Volume 5 Issue IV, April 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Optimization of Association Rule Mining using FP_Growth Algorithm with GA

Abhishek Kumar Singh[1], Deepak Sinwar[2]
[1]M.Tech Scholar, [2]Assistant Professor, Department of Computer Science & Engineering
BRCM College of Engineering & Technology, Bahal, Bhiwani, Hry. (INDIA)

Abstract: ***Frequent pattern mining is one of the active research themes in data mining which covers a broad spectrum of data mining tasks viz. Association rules, correlations, causality, ratio rules, emerging patterns etc. In this paper, we expand the horizon of frequent pattern mining by introducing an efficient algorithm for mining multi-level and multi-dimensional frequent patterns with flexible support constraints. We analyze the scalability of our algorithm and study its performance on different data sets. We have tested our two hybrid algorithms viz. Apriori+fpga and firefly+fpga with traditional apriori and ga. The objective of this paper is to compare the performance of the genetic algorithm for association rule mining by combining it with other algorithms. The algorithms when tested on abalone dataset that indicates that the accuracy depends mainly on the fitness function which is the key parameter. The crossover probability brings changes in convergence rate with minimal changes in accuracy. The size of the dataset and relationship between its attributes also plays a role in achieving the optimum accuracy. Theoretical analysis and experimental results show that the performance of firefly+fpga is better than other algorithms, however the performance of apriori+fpga was found better than traditional apriori and ga.***
Keywords: apriori algorithm, genetic algorithm, association rule mining, fp-growth

I. INTRODUCTION

Data mining is a promising and relatively new technology and it is defined as a process of discovering hidden, valuable information by analyzing large amount of data storing in databases or data warehouse by means of different techniques such as machine language, Artificial Intelligence (AI), and statistics etc. Machine Learning, a branch of Artificial Intelligence used for learning/ training purpose. Some types of learning are supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning.

Generally there are three major techniques through which we can achieve Data Mining viz. Association Rule Mining, Classification and Clustering. These techniques have richer application areas i.e., retail industry, telecommunication industry, biological data analysis, intrusion detection and aerospace etc. This paper belongs to one of very famous technique of Data Mining i.e. Association Rule Mining which is used to find interesting yet hidden associations between set of items. Large-scale organizations apply various data mining techniques on their data, to extract useful information and patterns. Knowledge discovery in database is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in the data [7]. Let's have a brief introduction about Association Rule Mining technique as follows:

A. Association Mining

Association mining searches for interesting relationship among items in a given database and displays it in a rule form, i.e. *A=>B*. With the massive amounts of data continuously being collected and stored in databases, many industries are becoming interested in mining associations among data. Market basket analysis is a typical example among the various applications of association mining.

1) *Support:* The support of an association pattern is the percentage of task-relevant data transactions for which the pattern is true
2) *Confidence:* Confidence is defined as the measure of certainty or trustworthiness associated with each discovered pattern

B. Importance of Association Rule Mining

As stated above there are numerous applications of data mining techniques. Some of the importance of association rule mining is as follows:

1) The association rule mining helps in finding particular relationships between various data elements of the large database i.e. database having a large number of records.
2) It helps to search useful information and knowledge that can be used to enhance the business or scientific operations.

3) Helps in finding the outlier entries, which may be useful in some cases such as fraud detection.

C. Apriori Algorithm

This is one of the traditional approaches used to find frequent item sets. The algorithm performs a level-by-level search starting from the single-item sets and proceeding to larger sets. The variable holds the family of candidate itemsets, whose frequencies are counted in one database the frequent candidates are selected into D. If the sets in D have size k, the new candidates are those sets of size k + 1 whose all size-k subsets are in D. It is possible to reduce the number of pairings significantly. The Apriori algorithm examines not only all frequent itemsets but also those non-frequent itemsets whose all subsets are frequent.

D. Genetic Algorithm

Genetic Algorithm is a search heuristic that mimic the process of natural evolution. This heuristic is routinely used to generate useful solutions for 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 etc. 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].

