IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

Eclat - Genetic Approach for Finding Association Rule

Santosh Kumar Satapathy[1], Santosh Kumar Moharana[2], Narendra Kumar routr[3],Avaya Kumar Ojha [4]

Assistant Professor, CSE, Gandhi Engineering College, Bhubaneswar, India [1,2,3,4]

Abstract: Data mining is a process which finds useful patterns from large amount of data The process of extracting previously unknown, comprehensible and actionable information from large databases and using it to make crucial business decisions - Simoudis 1996 This data mining definition has business flavor and for business environments. However, data mining is a process that can be applied to any type of data ranging from weather forecasting, electric load prediction, product design, etc. Data mining also can be defined as the computer-aid process that digs and analyzes enormous sets of data and then extracting the knowledge or information out of it. By its simplest definition, data mining automates the detections of relevant patterns in database. Data Mining is an analytic process designed to explore data (usually large amounts of data - typically business or market related) in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction - and predictive data mining is the most common type of data mining and one that has the most direct business applications. The process of data mining consists of three stages: (1) the initial exploration, (2) model building or pattern identification with validation/verification, and (3) deployment (i.e., the application of the model to new data in order to generate predictions). 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.

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

I. INTRODUCTION

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.

II.DATA MINING METHODOLOGIES

Frequent pattern mining is an important area of Data mining research. The frequent patterns are patterns (such as itemsets, subsequences, or substructures) that appear in a data set frequently. For example, a set of items, such as milk and bread that appear frequently together in a transaction data set is a frequent itemset. A subsequence, such as buying first a PC, then a digital camera, and then a memory card, if it occurs frequently in a shopping history database, is a (frequent) sequential pattern. A substructure can refer to different structural forms, such as subgraphs, subtrees, or sublattices, which may be combined with itemsets or subsequences. If a substructure occurs frequently, it is called a (frequent) structured pattern. Finding such frequent patterns plays an essential role in mining associations, correlations, and many other interesting relationships among data. Moreover, it helps in data classification, clustering, and other data mining tasks as well.

The process of discovering interesting and unexpected rules from large data sets is known as association rule mining. This refers to a very general model that allows relationships to be found between items of a database. An association rule is an implication or if-then-rule which is supported by data. The association rules problem was first formulated in [3] and was called the market-basket problem. The initial problem was the following: given a set of items and a large collection of sales records, which consist in a transaction date and the items bought in the transaction, the task is to find relationships between the items contained in the different transactions. A typical association rule resulting from such a study could be “90 percent of all customers who buy bread and butter also buy milk" – which reveals a very important information. Therefore this analysis can provide new insights into customer behaviour and can lead to higher profits through

better customer relations, customer retention and better product placements. The subsequent paper [4] is also considered as one of the most important contributions to the subject.

Mining of association rules is a field of data mining that has received a lot of attention in recent years. The main association rule mining algorithm, Apriori, not only influenced the association rule mining community, but it affected other data mining fields as well. Apriori and all its variants like Partition, Pincer-Search, Incremental, Border algorithm etc. take too much computer time to compute all the frequent itemsets. The papers [10, 11] contributed a lot in the field of Association Rule Mining (ARM). In this paper, an attempt has been made to compute frequent itemsets by applying genetic algorithm so that the computational complexity can be improved.

III.ASSOCIATION RULE MINING (ARM)

Association Rule Mining [2] techniques can be used to discover unknown or hidden correlation between items found in the database of transactions. An association rule is a rule, which implies certain association relationships among a set of objects such as occurs together or one implies to other in a database. Association rules identify relationships among sets of items in a transaction database. Ever since its introduction in (Agrawal, Imielinski and Swami 1993), Association Rule discovery has been an active research area. Association Rule Mining finds interesting association or correlations among a large set of data items.

Association Rule Mining aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories [8]. The major aim of ARM is to find the set of all subsets of items or attributes that frequently occur in many database records or transactions, and additionally, to extract rules on how a subset of items influences the presence of another subset. ARM algorithms discover high-level prediction rules in the form: IF the conditions of the values of the predicting attributes are true, THEN predict values for some goal attributes.

In general, the association rule is an expression of the form X=>Y, where X is antecedent and Y is consequent. Association rule shows how many times Y has occurred if X has already occurred depending on the support and confidence value.

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]
s(X->Y) =(XUY)/N
Confidence, c(X->Y) =Ω(XUY)/Ω(X)
Where, X and Y disjoint item set.

IV. GENETIC ALGORITHM

Genetic algorithms (GAs) have emerged as potentially robust optimization tools in the last decades. Genetic algorithms (GAs) are a search heuristic that mimics the process of natural evolution. Genetic algorithms (GAs) can be applied to the process controllers for their optimization using natural operators viz. mutation and crossover [18]. Well established methodologies have been discussed in literature for integrating soft computing techniques to realize synergistic or hybrid models with which better results could be obtained. Simulation is the computational realization of a model. They are executable, live representation of models that can be as meaningful as the real experiments. Simulation allows an engineer to reason if a model makes sense or not and how the model behaves for the certain parameter variations. Simulations can be carried out for designing and implementation of conventional proportional integral derivative (PID) controllers, fuzzy logic controllers (FLC) and hybrid fuzzy logic genetic algorithms (HFLGA) controllers. Simulation applications can dynamically adjust the various process control parameters at running state of the plant.

GAs are one of the best ways to solve a problem for which little is known. They are a very general algorithm and so will work well in any search space. The Genetic Algorithm [5] was developed by John Holland in 1970. GA is stochastic search algorithm modeled on the process of natural selection, which underlines biological evolution [6].

