of transactions}}.$$

Confidence

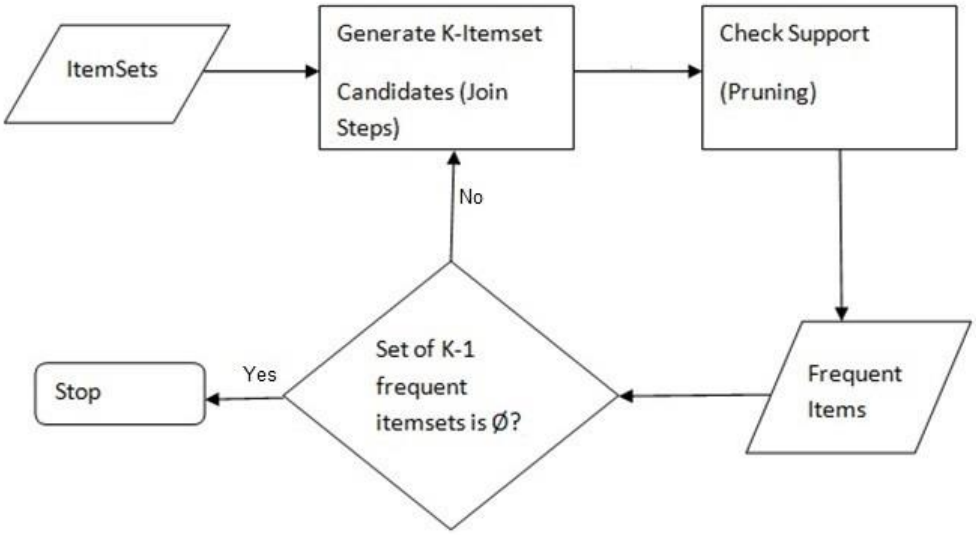
Confidence of a rule is defined the percentage of transactions in T, containing *milk*, that also contain *eggs*. In other words, the probability of having *eggs*, given that *milk* is already in the basket. (65% of all those who bought *milk*, also bought *eggs*.)

$$conf(X \longrightarrow Y) = \frac{supp(X \cup Y)}{supp(X)}$$

General Process of the Apriori algorithm

The entire algorithm can be divided into two steps:

Step 1: Apply minimum support to find all the frequent sets with k items in a database.
Step 2: Use the self-join rule to find the frequent sets with k+1 item with the help of frequent k-itemsets. Repeat this process from k=1 to the point when we are unable to apply the self-join rule.
This approach of extending a frequent itemset one at a time is called the "bottom up" approach.



Pros of the Apriori algorithm

1. It is an easy-to-implement and easy-to-understand algorithm.
2. It can be used on large itemsets.

Cons of the Apriori Algorithm

1. Sometimes, it may need to find a large number of candidate rules which can be computationally expensive.
2. Calculating support is also expensive because it has to go through the entire database.

All the traditional association rule mining algorithms were developed to find positive associations between items. Positive associations refer to associations between items existing in transactions. In addition to the positive associations, negative associations can provide valuable information. In practice there are many situations where negation of products plays a major role. By using Genetic Algorithm (GA) the system can predict the rules which contain negative attributes in the generated rules along with more than one attribute in consequent part.

Proposed system:

The genetic algorithm is a method for solving both constrained and unconstrained optimization problems that is based on natural selection, the process that drives biological evolution. The genetic algorithm repeatedly modifies a population of individual solutions. At each step, the genetic algorithm selects individuals at random from the current population to be parents and uses them to produce the children for the next generation. Over successive generations, the population "evolves" toward an optimal solution. It is frequently used to find optimal or near-optimal solutions to difficult problems which otherwise would take a lifetime to solve. It is frequently used to solve optimization problems, in research, and in machine learning.

GA has been successfully applied in many search, optimization, and machine learning problems. GA works in an iterative 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.

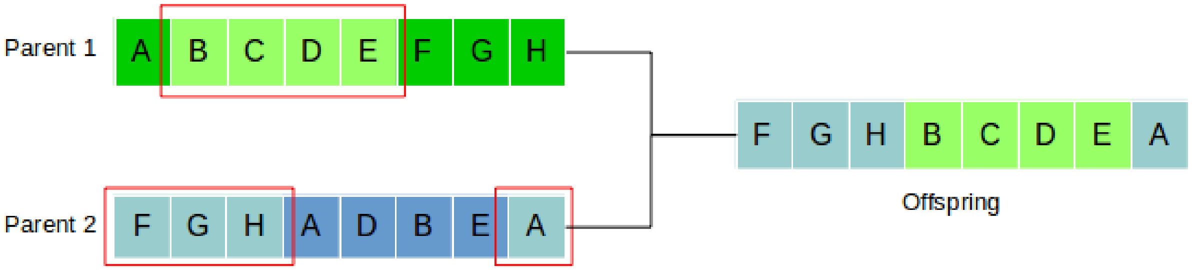
Genetic algorithm applies genetic operators on a random population in order to generate new itemsets. GA runs to generate solutions for successive generations. The probability of an individual reproducing is proportional to the goodness of the solution it represents. Hence the quality of the solutions in successive generations improves. The process is terminated when an acceptable or optimum solution is found. GA is appropriate for problems which require optimization, with respect to some computable criterion.

