International Journal of Advance Research in Engineering, Science & Technology (IJAREST),
ISSN (O):2393-9877, ISSN (P): 2394-2444,
Volume 2, Issue 9, September – 2015, Impact Factor: 2.125

Comparative study and analysis on Hard and Soft time Association Algorithm in Data Mining.

Monika Verma[1] , Asst. Prof. Roopal Lakhwani [2]

1Department of Computer Science and Engineering, DIMAT Raipur (C.G.), monika17verma@gmail.com
2Department of Computer Science and Engineering, DIMAT Raipur (C.G.), roopal.lkhwani@dishamail.com

Abstract

In this paper we survey about the main tasks in data mining are classification, clustering, regression and association rule mining and outlier detection. Association rule mining finds interesting associations or correlation relationships among a largest of data items. With massive amounts of data continuously being collected and stored in databases, many industries are becoming inquisitive about mining association rules from their databases. Association rule mining is the data mining task employed to solve an imperative issue in marketing parlance viz., market basket analysis. In this paper, Apriori algorithm, FP growth algorithm, Genetic Algorithm and Particle Swarm Optimization Algorithm are analyzed deeply. This research paper presents the thorough survey of algorithms for evaluating threshold values for support and confidence.

Keywords- Data Mining, Association Rule, Apriori algorithm , FP growth algorithm , Genetic Algorithm and Particle Swarm Optimization Algorithm.

I. INTRODUCTION

Data Mining infers programming uses some insight over simple grouping and partitioning of data for new information. Nowadays, the quick advancement in the size and number of databases is a great need for discovering knowledge hidden in large databases. Knowledge Discovery or Learning Prediction, information mining takes data that was once quite difficult to detect and presents it in an easily understandable format (ie: graphical or statistical) [1].Data mining Techniques involve advanced calculations, including Decision Tree Classifications, Association detection, and Clustering. Market basket examination to gather a few information about customer's market baskets approach was first introduced by Agrawal and Srikant[2].A basket indicates items purchased by a customer at a particular time. Customer's purchases can be prospected by analyzing market baskets. Data mining software uses some intelligence over simple grouping and partitioning of data to infer new information. Data Analysis is more standard statistical software (i.e.: web stats).This usually present information about subsets and relations within the recorded data set (i.e.: browser/search engine usage, average visit time, etc.). In different kinds of information database such as investigative data, medical data, financial data and marketing transaction data; analysis and finding critical hidden information has been a focused area for researchers of data mining. In this chapter, a hybrid of particle swarm optimization and the genetic algorithm developed previously by authors [3] is described and used to design the parameters of the linear state feedback controller for the system of a parallel-double-inverted pendulum. In elaboration, the utilized administrators of the hybrid algorithm include mutation, crossover of the genetic algorithm and particle swarm optimization formula. The established classical crossover and the multiple-crossover are two parts of the crossover operator. Agrawal et al. [4]. First proposed the issue of the mining association rule in 1993. They pointed out that some hidden relationships exist between purchased items in transactional databases. Therefore, mining results can help decision-makers understand customers' purchasing behavior. An association rule is in the form of X→Y, where X and Y represent Item set (I), or products, respectively and Item set includes all possible items{i1,i2, . . .,im}. The general transaction database (D= *{T1,* T2. Tk}) can represent the possibility that a customer will buy product Y after buying product X . Association rule mining introduced by Agrawal et al in 1993 [5], It is an important data mining model studied extensively by the database and data mining group. Association principle mining model among data mining several models, including association rules, clustering and characterization models is the most broadly connected technique. An arrangement of items is referred as an item set that contains k items is a k-item set. The support s of an itemset X is the percentage of transactions in the transaction database D that contain X. The support of the rule X Þ Y in the transaction database D is the support of the items set X È Y in D. The confidence of the rule X È Y in the transaction database D is the ratio of the number of transactions in D that contain X È Y to the number of transactions that contain X in D[6]. The Apriori algorithm is most illustrative calculation for association rule mining. This algorithm has a major shortcoming which can't ascertain the negligible value of support and confidence and these parameters is estimating intuitively. It comprises of numerous modified algorithms that focus on improving its efficiency and accuracy [7]. However, two parameters, minimal support and confidence, are always determined by the decision-maker him/herself or through experimentation; and in this manner, the algorithm lacks both objectiveness and efficiency. Therefore, the principle motivation behind my study is to propose an improved algorithm that can provide feasible threshold values for minimal support and confidence. So we need a method to find best values of support and confidence parameters automatically specially in large databases. In Association Rule mining find rules that will predict the occurrence of a thing taking into account the event of alternate things in the transaction.

