FREQUENT PATTERN MINING USING GENETIC ALGORITHM

Mamta[1], Sunil Kumar[2], Sunita Beniwal[3*]

Article History: Received: 05.05.2023 Revised: 15.06.2023 Accepted: 11.07.2023

Abstract

Data Mining is one of the most important tools in discovering and analyzing knowledge from the data. Association Rule Mining is one technique of data mining. Many algorithms are used for mining association rules. The main disadvantage of using classical approaches for frequent pattern mining is that these are time consuming and the rules generated from these are not interesting and strong. The research aims to overcome the above shortcomings. To generate association rules, Apriori algorithm is combined with genetic algorithm. Genetic algorithm is applied on the frequent patterns which are obtained by the Apriori algorithm. Lift measure is used to measure the correlation in itemsets i.e. whether they are correlated negatively, positively or are independent of each other. The rules having a lift value greater than one are considered as having a positive dependence. Lift measure helps the users to discover their choice of rules. The use of lift parameter has reduced the number of rules generated and also reduced the time required.

Keywords: Support, confidence, crossover, mutation, frequent itemset.

[1,2,3*]Department of Computer Science and Engineering, Guru Jambheshwar University of Science and Technology

***Corresponding Author:**
Sunita Beniwal[3*]
[3*]Department of Computer Science and Engineering, Guru Jambheshwar University of Science and Technology
Email: [3*]sunitabeniwalcse@gmail.com

DOI: 10.31838/ecb/2023.12.6.168

1. Introduction

Data mining is the process to extract knowledge from large amounts of data [1]. "It is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large databases "[2]. "Data mining is the non trivial process that automatically collects the useful hidden information from the data and is taken on as forms of rule, concept, pattern and so on [3]. To find frequent patterns, Association rule mining is used. Traditional approaches used for frequent pattern mining are time consuming and the rules generated are less interesting. These problems can be reduced by combining Apriori algorithm with genetic algorithm. In this paper, Apriori and genetic algorithm are used on 1984 United States Congressional Voting Records dataset [4] for evaluation and the proposed algorithm is implemented in MATLAB.

Association rule mining has been widely used in last few decades. Association rules are used to identify and represent dependencies in data [5]. The main objective is to find sets of attributes that occur together frequently to find comprehensive rules that are able to explain the relationship among attributes [6]. Association rules are generally used to find the relationship between the attributes and transactions which are part of a system. These rules help in decision making. An association rule is an implication of the form A=>C, where A, C belongs to itemset and $A \cap C = \emptyset$. A is antecedent and C is called consequent of the rule respectively. The classical approaches used are Apriori, FP- Growth, ECLAT etc. which have some shortcomings like these are relatively time consuming and inefficient. So there is always a need to improve the performance related issues and to enhance the efficiency.

Genetic Algorithms (GA) are search algorithms based on natural genetics that provide robust search capabilities in complex spaces, and thereby offer a valid approach to problems requiring efficient and effective search processes [7]. Genetic algorithms were first introduced by John Holland [7]. He defined Genetic Algorithm as a process of travelling between the populations i.e. to pass from one population to other population. Populations are made up of chromosomes. Many steps are carried out like selection, crossover, mutation and inversion [7, 8].

2. Literature Review

Aggarwal et al. [9] used buffer management and pruning techniques to generate all significant association rules. Han et al. [10] discussed about the current status of the frequent pattern mining and all the future directions of the data mining. Le and Ong [6] presented an approach for extracting knowledge on search dynamics of binary GA. Ghosh, and Nath [11] used Pareto based Genetic Algorithm for extracting useful and interesting rules using support count, comprehensibility and interestingness as multiple objectives. Fidelis et al. [12] proposed a GA based classification algorithm which discovered comprehensible if-then-else rules using flexible chromosome encoding where each chromosome represented a classification rule. Alcala-Fdez et al. [13] studied three method of generating association rules using genetic algorithms for extraction of quantitative association rules. Vaani and Ramaraj [14] discussed about the Apriori algorithm using multiple minimum supports to mine offer interesting rules. Wakabi- Waiswa et al. [15] designed an algorithm using a combination of genetic algorithm and a modified Apriori algorithm which yielded fast results. Martinez et al.[16] used principal component analysis to reduce the number of features and reported improvement as compared to when principal component analysis is not used. Ghosh et al. [17] discussed the advantages of using GA in the extracting frequent itemets.