Operators of Genetic Algorithms

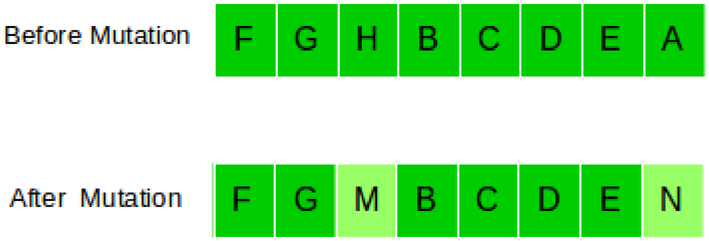
Once the initial generation is created, the algorithm evolves the generation using following operators

1) Selection Operator: The idea is to give preference to the individuals with good fitness scores and allow them to pass their genes to the successive generations.

2) Crossover Operator: This represents mating between individuals. Two individuals are selected using selection operator and crossover sites are chosen randomly. Then the genes at these crossover sites are exchanged thus creating a completely new individual (offspring). For example –



3) Mutation Operator: The key idea is to insert random genes in offspring to maintain the diversity in population to avoid the premature convergence. For example –



PSEUDO-CODE FOR GENETIC ALGORITHM:

```
1. Choose the random population of individuals.
2. [fitness] Evaluate the fitness f(x) of each chromosome x in the population.
3. [New population] create a new population by repeating the following steps until the new population is complete.
   a) Select the best-fit individuals for reproduction
   b) Breed new individuals through crossover and mutation operations to give birth to offspring
   c) Evaluate the individual fitness of new individuals
   d) Replace least-fit population with new individuals
```

(1) *Fitness Evaluation:* The fitness (i.e., an objective function) is calculated for each individual.
(2) *Selection:* Individuals are chosen from the current population as parents to be involved inrecombination.
(3) *Recombination:* New individuals (called offspring) are produced from the parents byapplying genetic operators such as crossover and mutation.
(4) *Replacement:* Some of the offspring are replaced with some individuals (usually with theirparents).

One cycle of transforming a population is called a generation. In each generation, a fraction ofthe population is replaced with offspring and its proportion to the entire population is called thegeneration gap (between 0 and 1).

CONCLUSION:

Mining profit pattern mixes the statistic based pattern extraction with value-based decision making to achieve the business goals. Using Genetic Algorithm to optimized rules not only improves the mining process but also provide the accuracy and efficiency to association rule mining. Although a many researches has been carried out in association rule mining but still it requires more attention for defining the notion of profit which would help in improving business strategies.

REFERENCES

[1] J. Han and M. Kamber, “Data Mining: Concepts and techniques”, Morgan Kaufmann Publishers, Elsevier India, 2001.

[2] R Agrawal, T.Imielinski, and A.Swami, 1993. “Mining association rules between sets of items in large databases”, in proceedings of the ACM SIGMOD Int'l Conf. on Management of data, pp. 207-216.

[3] Melanie Mitchell, An Introduction to Genetic Algorithms, PHI, 1996

[4] A. Tiwari, R.K. Gupta and D.P. Agrawal “A survey on Frequent Pattern Mining : Current Status and Challenging issues” Information Technology Journal 9(7) 1278-1293, 2010.

[5] Ke Wang, Senqiang Zhou, and Jiawei Han, Profit Mining: From Patterns to Actions, C.S. Jensen et al. (Eds.): EDBT 2002, LNCS 2287, pp. 70–87, 2002.Springer-VerlagBerlin.

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

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

[8] L. I. Kuncheva, J.C. Bezbek, R.P.W Duin, ―Decision template for multiple classifier fusion: an experimental comparison‖, Pattern Recognition, Vol-34, pp.299-314, 2010.

[9] M. Re, G. Valentini, ―An ensemble based data fusion for gene function prediction, Multiple Classifier Systems‖, Springer, pp.448-457, 2009. [10] H.R. Albert, R. Ko, R. Sabourin, A. S. Britto, L. Oliveira, ―Pair wise fusion matrix for combining classifiers‖, Pattern Recognition, Vol-40, pp. 2198-2210, 2007.

[11] J. Kennedy, R. Eberhart, ―Particle Swarm Optimization‖, Proc. of IEEE Int. Conf. on Neural Networks, pp.1942-1948, 1995.

[12] A.M. Sarhan, ―Cancer classification based on micro array gene expression data using DCT and ANN‖, Proc. Of Int. Conf. on General of Theoretical and Applied Information Technology, pp. 208-216, 2009.

[13] R. Kumar, M.S.B. Saithij, S. Vaddadi, S.V.K.K. Anoop, ―An intelligent functional link artificial neural network for channel equalization‖, Proc. of Int. Conf. on Signal Processing Robotics and Automation, pp. 240-245 2009.

[14] E.Peterson, ―Partitioning large –sample microarray –based gene expression profile using principal component analysis‖, Computer Methods Programming in Biomedicine, pp107-109 2003.