Example of Association Rules:

{Diaper} → {Beer},
{Bread, Milk} → {Egg, Coke},
{Bread, Beer} → {Milk},

Implication means co-occurrence, not causality. Association rule is an implication expression of the form *X* →Y, where *X* and *Y* are item sets.

Example: {Milk, Diaper} → {Beer}
The two best known algorithms:
Frequent Itemset Property: Any subset of a frequent item set is frequent.
Contrapositive: If an item set is not frequent, none of its supersets are frequent.

II. Related Work:

Mining of frequent item sets is an essential phase in association mining which discovers frequent item sets in transactions database. It is the center in numerous undertakings of information mining that try to find interesting patterns from datasets, such as association rules, scenes, classifier, bunching and correlation, etc [13]. In data mining, association rule learning is a popular and well researched method for discovering intriguing relations between variables in huge databases. It analyzes and present strong rules discovered in databases utilizing diverse measures of interestingness. In these paper, Many algorithm Aprior algorithm ,FP growth , GA and PSO are analyze deeply .Agrawal et al. Proposed an algorithm[1], many researchers have been done to make frequent itemsets mining scalable and efficient. The Apriori-based algorithms find frequent itemsets based upon an iterative bottom-up approach for generating frequent itemsets. An effective Boolean algorithm for mining association rules in large sales transaction databases. The major advantage of the Boolean algorithm over the Apriori algorithm is that the Boolean algorithm generates frequent itemsets without constructing candidate itemsets. In contrast, construction of candidate itemsets is required by the Apriori algorithm. In these research paper an improved Apriori[14] is proposed through reducing the time consumed in transactions scanning for hopeful itemsets by lessening the quantity of transactions to be scanned. The time consumed to generate candidate support count in our improved Apriori is less than the time consumed in the original Apriori; our improved Apriori reduces the time consuming by 67.38%[15]. In this paper the author proposed that the Apriori-Growth algorithm based on the Apriori algorithm and FP-Growth algorithm can be combine this method only scans twice passes over dataset compresses dataset and more faster than Apriori algorithm[9]. In this paper show that FP-Growth algorithm has a higher operating efficiency and better scalability and extensibility. It can effectively analysis and deal with large data sets [16]. Particle Swarm Optimization (PSO) is a biologically inspired computational search and optimization method created in 1995 by Eberhart and Kennedy based on the social behaviors of birds flocking or fish schooling. Various fundamental varieties have been developed due to improve speed of convergence also, nature of arrangement found by the PSO. Then again, basic PSO is more appropriate to process static, simple optimization problem [12]. In this paper, we have drawn a simile of PSO with a recombinative archive-based evolutionary optimization algorithm which is algorithmically identical to a standard PSO algorithm. This is done not to demonstrate that a PSO is an EA (and possibly an EA is also a PSO), but to highlight the fact that such a similarity analysis allows us to borrow useful operations from one algorithm in enhancing the performance of another [17]. The particle swarm optimization algorithm first searches for the optimum fitness value of each particle and afterward discovers comparing backing and certainty as negligible threshold values after the data are transformed into binary values.

III. METHODOLOGY:

3.1 Apriori Algorithm

Data mining is the most instrumental device in finding information from transactions. Nowadays, improvement in technology allows shops to collect several data about customer's market baskets.A basket indicates items purchased by a customer at a particular time. Client's buys can be prospected by analyzing market baskets. The algorithm is proposed by R. Agrawal and R. Srikant in 1994 for mining frequent itemsets for Boolean association rules. The algorithm name is in view of the reality that the algorithm uses prior knowledge of frequent itemset properties. In software engineering and data mining, Apriori is a classic algorithm for learning association rules. Apriori is designed to work on database containing exchanges (for example, collections of items bought by customers, or details of a website frequentation). The algorithm endeavors to discover subsets which are common to at least a minimum number C (the cutoff, or confidence threshold) of the itemsets. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time a stage known as hopeful era, and gatherings of competitors are tested against the data. The algorithm end when no further successful extensions are found. Apriori uses breadth-first search and a hash tree structure to tally hopeful thing sets proficiently.The Apriori Algorithm is an influential algorithm for mining frequent itemsets for boolean association rules. Frequent Itemsets of item which has minimum support. Any subset of frequent itemset must be frequent. An itemset whose relating hashing container number is underneath the edge cannot be frequent.Transaction reduction that does not contain any frequent k-itemset is useless in subsequent scans. Partitioning itemset that is potentially frequent in DB must be frequent in at least one of the partitions of DB. Dynamic itemset counting an add new candidate itemsets only when all of their subsets are estimated to be frequent. Frequent itemsets are mined utilizing using Apriori algorithm or Frequent-Pattern Growth method [8]. Apriori property states that all the subsets of frequent itemsets should likewise be frequent. Apriori algorithm uses frequent itemsets, join & prune methods and Apriori property to derive strong association rules. Frequent-Pattern Growth method avoids repeated database scanning of Apriori algorithm.