3. Methodology

The 1984 United States Congressional Voting Records Dataset is used for evaluation [4]. This data set includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the CQA. The workflow of proposed algorithm is as follows:

a. Load a dataset in the system which fits in the memory,
b. Generate the frequent itemsets using Apriori algorithm,

c. Encode the generated item sets in binary form by using encoding scheme,
d. Apply genetic algorithm to mine association rules using ranks assigned using non-dominance property.
e. Calculate the efficiency of the proposed approach.

Many parameters are used to evaluate the rules generated. In this work the parameters used are lift, confidence andsupport. The formulas for the same are given below:

$$confidence = \frac{\text{number of transactions containing the items in A and C}}{\text{Number of transactions containing the items in A}}$$

$$support = \frac{\text{number of transactions containing the items in A and C}}{\text{total number of transactions}}$$

In classical association rule mining rules with higher support and confidence values are discovered. A rule is said to be strong if its support and confidence is greater than the user defined threshold value. But this is not always true as the confidence value can be high even if the items do not depend on each other. So to find rules having positive correlation between item sets, lift measure is used which is derived by the given formula:

$$lift = \frac{\text{support}(A \rightarrow C)}{\text{support}(A) * \text{suppord}\ (C)}$$ [18]

Lift interestingness measure defines the number of transactions that contain the items used to find interesting patterns [18]. The lift of the rule is used to relate the frequency of co-occurrence i.e. the confidence to the expected frequency of co-occurrence i.e. expected confidence under the assumption of conditional independence. The lift values can be 1, greater than 1 or smaller than 1 indicating no correlation, a positive correlation or a negative correlation respectively.

4. Results & Discussion

The proposed work had been carried out using MATLAB. Initially population is taken 20. The probability of mutation (pm) is taken 0.5. The numbers of rules generated are analyzed on the lift parameter for the strength evaluation. Table 1 is evaluated by keeping the support constant.

It can be concluded from table 1 that when support is kept constant and confidence is gradually increased, number of rules gets decreased. Time required to generate rules is also decreased. From table 2 it is clear that when value of support is increased with confidence kept constant, lesser rules are generated. The time it takes to mine these rules also start decreasing. It can be seen from the table that on a support value 0.4988, only three rules are generated. On decreasing support to 0.3508 the number of rules increased to 679.

Minimum Support (threshold)	Number of rules generated	Total time taken (in seconds)
0.2567	12231	260.306
.2800	5378	121.324
.3311	1105	22.610
.3508	679	16.838
.4988	3	2.973

TABLE 1(When support is kept constant=.2988)

Minimum confidence	No. of rules generated	Total time taken (in seconds)
0.6	8103	84.687
0.7	7335	80.331
0.8	5626	66.235
0.9	2942	54.059
1.0	110	45.547

TABLE 2 (When confidence is kept constant = 0.9)