E. FP-Growth Algorithm

FP-growth algorithm is proposed by Jiawei Han that finds the association rules more efficiently compared to Apriori algorithm without the generation of candidate itemsets. The Apriori algorithm will require n+1 scans, where n is known as the length of the longest pattern. The FP-growth algorithm will require only two scans of the database for finding frequent patterns. FP-growth algorithm will adopts divide and conquer strategy. Initially, it will construct a FP-tree by using the data present in transactional database and then it mines all the frequent patterns taken from FP-tree. The association rules can be generated easily after mining of frequent patterns.

II. PROBLEM STATEMENT

The explosive growth of many business, government and scientific databases has far outpaced our ability to interpret and digest the data. We are drowning in information yet starving for knowledge. Data mining therefore appears as a useful tool to address the need for sifting useful information such as hidden patterns from databases. Frequent pattern mining is one of the active research themes in data mining. It covers a broad spectrum of data mining tasks, including mining various kinds of association rules, correlations, causality, ratio rules, emerging patterns and etc.

In this paper, we expand the horizon of frequent pattern mining by introducing an efficient algorithm for mining multi-level and multi-dimensional frequent patterns with flexible support constraints. We analyze the scalability of our algorithm and study its performance on different data sets. Multi-level and multi-dimensional frequent pattern mining is a very promising research topic and plays an invaluable role in real life applications. In this section, we review related concepts and give the definition of multi-level multi-dimensional frequent pattern mining.

III. IMPLEMENTATION

In this implementation, we have used Genetic Algorithm (GA) for mining association rules from the prepared database. The genetic algorithm is a heuristic (sometimes called meta-heuristic) which utilized to solve optimization problems by using approaches inspired from nature. This algorithm works as below:

A. Begin by choosing initial population using Apriori algorithm.
B. Evaluate the individual fitness of a certain proportion of the population.
C. Select pairs of best-ranking individuals to reproduce.

D. Apply crossover operation.
E. Apply mutation operation.
F. Apply local search until terminating condition.
G. End.

This method has two main differences as compared with traditional Genetic Algorithm. First, in this approach the initial population is not generated randomly but in the GA the initial population is randomly selected, which is less effective. The initial population is selected here by using the Apriori algorithm. Secondly this model utilizes micro genetic local search which is a small version of genetic algorithm. In this model, the output of genetic algorithm on the initial operation (here we call this Current Population), would be the input of micro genetic algorithm then the following operations is done on this population which is called current population in order to generate a new population. Here, every chromosome is defined as a row of this database. So a population or a chromosome consists of a binary array of items which consists of 'One' (i.e. the customer purchased that item) and 'Zero' (i.e. the customer didn't purchase that item).Therefore, first we need to create an initial population for genetic algorithm which (as we mentioned) is done by using the Apriori algorithm i.e., The frequent patterns generated by Apriori. We define objective function as 'Support'. The overall hybrid algorithm is given below:

H. Algorithm 1: Apriori+FPGA

Step 1: Start Step
Step 2: Load a sample of records from the database that fits in the memory.
Step 3: Apply Apriori algorithm to find the frequent item sets with the minimum support. Suppose A is set of the frequent item set generated by Apriori algorithm.
Step 4: Set Z= 0 where Z is the output set, which contains the association rule.
Step 5: Input the termination condition of FP-GA.
Step 6: Represent each frequent item set of A as a binary string using the combination of representation.
Step 7: Select the two members from the frequent item set using Roulette Wheel sampling method.
Step 8: Apply the crossover and mutation on the selected members to generate the association rules.
Step 9: Find the fitness function for each rule X*Y and check the following condition.
Step 10: If (fitness function > min confidence)
Step 11: Set Z = Z U {X * Y}
Step 12: If the desired number of generations is not completed, then go to Step 3.
Step 13: Stop.

