Performance Evaluation of Frequent Pattern Mining using Genetic Algorithm: The Review

Mamta Sharma[1]**, Sunil Kumar**[2]

[1]Student, Computer Science & Engineering, Guru Jambheshawar University of Science & Technology, India

[2]Assistant Professor, Computer Science & Engineering, Guru Jambheshawar University of Science & Technology, India

***Abstract*—Frequent pattern mining is one of the essential tasks in the data mining. it is the core process used by other data mining processes for decision making like data indexing ,organization, clustering, etc. it show an central role in association rule mining. Classical approaches of Frequent Pattern Mining that's takes pretty time and there is more than one objective in frequent pattern mining. So, we make use of genetic algorithm to prevail over these issues.Genetic algorithm works on discovery the global best promising solutions. Then several frequent patterns are mined. Based on these frequent patterns we try to pit association rules. Along with we evaluate effectiveness of these rules with previously generated association rules.**

***Keywords*—Data Mining, Frequent pattern mining, Rules, genetic algorithm, evaluation.**

I. Introduction

Frequent pattern mining[1] is the process of finding of the items which occurs frequently in a database. Associations allow capturing possible rules that explain the presence of some attributes according to the presence of other attributes in the same transaction.

Let I = {i1,i2,i3……..in} be a set of every items, A k-item set α which consisting of k items from set I is frequent if α occurs in a transaction database D no lower than θ times, where θ is user specified threshold.

Association rule mining becomes a vast locale of research in last some decades. The fundamental idea behind association rule mining is to mine positive (interesting) and negative(uninteresting) rules from a transaction database. Association rule mining is the process of finding the relation between the frequently occurring item-set which are used for decision making. Association rule is of the form A->C where A and C are set of items and A∩C=Ø. This means that if all the items in A exists in transaction then all the items in C with a high probability are also in the transaction and A, C should not have any common items. A is known as antecedent and C is known as consequent. The antecedent and consequent are constructed from attribute tests (AT). Rule antecedent are conjuction of AT and are different from the attribute in consequent. The attribute takes either numerical or categorical values. The strength of such a rule is usually evaluated by means of statistical a procedure, that is for instance the confidence and support, which are defined as follows:

Support(A->C)=support(A∩C) also known as minimum support or minsup

Confidence(A->C)=support(A∩C)/Support(A) also known as minimum confidence or minconf

The rule is said to be strong if its support and confidence is greater than the minsup and minconf. Discovering strong rule is a two-step process :

i. First step is to find all the frequent itemsets w.r.t minsup.

ii. Second step consist of discovering the association rule based on minconf.

First step in association rule mining is more complex than the second step. There are many methodologies which are used for frequent pattern mining like Apriori, FP growth, vertical(Eclat) etc.Many previous studies for mining association rules focused on database with binary or discrete values, however data values in real-life applications usually consists of quantitative values. Methods like apriori, fp growth and other classical approaches are good for these discrete values but not much efficient for these quantitative values. So we use partitioning the domain introducing the new attributes with intervals. Currently apriori algorithm plays a major role in identifying frequent item-set and driving rule-set out of it. The majority of existing techniques to mine association rules typically use the support and the confidence to evaluate the quality of rule obtained. However these two measures may not be sufficient to properly access their quality due to some inherent drawbacks they present. We can also use genetic algorithm with the other frequent pattern mining methodologies like apriori, fp expansion, Eclat to prevail over the disadvantages of these methodologies. Apriori algorithm is quite time consuming since we have to scan whole database each time the contestant is generated and the no. of candidate are very large so it takes bunch of time for whole process. So we use genetic algorithm to overcome this disadvantage. Genetic algorithms use chromosomes for the representation of rules. These chromosomes are represented with the help of bits 0,1. Each bit in chromosome is known as gene and the collection of chromosomes is known as population. There are a specific no of population in a run. Some of the operators which are used in the genetic algorithm are selection, crossover, and mutation. Selection is used to select the chromosomes which are fit for reproduction while crossover exchange subparts of two chromosomes it is used to produce a better offspring from the two good parents. Crossover may be single point or multipoint. Mutation randomly flip the bit value of some locations into the chromosomes. Genetic algorithm frequently requires a “fitness function” that assigns fitness to each chromosome in the current population. The fitness of a chromosome depends of the how well that chromosome solve the problem. The more the chromosome is fit the more the chances it is selected for next population.

The procedure of integrating two or more than two techniques to improve the overall performance is known as hybridization. The output of a GA as a rule induction system is a simple “if….then” rule for each individual if the Michigan style approach is used each rule representing a class, in the case of a Pittsburg approach the output is a complex “if…then….else if” rule which encode the entire system of knowledge base. Most of these algorithms for mining quantitative association rules focus on positive dependencies without paying particular attention to negative dependencies.