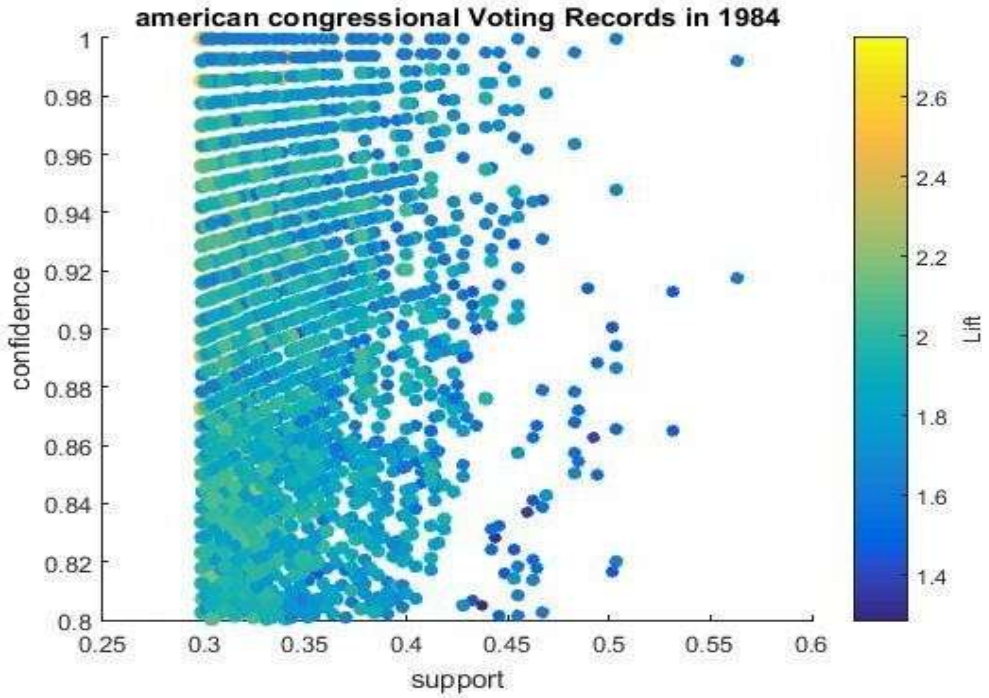


Figure 1: When Support is constant and Confidence is 0.8 and above

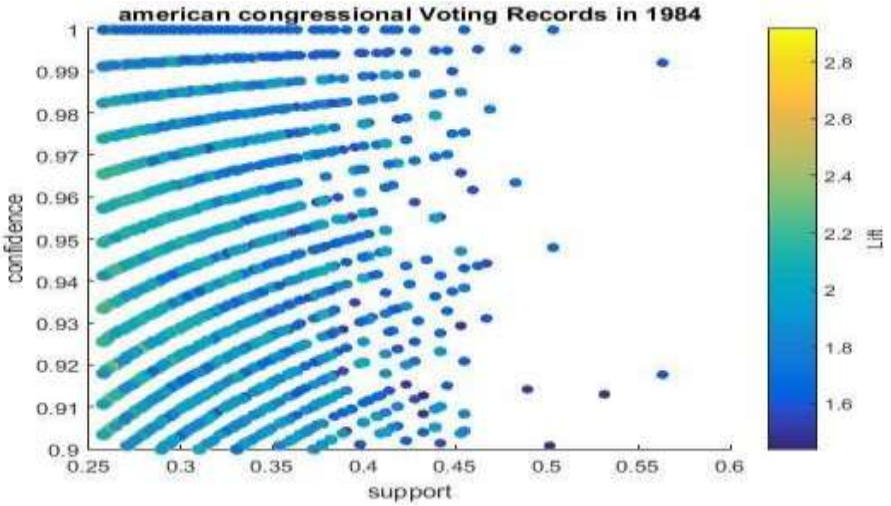


Figure 2: When Confidence is constant and Support is 0.2567 and above

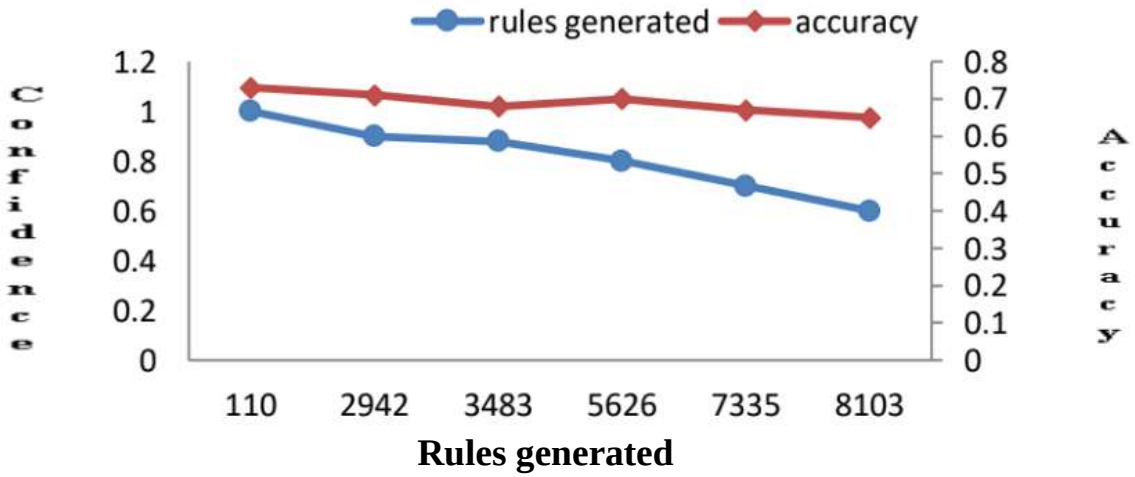
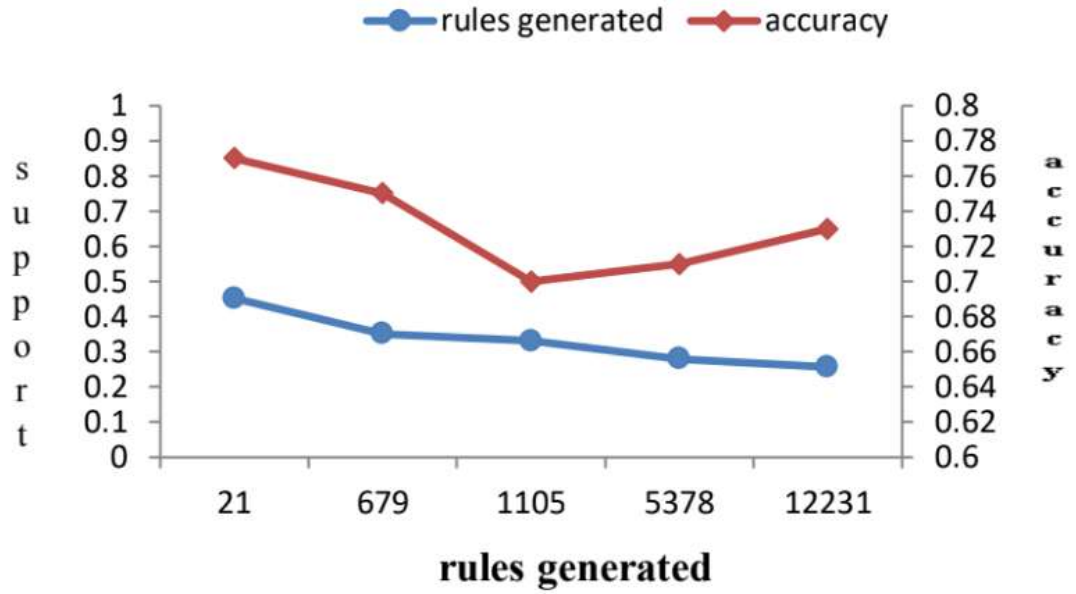


