ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 9, September 2015

Eclat-Genetic Approach for Finding Association Rule

Mr. Omkar B. Bhalerao[1], Prof. L. M. R. J. Lobo[2]

Computer Science and Engineering Department, Walchand Institute of Technology, Solapur, India[1,2]

Abstract: Data mining is the process of knowledge discovery in databases. It is the art and science of intelligent analysis of existing large data sets, for finding meaning and previously unknown insights and transforms it into a flexible structure. With the help of Database and data mining we can extract meaningful data sets from huge data. When we apply data mining techniques on huge datasets results of improved quality are achieved. In data mining algorithms, association rule mining finds an important place; it is an easy and popular method to find out association rules from an existing large datasets. In general frequent itemsets are generated from large data sets by applying association rule mining this task takes a lot of computer time to compute all the frequent itemsets.
In this paper, the main area of concentration was to optimize the rules that are generated by an Association Rule Mining algorithm (Eclat) by using a Genetic Algorithm. Here we generate more accurate and complete rules. The advantage of using genetic algorithm is to discover high level prediction rules.

Index Terms: Eclat, Genetic Algorithm, Association Rule, Data Mining.

I. INTRODUCTION

I In this high technical era, nowadays all fields are computerized. Each and every sector has a huge amount of data which becomes a candidate for data mining. If we apply the proper technique on it, this data no longer remains only pieces of data but, it gives lot of pattern from it. In this we need to make inferences from the immense data so that we can make decisions driven by knowledge. [1]
Data mining is the process of knowledge discovery in database. It is art and science of intelligent analysis of large data sets for meaning and previously unknown insights and is nowadays actively applied. With the help of Knowledge Discovery in Database (KDD) and data mining we extract the meaningful data sets from the large amount of data [2]. So, on the large data sets when we applied data mining techniques then it gives results into improved quality of mined data. Data mining is popularly known as "Knowledge Discovery in Database (KDD)" [2]. Data mining tools are powerful generating rules from vast & diversified datasets which are in the huge amount. Generally, data mining is the process of analyzing data from a different perspective and summarizing it into useful information.
In data mining there are various methods which are applied over the huge amoung of data and we get some pattern or knowledge from it. For optimization of solution or result we use Genetic algorithm. Genetic Algorithm is a randomized algorithm that could be run for a very long time to obtain an optimal solution. The main purpose of association rule mining is to find out the hidden relationship between different data item sets in the database. [3, 4]

Association Rule Mining

Association Rule Mining is the process of finding new interesting Correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories [5]. In data mining, association rule mining is an important and easy method to find frequent item sets from large dataset.
It is intended to identify strong rules discovered in databases using two different measures of interestingness. The first one is support which generates frequent item set from the provided database and the other one is confidence which is focuses on rule generation. [6, 7, 8]

Frequent item sets- A set of attributes is termed as frequent item set if the occurrence of the set within the database is more than a user given threshold.

Support- Support determines how often a given rule is applicable to a given data set.

Confidence- Confidence determines how frequently items in Y appear in transactions that contain X.[9]

$$\text{Support,} \quad s(X\text{->}Y) = \frac{\sigma(X \cup Y)}{N}$$

$$\text{Confidence,} \quad c(X\text{->}Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

Where, X and Y disjoint item set.

Genetic Algorithm-

Genetic Algorithms (GA) are direct, parallel methods for global search and optimization. GA is one of the most commonly used Evolutionary Algorithms (EA). Genetic Algorithms are efficient method for optimization. As they use populations with allowed number of solutions (individuals), they are added in the group of parallel algorithms. [10] Main ingredients of GA are Chromosomes, Selection, Recombination and Mutation.

Selection - During each successive generation, a proportion of the existing population is selected to breed a new generation. Fitness-based process is used to select individual solutions where fitter solutions (as measured by

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 9, September 2015

a fitness function) are typ ically more likely to be selected. At this stage elitism could be used – the best n individuals are directly transferred to the next generation. The elitism ensures, that the value of the optimization function cannot get worst (once the extremism is reached it would be kept).