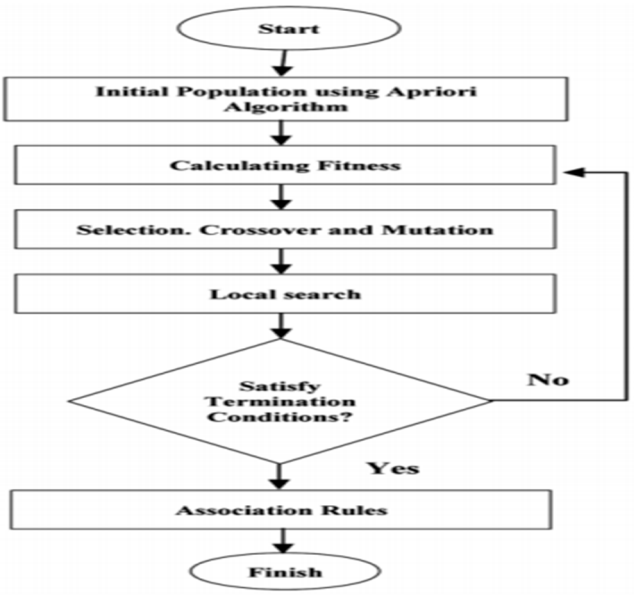

Figure 1: Outline of proposed Genetic Algorithm

Volume 5 Issue IV, April 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

I. Algorithm 2: Hybrid Firefly and FPGA

In this method, each firefly deliberates as a rule and computes fitness value of every firefly. Using this method tried to discovery out the high-frequency association rules. The projected procedure covers two fragments. The first part provides procedures connected to encoding and calculating the fitness values of the firefly swarm. In the second part of the procedure, which is the main influence of this study, the firefly algorithm is working to mine the association rules. The pseudo code of this method is revealed in below steps. In this, first specify t as an iteration number. The algorithm is run as number of iterations or number of desired rules. At the beginning of the algorithm, the Discovered Rules set is empty. At the first iteration of algorithm, each firefly is initialized randomly as a rule. In each of iterations, until reaching the termination conditions, calculate fitness of each fireflies. If light intensity of j^{th} firefly is larger than i^{th} firefly then calculate distance between them. The individuals of population are sorted in descending order according to their distance and then select minimum distance between i^{th} and j^{th} firefly and then move i towards j and increase the intensity of j^{th} firefly

Input: Database of Abalone

Output: Best discovered association rules

1. t=0 // here t is iteration number

2. Discovered Rules = ∅

3. Generate initial population of fireflies xi (i=1,2,...n)

4. While (t < no of iterations)

5. Compute objective function of each firefly

6. for i=1: n (all n fireflies)

7. for j=1: i (all n fireflies)

8. if(fit(j) > fit(i))

9. Calculate distance rij = || xi – xj ||

10. end if

11. sort distance in descending order

12. min ← firefly[min distance index]

13. end for j 14. Movement phase:

15. j = min distance index

16. xi = xi + f(xi)×(xj-xi) + α(rand -1/2)

17. Ij = Ij + xi //increase the intensity of firefly j

18. End for i

19. Best = firefly [Bestindex]

20. Until not terminate

21. Discovered Rules = Best ∪ Discovered Rules

22. t++ // End of algorithm

Volume 5 Issue IV, April 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

IV. RESULT

In each iteration the best discovered rule is added to Best vector and combines with Discovered Rules. The rule is valid if it has at least one attribute in the antecedent of the rule and one in the consequence of the rule. This process is continued until termination conditions not occur.

A. Result Analysis

This section describes the experimental results obtained by applying the proposed algorithms to a variety of data sets. This research proposes FP- Association Rule Mining (Hybridization) through Genetic Algorithm with sorted initial population and applies it to solving real time Abalone problem. Results shows that Firefly algorithm is more useful and faster in discovering association rules in a transactional database and it also reach the answer in less repetition. It obtains better answers in compare with other algorithms like GA, Apriori and FP-Growth. We have compared our model with Genetic, Apriori and Apriori FPGA with Firefly- FPGA algorithm. As we can see in the figure 2, Firefly- FPGA created better rules than Apriori and Genetic Time consumption for creating model using Hybrid Apriori FP-GA algorithms is very low although both FP-Growth and Apriori algorithms are successful in some situations, but they use a lot of memory. According to this experiment, our method has a good performance in compare with Apriori and GA.