Figure 3: Accuracy and number of rules generated when Support =0.2988



4: Accuracy and number of rules generated when Confidence =0.9

It can be seen that when support is kept constant and confidence is increased, the relative accuracy starts increasing but at some point it drops by some percentage. But as we kept on increasing the confidence, the accuracy again starts increasing. Accuracy fluctuates between 65 percent and 73 percent. The scale taken for accuracy is [0, 1]. It is easily analyzed that when confidence is kept constant the accuracy initially decreases when we increase the support but it then starts increasing gradually. Numbers of rules generated are comparatively low when support is high. The accuracy fluctuates between 70 percent and 80 percent. The scale taken for accuracy is [0, 1].

On applying the described approach different results are obtained. It can be observed that increase in confidence or support keeping the other constant criteria results in the reduction of number of rules. But the rules generated are interesting which is measured using the value of lift.

Some of the rules which are discovered are as follows:

R1: {E1 Salvador=Yes, Budget Resolution=No, Mx Missile= No} => {Republican}i.e.

Lift= 2.36. As the value of the lift parameter for this rule is high it is clear that the rule is quite strong.R2: {E1 Salvador= No, Budget Resolution=Yes, Mx Missile= Yes} => {Democrat}
Lift=1.59.

R3: {Physician Fee Freeze=No, Right to Sue=No, Crime=No} => {Democrat}

Lift 1.65 .As the value of the lift parameter for these rules is below 2 which is not very high, it implies that the rule generated is an average rule.

It can be seen from above that sometimes the rules are accurately defined but these are not interesting enough to be taken into consideration. Also if large number of rules are discovered, it becomes difficult to interpret the rules and time taken is also more. So using the value of lift can help user to mine those rules which are more useful and interesting for the user and hence reducing the time taken.

Hussein et al [19] reported better performance on data provided by applied science university by applying association rule mining and using lift parameter for analysis of rules led to generation of only those rules in which user was interested. Deora et al [20] concluded that on large datasets lift achieved a better reduction in number of rules was done by lift parameter. Montella [21] used association rule mining on accident data to mine contributory factors for crash and the relationships between them. Ordonez [22] used association rule mining on heart disease dataset to find factors responsible for heart disease and lift parameter was used by the author to prune rule set and also to check which attribute was more linked to presence or absence of disease. Ordonez et al [23] reported that lift helps selecting rules with high predictive power and is used in conjunction with confidence to evaluate the significance of discovered rules on heart disease dataset collected patients admitted to hospital .