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 [7].

Standard GA apply genetic operators such selection, crossover and mutation on an initially random population in order to compute a whole generation of new strings. 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.This generational process is repeated until a termination condition has been reached.

Common terminating conditions are:
- A solution is found that satisfies minimum criteria
- Fixed number of generations reached
- Allocated budget (computation time/money) reached
- The highest ranking solution's fitness is reaching or has reached a plateau such that successive iterations no longer produce better results
- Manual inspection
- Combinations of the above

Simple generational genetic algorithm pseudo-code:
- Choose the initial population of individuals

IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

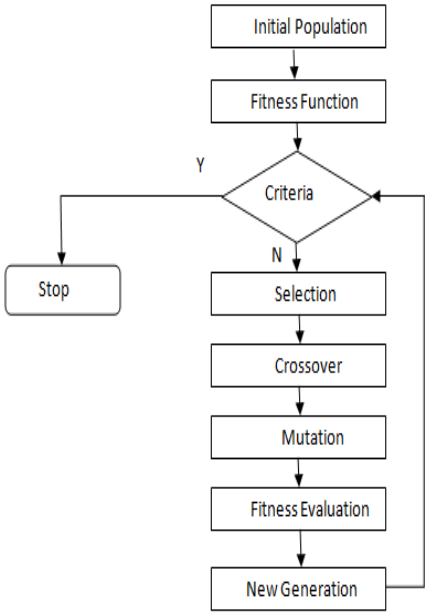
- Evaluate the fitness of each individual in that population
- Repeat on this generation until termination: (time limit, sufficient fitness achieved, etc.)
- Select the best-fit individuals for reproduction
- Breed new individuals through crossover and mutation operations to give birth to offspring
- Evaluate the individual fitness of new individuals
- Replace least-fit population with new individuals

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 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 oneparent to the other. The children take one section of thechromosome 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 occursand 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:



V. 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.

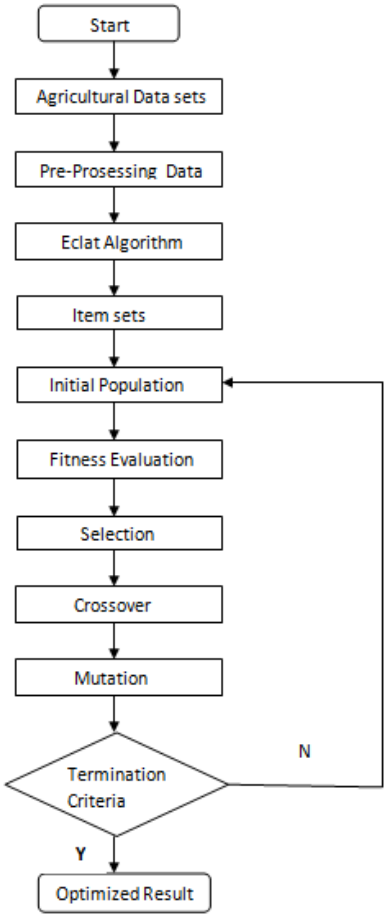


Figure 2: Architecture

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.

DOI 10.17148/IJARCCE.2015.41291

IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

VI. 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 chartsthat 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 28.31good. The scroll down window on the left of the MS Word Formatting toolbar.

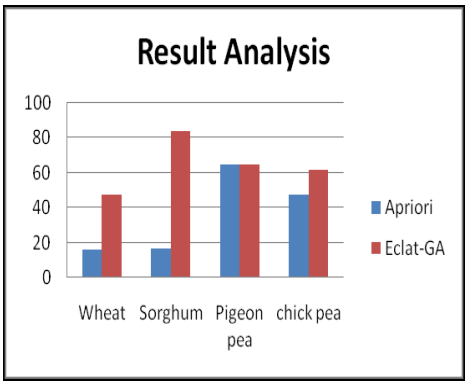


Figure3. Result Analysis

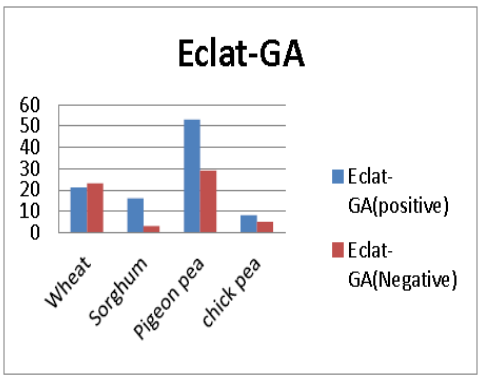


Figure4. Eclat-GA Rule Analysis

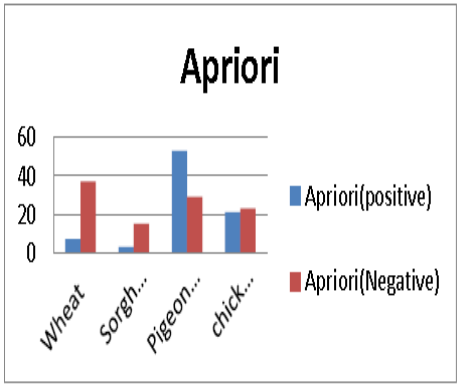


Figure5. Apriori Rule Analysis

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.

VII.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 GeneticAlgorithms by using distributed computing

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. Venansius Baryamureeba, "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.
[15] W.Chen, S.Chen, C.Lin, ―A Speech Recognition Method Based on The Sequential Multi-layer Perceptrons‖, Neural Networks, Vol-9, pp.655-699, 1996