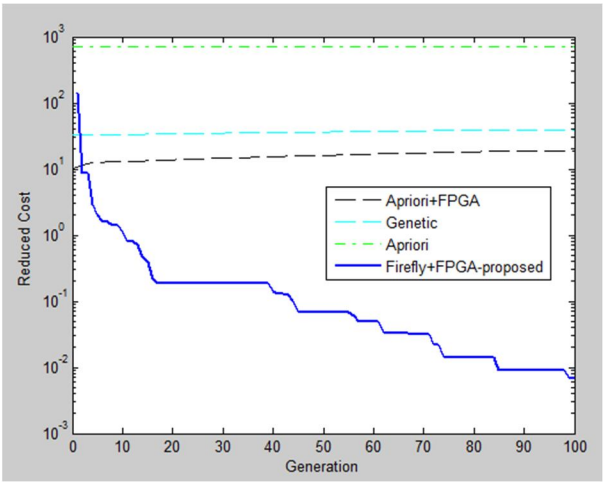

Figure 2: Comparison between Apriori, GA, Apriori + FPGA and Firefly+FPGA for the generation and fitness

The experiment is done on Abalone dataset obtained from UCI machine learning repository. The data set has 4177 samples. This paper has implemented a hybrid approach called “Hybrid Apriori FPGA (FP_Growth with GA)” for frequent Strong rule generation in data. Generally Apriori algorithm is used to generate the rules, and we have to implement any optimization technique in order to refine the rules. Here ‘Genetic Algorithm’ is one of the best ways to optimize the rules from the qualitative and quantitative results. As shown in Figures 2 and 3 ‘Hybrid Firefly FPGA” works properly and generated rules very efficiently.
The figure 2 and 3 shows the comparison between Apriori, GA and Firefly+ FPGA for the cost, generation and support/confidence.

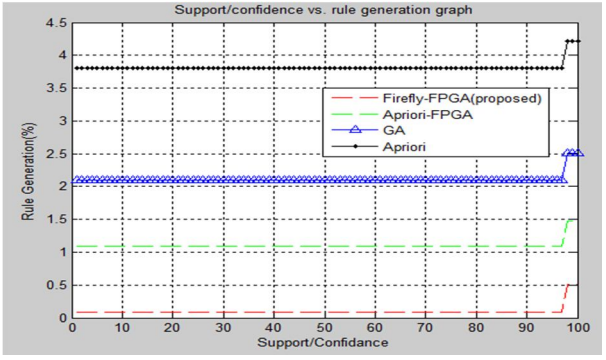

Figure 3: Comparison between Apriori, GA, Apriori + FPGA and Firefly+FPGA for the Rule generation ratio for support/confidence

Figure 3 show the number of rules generated. The number of rules generated in hybrid (Apriori+ FP) with GA algorithm is approximately 60% less than the number of rules generated in Apriori algorithm with GA. Thus the proposed hybrid (Firefly+FPGA) with GA algorithm gives the better results.

V. CONCLUSION

The work reported in this paper presents an efficient approach for exploring high quality association rules. The modified approaches are called Hybrid "Apriori+FPGA" and "Firefly+FPGA". In these approaches, chromosome of matrix is applied to extract rules from database. For this purpose, fitness value of individual rule is computed instead of minimum support and minimum confidence thresholds. This scheme has an advantage that the database is scanned once only which enhances efficiency of the approach in terms of CPU time and memory consumption. This approach was implemented in MATLAB and the result was compared with existing approaches of the same kind. Experimental results show that "Apriori+FPGA" outperforms other traditional algorithms in terms of relationship between population size and computation time, whereas "Firefly+FPGA" outperforms all other algorithms in terms of conciseness and cost reduction.