5. Conclusion

From this work it can be concluded that the approach which is described is better in terms of time and rules generated. A problem with this approach is that accuracy is not too high, so in future accuracy can be improved with the help of better fitness function. This approach can be extended by taking other parameters into account to check whether rule is strong or interesting. The efficiency of this approach can be measured on other datasets. Other criteria's can also be implemented to increase rule comprehensibility. This technique can be combined with other techniques to give better results.

6. References

1. Han J, Pei J, Tong H. Data mining: concepts and techniques. Morgan kaufmann; 2022 Jul 2
2. Witten IH, Frank E. Data mining: practical machine learning tools and techniques with Java implementations. Acm Sigmod Record. 2002 Mar 1;31(1):76-7.
3. Kumar D & Beniwal S. Genetic algorithm and programming based classification: A survey. Journal of Theoretical and Applied Information Technology. (2013)54. 48-58.
4. *https:/s/archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records*
5. Zhang C, Zhang S, editors. Association rule mining: models and algorithms. Berlin, Heidelberg: Springer Berlin Heidelberg; 2002 Jul.
6. Le MN, Ong YS. A frequent pattern mining algorithm for understanding genetic algorithms. In Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence: 4th International Conference on Intelligent Computing, ICIC 2008 Shanghai, China, September 15-18, 2008 Proceedings 4 2008 (pp. 131-139). Springer Berlin Heidelberg.
7. Holland JH. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT press; 1992 Apr 29.
8. Mitchell M. An introduction to genetic algorithms. MIT press; 1998 Mar 2.
9. Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. InProceedings of the 1993 ACM SIGMOD international conference on Management of data 1993 Jun 1 (pp.207-216).
10. Han J, Cheng H, Xin D, Yan X. Frequent pattern mining: current status and future directions. Data mining and knowledge discovery. 2007 Aug;15(1):55-86.
11. Ghosh A, Nath B. Multi-objective rule mining using genetic algorithms.

Information Sciences. 2004 Jun 14;163(1-3):123-33.
12. Fidelis MV, Lopes HS, Freitas AA. Discovering comprehensible classification rules with a genetic algorithm. InProceedings of the 2000 congress on evolutionary computation. CEC00 (Cat. No. 00TH8512) 2000 Jul 16 (Vol. 1, pp. 805-810). IEEE.
13. Alcala-Fdez J, Flugy-Pape N, Bonarini A, Herrera F. Analysis of the effectiveness of the genetic algorithms based on extraction of association rules. Fundamenta Informaticae. 2010 Jan 1;98(1):1-4.
14. Vaani MK, Ramaraj E. An integrated approach to derive effective rules from association rule mining using genetic algorithm. In2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering 2013 Feb 21 (pp. 90-95). IEEE.
15. Wakabi-Waiswa PP, Baryamureeba V, Sarukesi K. Optimized association rule mining with genetic algorithms. In2011 Seventh International Conference on Natural Computation 2011 Jul 26 (Vol. 2, pp. 1116-1120). IEEE.
16. Martínez-Ballesteros M, Martínez-Álvarez F, Troncoso A, Riquelme JC. Selecting the best measures to discover quantitative association rules. Neurocomputing. 2014 Feb 27;126:3-14.
17. Ghosh S, Biswas S, Sarkar D, Sarkar PP. Association rule mining algorithms and Genetic Algorithm: A comparative study. In2012 Third International Conference on Emerging Applications of Information Technology 2012 Nov 30 (pp. 202-205). IEEE.
18. Oweis NE, Fouad MM, Oweis SR, Owais SS, Snasel V. A novel Mapreduce lift association rule mining algorithm (MRLAR) for big data. International Journal of Advanced Computer Science and Applications. 2016;7(3).
19. Hussein N, Alashqur A, Sowan B. Using the interestingness measure lift to generate association rules. Journal of Advanced Computer Science & Technology. 2015 Jan 1;4(1):156.
20. Deora CS, Arora S, Makani Z. Comparison of interestingness measures: support-confidence framework versus lift-irule framework. IJERA International Journal of Engineering Research and Applications. 2013 Mar;3(2).
21. Montella A. Identifying crash contributory factors at urban roundabouts and using association rules to explore their relationships to different crash types. Accident Analysis & Prevention. 2011 Jul 1;43(4):1451- 63.
22. Ordonez C. Comparing association rules and decision trees for disease prediction. InProceedings of the international workshop on Healthcare information and knowledge management 2006 Nov 11 (pp. 17-24).
23. Ordonez C, Ezquerra N, Santana CA. Constraining and summarizing association rules in medical data. Knowledge and information systems. 2006 Mar;9:1-2.