3.2 FP-Growth

FP-Growth allows frequent itemset discovery without candidate itemset generation [9]. Scan DB once; find frequent 1-itemset (single item pattern).Sort frequent items in frequency descending order. Scan DB again then construct FP-tree two step approaches:
Step 1: Build a compact data structure called the FP-tree Built using 2 passes over the data-set.

Step 2: Extracts frequent itemsets directly from the FP-tree.

Completeness: Preserve complete information for frequent pattern mining never break a long pattern of any transaction.

Compactness: Reduce irrelevant info—infrequent items are gone Items in frequency descending order: the more frequently occurring, the more likely to be shared.

Divide-and-conquer: Break down both the mining assignment and DB as indicated by the frequent patterns obtained to focused search of smaller databases Other components no candidate generation, no candidate test compressed database: FP-tree structure no repeated scan of whole database basic ops-counting local database basic ops—counting local database basic ops—counting local frequent items and building sub FP-tree, no pattern search and matching.

3.3 Genetic Algorithms

Genetic algorithms are executed as a PC simulation in which a population of abstract representations (called chromosomes or the genotype or the genome) of candidate solutions (called individuals, creatures, or phenotypes) to an advancement issue advances toward better arrangements. The Genetic Algorithm [10] was developed by John Holland in 1970. The fundamental idea of GAs is intended to reproduce forms in natural system necessary for evolution, specifically those that take after the standards first set around Charles Darwin of survival of the fittest .A genetic algorithm is a pursuit procedure utilized as a part of computing to find true or approximate solutions to advancement and search problems. Genetic algorithms are classified as global search heuristics. Genetic algorithms are a particular class of evolutionary algorithms that utilization systems motivated inspired by evolutionary biology such as inheritance, mutation, selection, and crossover (also called recombination) [11]. GA is stochastic search algorithm demonstrated on the procedure of natural selection, which underlines biological evolution. Customarily, arrangements are represented in binary as strings of 0s and 1s, but other encodings are also possible. The new populace is then utilized as a part of the following emphasis of the algorithm. The algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population. If the algorithm has ended due to a greatest number of generations, a satisfactory solution may or may not have been reached. The functions of genetic operators are as follows:

Selection: Replicates the most successful solutions found in a population at a rate proportional to their relative quality.

Recombination: Decomposes two distinct solutions and then randomly mixes their parts to form novel solutions.

Mutation: Randomly perturbs a candidate solution.

3.4 PSO algorithm

Particle Swarm Optimization Proposed by James Kennedy & Russell Eberhart in 1995. Inspired by social behavior of birds and fishes Joins self-involvement with social experience Population-based optimization.PSO is a robust stochastic optimization technique based on the movement and intelligence of swarms .Uses a number of particles that constitute a swarm moving around in the quest space searching for the best solution. Each particle in search space adjusts its "flying" as indicated by its own particular flying background too as the flying experience of other particles. The Particle Swarm Optimization (PSO) algorithm is a versatile algorithm made of population of individuals (commonly referred to as particles), adapting through returning stochastically toward previous successful regions. The two primary operators in PSO are Velocity update and Position update. In particle swarm optimization, particles are flying through hyper-dimensional search space and the changes in their way are based upon the social-psychological tendency of people to imitate the achievement of different people. Here, the PSO operator adjusted the value of positions of particles which are not chosen for genetic operators [12].

The algorithm for Particle Swarm Optimization:-

Step1. Initialize the population with locations an velocities.
Step2. Evaluate the fitness of the individual particle term as“lBest”.
Step3. Keep track of the individual highest fitness termed as “gBest”.
Step4. Modify the velocities based on velocity update equation.
Step5. Update the particles position based on position update equation.
Step6. Terminate if the termination condition is met.
Step7. Go to Step 2

IV. Conclusion

In this paper related to Association Rule Mining has been carried out. Various research works have already been done on Association Rule Mining mostly using publicly dataset with different technology. Here, the survey is done on four method(Apriori algorithm, FP growth algorithm, Genetic Algorithm and Particle Swarm Optimization Algorithm) .On this survey, it is found that the PSO and GA is good to apply for the Association Rule mining purpose as it has various advantages over the other algorithm. Further ,direction include devising more robust strategies for association rule mining by combination two evolutionary algorithms and suggesting newer fitness function.