Crossover – The most common type is single point crossover. In single point crossover, we choose a locus point at which you swap the remaining alleles from one parent to the other. The children take one section of the chromosome from each parent. Chromosome is broken based on the randomly selected crossover point. This particular method is called single point crossover because only one crossover point exists. Sometimes only one child is created, but generally both offspring are created and put into the new population. Crossover does not always occur. Sometimes, based on a set probability, no crossover occurs and the parents are copied directly to the new population.

Mutation – After selection and crossover, we have a new population full of individuals where some are directly copied, and others are produced by crossover. In order to ensure that the individuals are not all exactly the same, we allow a small chance of mutation. Mutation is fairly simple. Mutation is, however, vital to ensuring genetic diversity within the population. [6, 11]
Basic block diagram of Genetic Algorithm is:

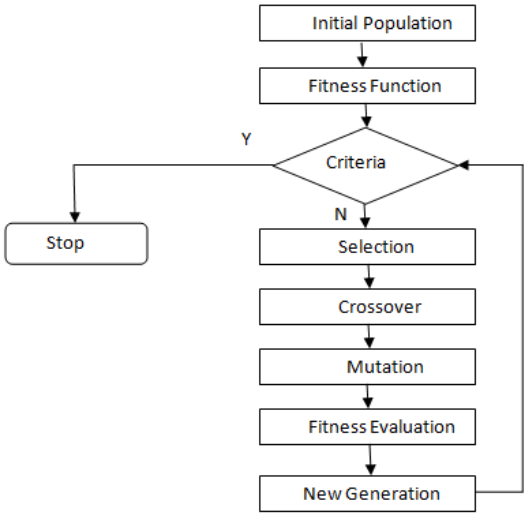

Figure 1: Genetic Algorithm

Genetic Algorithm is a randomized algorithm that could be run for a very long time to obtain an optimal solution.

II. LITERATURE REVIEW

In the previous research, different techniques were presented for historical agriculture data analysis-

Mahesh Saggar, Ashish Kumar Agrawal, Abhimanyu Lad[3], have considered Association Rule Mining and tried to improve this technique by applying Genetic Algorithms. On the rules generated by Association Rule Mining optimize the rules generated by association rule mining (using apriori technique). They have not consider the negative occurrences of attributes in them, but by using Genetic Algorithms (GA) over these rules the system can predict the rules which contains negative attributes. The main motivation for using GA in the discovery of high-level prediction rules is that they perform a global search and cope better with attribute interaction than the greedy rule induction algorithms open used in data mining. The improvements applied in GA are definitely going to help the rule based systems used for classification as described in results and conclusions.

Peter P. Wakabi–Waiswa, Venansius Baryamureeba, Karunakaran Sarukesi [4] worked on Optimized Association Rule Mining with Genetic Algorithms. Here the authors propose a multi–objective approach to generating optimal association rules using two new rule quality metrics: transactional superiority (Ts) and syntactic superiority (Ss). These two metrics ensure that dominated but interesting rules are returned to not eliminate from the resulting set of rules. Experimental results show that when they modify the dominance relations new interesting rules emerge implying that when dominance is solely determined through the raw objective values there is a high chance of eliminating interesting rules.

Arvind Jaiswal, Gaurav Dubey [6] finds out the best association rules and their optimization using genetic algorithm. They proposed a genetic algorithm based method for finding frequent item sets. Repeatedly transforms the population by executing the following various steps: Fitness Evaluation: The fitness (i.e., an objective function) is calculated for each individual. Selection: Individuals are chosen from the current population as parents to be involved in recombination. Recombination: New individuals (called offspring) are produced from the parents by applying genetic operators such as crossover and mutation. Replacement: Some of the offspring are replaced with some individuals (usually with their parents).
In a literature review we have reviewed different ideas of machine learning techniques for obtaining Association rule mining and have selected Eclat algorithm for obtaining the best item sets only.