REFERENCES

[1] M., Suraj Kumar Sudhanshu, Ayush Kumar and Ghose M.K., "Optimized association rule mining using genetic algorithm Anandhavalli Advances in Information Mining" , ISSN: 0975–3265, Volume 1, Issue 2, 2009, pp-01-04
[2] Margret H. Dunham, S.Sridhar, "Data mining Introductory and advanced topics", Pearson Education,Second Edition,2007.
[3] Daniel Hunyadi, "Performance comparison of Apriori and FP-Growth algorithms in generating association rules", Proceedings of the European Computing Conference
[4] Time.Dr (Mrs).Sujni Paul, "An Optimized Distributed Association Rule Mining Algorithm In Parallel And Distributed Data Data Mining With XML Data For Improved Response", International Journal of Computer Science and Information Technology, Volume 2, Number 2, April 2010
[5] S. Rangaswamy, Shobha G., "Optimized Association Rule Mining Using Genetic Algorithm," Journal of Computer Science Engineering and information Technology Research (JCSEITR), Vol.2, Issue 1, pp 1-9, 2012.
[6] S. Jain, S. Kabra. "Mining & Optimization of Association Rules Using Effective Algorithm," International journal of Emerging Technology and Advanced Engineering (IJETAE), Vol.2, Issue 4, 2012
[7] Jun Gao, "A New Association Rule Mining algorithm and Its Applications", IEEE 3rdInt. Conf. on Advanced Computer Theory and Engineering (ICACTE), vol 5, pp. 122-125,2010.
[8] Li Juan and Ming De-ting, "Research of an association rule mining algorithm based on FP tree", IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS), Vol. 1, pp. 559-563,2010.
[9] Zhi Liu, Mingyu Lu, Weiguo Yi, and HaoXu, "An Efficient Association Rules Mining Algorithm Based on Coding and Constraints", Proceedings of the 2nd International Conference on Biomedical Engineering and Informatics, pp. 1-5, 2009.
[10] WanjunYu ,XiaochunWang, and Fangyi Wang, "The Research of Improved Apriori Algorithm for Mining Association Rules", 11th IEEE International Conference on Communication Technology Proceedings, pp. 513-516, 2008.
[11] X. Wu, C. Zhang, and S. Zhang, ―Efficient Mining of Both Positive and Negative Association Rules,‖ ACM Transactions on Information Systems, Vol. 22, No. 3, 2004, pp. 381–405.
[12] RupeshDewang et al.‖ A New Method for Generating All Positive and Negative Association Rules‖ International Journal on Computer Science and Engineering (IJCSE), ISSN: 0975-3397 Vol. 3 No. 4 Apr 2011.
[13] Honglei Zhu, ZhigangXu, ―An Effective Algorithm for Mining Positive and Negative Association Rules‖ International Conference on Computer Science and Software Engineering 2004 IEEE Computer Society Press.
[14] Estefan G.M de L Manoelatl.,' Minimum number of switching operations via ant colony optimization', 19th internation conference on electricity distribution Vienna, 21-24 May 2007
[15] R. Agrawal,T. Imielinski, A. Swami,' Mining Association rules between sets of items in large database,' proceeding of 1993 ACM SIGMOD conference Washington DC, USA.
[16] AK Jain, MN Murthy, PJ Flynn,' Data clustering- A review', ACM Computing Surveys, Oct. 2001
[17] M Dorigo , T Stutzle, 'Ant colony optimization', The MIT press Cambridge, MA.
[18] Karla Taboada, S Mabu, E Gonzales,' Genetic Network programming for fuzzy association rule based classification, 2009
[19] Karla Taboada,ShingoMabu, Eloy Gonzales," Genetic Network Programming for Fuzzy Association Rule-Based Classification", 2009