REFERENCES

[1] Suh-Ying Wur and Yungho Leu "An Effective Boolean Algorithm for Mining Association Rules in Large Databases

", Department of Information Management National Taiwan University of Science and Technology {yhl,tammy}@cs.ntust.edu.tw 1999.

[2] Mehmet KAYA and Reda ALHAJJ "Multi-Objective Genetic Algorithm Based Approach for Optimizing Fuzzy Sequential Patterns",IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2004) .

[3] Mahmoodabadi, M. J., Safaie, A. A., Bagheri, A., & Nariman-zadeh, N."A novel combination of particle swarm optimization and genetic algorithm for pareto optimal design of a five-degree of freedom vehicle vibration model" 2013 Applied Soft Computing, 13(5), 2577–2591.

[4] R. Agrawal, T. Imielin´ ski, A. Swami, "Mining association rules between sets of items in large databases", ACM SIGMOD Record 22 (2) (1993) 207–216]

[5] Manish Saggar , Abhimanyu Lad and Ashish Kumar Agrawal "Optimization of Association Rule Mining using Improved Genetic Algorithms "0-7803-8566-7/04/$20.00 Q 2004 IEEE.

[6] Soumadip Ghosh+, Sushanta Biswas*, Debasree Sarkar*, Partha Pratim Sarkar* "Mining Frequent Itemsets Using Genetic Algorithm " International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.4, October 2010 .

[7] Abdoljabbar Asadi, Azad Shojaei,,Salar Saeidi , Salah Karimi ,and Ebad Karimi "A new method for the discovery of the best threshold value for finding positive or negative association rules using Binary Particle Swarm Optimization "IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 3, November 2012 ISSN (Online): 1694-0814 .

[8] Jiao Yabing "Research of an Improved Apriori Algorithm in Data Mining Association Rules "International Journal of Computer and Communication Engineering, Vol. 2, No. 1, January 2013 .

[9] M SUMAN, T ANURADHA, K GOWTHAM, A RAMAKRISHNA/"A FREQUENT PATTERN MINING ALGORITHM BASED ON FP-TREE STRUCTURE AND APRIORI ALGORITHM "International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 1, Jan-Feb 2012, pp.114-116 .

[10] Pei M., Goodman E.D., Punch F "Feature Extraction using genetic algorithm" Case Center for Computer-Aided Engineering and Manufacturing W. Department of Computer Science(2000).

[11] Ali Hadian Mahdi Nasiri, Behrouz Minaei-Bidgoli "Clustering Based Multi-Objective Rule Mining using Genetic Algorithm " International Journal of Digital Content Technology and its Applications Volume 4, Number 1, February 2010 .

[12] Russell Eberhart and James Kennedy "A New Optimizer Particle Swarm Theory" Sixth International Symposium on Micro Machine and Human Science 0-7803-2676-8/95$4.0001995 IEEE.

[13] S. Rao, R. Gupta, "Implementing Improved Algorithm Over APRIORI Data Mining Association Rule Algorithm", International Journal of Computer Science And Technology, pp. 489-493, Mar. 2012.

[14] Shilpi Singla1, Arun Malik2 "Survey on various improved Apriori Algorithms " International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 11, November 2014 .

[15] Huan Wu, Zhigang Lu, "An Improved Apriori-based Algorithm for Association Rules Mining " 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery.

[16] Lijuan Zhou and Xiang Wang "Research of the FP-Growth Algorithm Based on Cloud Environments" 2014 ACADEMY PUBLISHER doi:10.4304/jsw.9.3.676-683

[16] Kalyanmoy Deb and Nikhil Padhye "Improving a Particle Swarm Optimization Algorithm Using an Evolutionary Algorithm Framework" Kanpur Genetic Algorithms Laboratory Department of Mechanical Engineering Indian Institute of Technology Kanpur PIN 208 016, India deb@iitk.ac.in, npdhye@gmail.com KanGAL Report Number 2010003 February 21, 2010 .

[17] Shilpi Singla1, Arun Malik2 "Survey on various improved Apriori Algorithms " International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 11, November 2014 .