II. LITERATURE REVIEW:

In the paper given by Rakesh Agrawal et.al [2] introduces the concept of association rule mining in large database. He used the apriori algorithm for mining the frequent patterns. This paper studied the market behavior of the customer and tries to find frequent items which are purchased by the customer in a single trip of market. This paper tells that given a set of transaction T, we are interested in generating all rules that satisfy certain additional constraint syntactic and support.

J. Han et.al. [3] tells about the current issues and future aspects about the frequent pattern mining and gives a brief description about different methodologies used in frequent pattern mining. This paper describes different types of frequent patterns which can be mined from the large databases i.e. sequential patterns, structured patterns, frequent pattern based clustering. It provide brief introduction in relation to the methodologies used for frequent pattern mining like apriori and its associated extension, fp growth, vertical method, etc. it also give overview on mining multidimensional, multilevel and quantitative association rules.

Deepak Garget. al. [5].Various techniques have been proposed to improve the performance of frequent pattern mining algorithms. This paper presents a review of different frequent mining techniques including apriori based algorithms, partition based algorithms, SQL based algorithms and incremental apriori based algorithms. It tells that performance of particular technique depends on input data and available resources. It uses five algorithms for comparison purpose. These are apriori based algorithm, division based algorithm, incremental based algorithm, FP hierarchy algorithm, SQL based algorithm. It uses the kosarak dataset and mushroom dataset.

Frequent pattern mining for understanding genetic algorithm is given by Minh Nghia Le et. al.[8]. This paper present frequent schema analysis (FSA) approach as an instance of informatics for extracting knowledge on search dynamics of binary genetic algorithm using the optimization data generated throughout the search. SAFP- a new self adaptive algorithm for frequent pattern mining by Xin-Yin Wang et. al.[11].construct a robust algorithm by meticulously combining two different mining algorithm on FP-tree while adjusting the mining strategy dynamically and automatically during a complete process of frequent pattern mining.

Multi-objective evolutionary algorithm is explained by EckartZitzler et.al.[12]. He gives some instances in evolutionary algorithms and also done a comparative case study and the strength of pareto approach. It takes 0/1 knapsack as a basis for problem description. This study compared four multi-objective 0/1 knapsack with nine different problem settings.

Genetic algorithm and different application areas covered by them is defined by Mir AsifIquebal [13]. It uses matlab tool to compare different problem areas of genetic algorithms.

How effective genetic algorithms are in extracting the good association rules is studied by Jesus Alcala Fdezet. al.[15].The aim of this paper is to show the effectiveness of genetic algorithm for mining quantitative association rules. EARMGA, GAR, GENAR.

M. Martinez-Ballesteros [24] tells that the two measures for ruling interestingness i.e. support and assurance may not be adequate to properly access their quality due to some inherent drawbacks they present.

III. PROBLEM FORMULATION:

After studying different papers related to frequent pattern mining we come across several types of frequent patterns and different types of algorithms and techniques to mine the frequent patterns. We also studied about the different advantages and disadvantages related to each algorithm. Every algorithm need minimum support and minimum confidence as a measure of rule interestingness and how much the rule is strong. There is always a doubt in choosing the right confidence and right confidence value so that only relevant policy can be mined. Apriori algorithm be one of the first and basic algorithms adopted to mine the repeated item-sets. although it is pretty time consuming to practically adopt the apriori algorithm because the number of candidates generated is very large and each time to generate the candidates the whole database need to be scanned. thus it is essential to bound these disadvantages of apriori algorithm before adopting it practically. We study that the classical approaches for frequent pattern mining used to mine the discrete or categorical data. But as we know that real world problem not only contain the categorical data but also contain continuous and numeric data. So these classical approaches seem to be quite inefficient for mining these numerical dataset. The mining of these continuous item-sets is known as quantitative association rule mining. So we usesome techniques to mine these quantitative association rules. From previous papers we also studied that the frequent pattern mining does not be considered as a single objective problem but as a multi-objective problem there are many objectives which have to be achieved. Some of these objectives are unambiguousness, interestingness, correctness, sturdy etc. so efforts must be made to achieve a good tradeoff between all these objectives. From studying about the evolutionary algorithms we learnt thatthese evolutionary algorithms plays an important role in achieving all the objectives stated in the frequent pattern mining. It helps in achieving a good tradeoff between different objectives. Genetic algorithms are basically used for quantitative association rule mining. Genetic algorithms works in the direction of searching global optimal solution instead of finding local optimal solution. These algorithms can be used with the classical techniques of frequent pattern mining to mine the Quantitative association rules. Basically apriori algorithm is used as a reference and genetic algorithm is used as the main algorithm to mine QAR. We studied in a paper that QUANTMINER is used as an algorithm which use genetic algorithm in mining frequent patterns. It use rule templates. It follows a prototypicalgenetic algorithm approach. Then QAR-CIP-NSGA-II is also studied. It tries to achieve a tradeoff between accuracy and performance. The main problem in using genetic algorithm in mining the quantitative association rule is that how to decide different objective measures for which we need tradeoff. For deciding the criteria for when to stop the running. So we try to find some better objective measures which help us in mining better quantitative association rules. With the assist of the genetic algorithms we strive to mine some good quantitative association rules