III. METHODOLGY

In our methodology we use a two stage model. In the first stage, we apply association rule mining on the historical datasets and generate rules from frequent item sets by applying the proper support and confidence for each rule. The user then gives a minimum support and confidence and based on this initial best rules that form the initial population for GA are extracted. In the second stage, we apply Genetic algorithm to optimize the initial population rules which we get from association rule mining. So that, we will get best rules that predict output as an optimized rules. For demonstration of its utility we give historical agriculture datasets to the proposed model.

The figure 2 shows the first stage of our proposed model where we find out the item sets. By using, Eclat algorithm we obtain the best frequent item sets.

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 9, September 2015

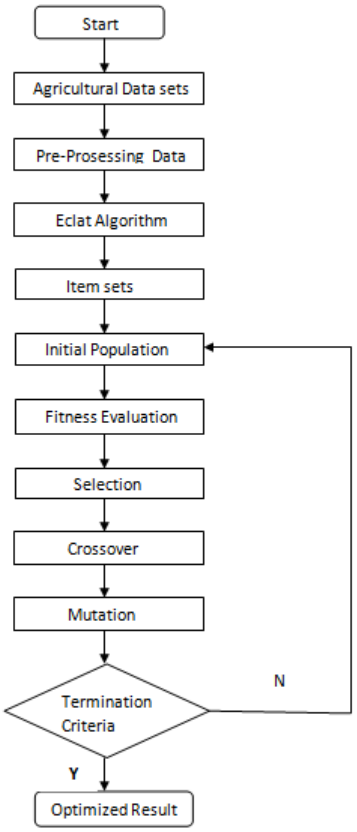

Figure 2: Architecture

In the second stage, frequent item sets are considered as population as an input to GA which initializes its population with frequent item sets. Then GA does the selection, crossover & mutation on population & returns the best population. The termination condition is then given for the genetic algorithm & it will test for the desired output . If desired output is found then it stops Genetic Algorithm otherwise GA continues until a termination condition is meet achieving the maximum fitness.

IV. EXPERIMENTAL SETUP & RESULTS

The system was developed using Java platform and R programming tools. The frequent set items were arrived at using Eclat algorithm of 'apriori' package of R. The GA was developed in JAVA language which was interfaced with the Eclat frequent item sets to generate best association rules.

The testing was performed on the agricultural datasets with different crop rules and result obtained was quite satisfactory. We are showing the chart of 4 crops with Apriori Algorithm Vs Eclat-Genetic Algorithm which is from the system developed. It can be seen from the charts that the prediction accuracy is 28.31% good.
On the basis of above Figure 3, we can say that Eclat-Genetic Algorithm perform 28.31% up well as the prediction accuracy which is quite satisfactory. It can be seen from the chart that the prediction accuracy is 28.31% good.

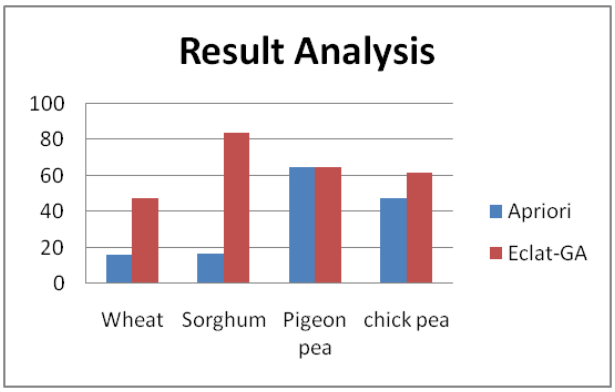

Figure3. Result Analysis

We also show the positive and negative rule generated comparison between Eclat-GA and Apriori Algorithm as shown in Figure 4 and Figure 5. This will help in better performance of predicting required rules.

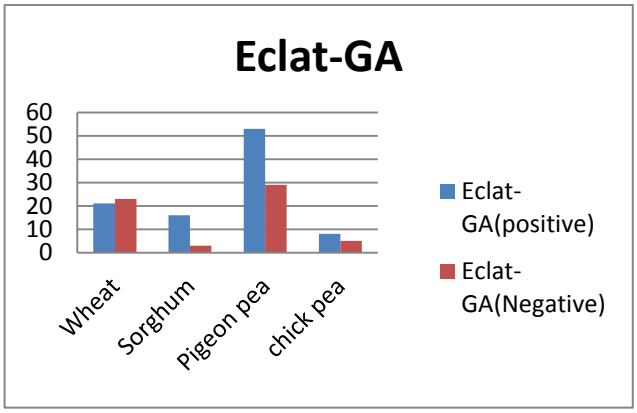

Figure4. Eclat-GA Rule Analysis

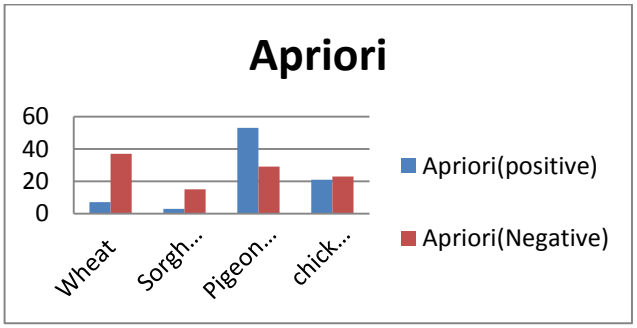

Figure5. Apriori Rule Analysis

V.CONCLUSION AND FUTURE SCOPE

Although a number of works are already published. These have not been applied to agriculture datasets, but in this paper we have tried to use the enormous robustness of Association rule by applying GA on frequent item sets.
The experimental result shows that, the developed model gives better result than the existing technique (Apriori technique) in terms of accuracy. We believe that the toolkit can also handle other databases, after minor modifications. As a future work, the author is currently working on the complexity reduction of Genetic Algorithms by using distributed computing

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 9, September 2015

REFERENCES

[1] V. Umarani, Dr. M. Punithavalli, "A STUDY ON EFFECTIVE MINING OF ASSOCIATION RULES FROM HUGE DATABASES", International Journal of Computer Science and Research Computer Science and Research, Vol.1 Issue 1, 2010.

[2] T. Karthikeyan, N. Ravikumar, "A Survey on Association Rule Mining", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 1, January 2014.

[3] Manish Saggar, Ashish Kumar Agrawal, Abhimanyu Lad, "Optimization Of Association Rule Mining using Genetci Algorithm",2004 IEEE .

[4]Peter P. Wakabi-Waiswa, Venasius Baryamureeba, Karunakaran Sarukesi, "Optimized Association Rule Mining with Genetic Algorithm" ©2011 IEEE ,2011.

[5] Indira K and Kanmani S., "Performance Analysis of Genetic Algorithm for Mining Association Rules", International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012.

[6]Arvind Jaiswal, Gaurav Dubey, "Identifying Best Association Rules and Their Optimization Using Genetic Algorithm", International Journal of Emerging Science and Engineering (IJESE), ISSN: 2319–6378, Volume-1, Issue-7, May 2013.

[7]Agrawal, R., Imielinski, T. and Swami, A.N. (1993) "Database mining: a performance Perspective." IEEE Transactions on Knowledge and Data Engineering, Vol. 5, 914–925.

[8]Ruchi Bhargava, Shrikant Lade, "Effective Positive Negative Association Rule Mining Using Improved Frequent Pattern Tree ", Volume 3,Issue 4,April 2013.

[9] Ms Shweta, Dr. Kanwal Garg , " Mining Efficient Association Rules Through Apriori Algorithm Using Attributes and Comparative Analysis of Various Association Rule Algorithms", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6,June 2013.

[10] Gopesh Joshi, "Review of Genetic Algorithm: An Optimization Technique", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 4, April 2014.

[11] Anubha Sharma, Nirupama Tivari, "A survey of Association Rule Mining Using Genetic Algorithm", International Journal of Computer Application And Information Technology, Vol.1, Issue 2, August 2012.