IV. Methodologies:

In this work first we study very carefully the previous papers based on the mining quantitative association rules using genetic algorithm. Then we start our work by loading the sample of records from the transactional database that fits in the memory. Then an initial population is created consisting of randomly generated transaction. Each transaction can be represented by a string of bits. We does our coding with the help of 0, 1. The 0 part indicates that attribute is in antecedent part, the 1 indicate that the attribute is in consequent part. -1 indicates that the attribute is neither form antecedent nor forms a part of consequent. Then we use LB(lower bound) i.e. forms the lower bound of the interval and UB(upper bound) i.e. forms the upper bound of the interval to make the intervals. A term known as amplitude is used to control the interval from spanning overall density. By using genetic algorithm we try to find the globally optimal solution i.e. by finding the globally optimal association rules.

Gene= (aci,lbi, ubi) where i=1,2,3,…….n.

CT= Gene1, Gene2,……,Genen. where CT is a chromosome coded in this way.

Our proposed genetic algorithm works as follows:

a) Fitness evaluation: first the fitness of each individual is calculated
b) Selection: individuals are chosen from the current population as parents to be involved in the recombination
c) Recombination: new individuals are produced from the parents by applying genetic operator such as crossover or mutation.
d) Replacement: some of the offspring are replaced with some individuals (usually with their parents).

One cycle of transforming a population is called a generation. In each generation, a fraction of population is replaced with offspring and its proportion to entire population is called generation gap. Then we iterate our search for finding association rules when we meet some good association rules or when some given criteria is achieved. We specify some confidence or support measures and determine how interesting rule is the. Then some frequent patterns are mined. Based on these frequent patterns we try to mine association rules. And then we compare efficiency of these rules with previously generated association rules.

Facilities required for proposed work:

The proposed work is planned to be done in one of the software which is best available at the time of practical implementation. The software may be MATLAB/KEEL/WEKA etc.

V. References:

[1]. J. Han, M. Kamber, Data Mining: Concepts and Techniques, 2nd edition Burlington, MA, USA: Morgan Kaufmann, 2006

[2]. R. Agrawal, T. Imielinski and A.Swami “ Mining Association Rules Between Sets of Items in Large Databases,” in Proc. SIGMOD, 1993, pp. 207-216

[3]. J. Han, Hong Cheng, Dong Xin, Xifeng Yan “ Frequent Pattern Mining: Current Status and Future Directions” in Data Mining and Knowledge Discovery(2007) pp.55-86

[4]. Deepak Garg, Hemant Sharma “ comparative Analysis Of Various Approaches Used in Frequent Pattern Mining” in International Journal Of Advance

Computer Science And Applications, Special Issue On Artificial Intelligence.

[5]. D. Usha, Dr. K.Rameshkumar " A Complete Survey On Application Of Frequent Pattern Mining And Association Rule Mining On Crime Pattern Mining" in International Journal Of Advances In Computer Science and Technology, Volume 3, No. 4, April 2014

[6]. Melanie Mitchell "Genetic Algorithm: An Overview" Adapted from An Introduction to Genetic Algorithms, Chapter 1, MIT Press 1(1) 31-39, 1995

[7]. Ashish Ghosh, BhabeshNath " Multi-objective Rule Mining Using Genetic Algorithms" in Information Sciences 163(2004) pp.123-133

[8]. Xin-Yin Wang, Xin Zhang, Hai- Bing Ma, Yun- Fa Hu "SAFP: A New Self Adaptive Algorithm For Frequent Pattern Mining" in Proceedings of Fifth International Conference on Machine Learning and Cybernetics, Dalian, 13-16 August 2006

[9]. EckartZitzler, Lothar Thiele "Multiobjective Evolutionary Algorithms: a Comparative Case Study and Strength Pareto Approach"

[10]. Mir AsifIquebal "Genetic Algorithms and Their Applications: An Overview"

[11]. A.A. frietas, H.S. Lopes, M.v. Fidelis "Discovering Comprehensible Classification Rules With a Genetic Algorithm"

[12]. Jesus Alcala-Fdez, NicoloFlugy – Pape, Andrea Bonarini, Francisco Herrera " Analysis of the Effectiveness of the Genetic Algorithms Based on the Extraction of Association Rules" in FundamentaInformaticae 98(2010) pp 1-14, IOS Press

[13]. Sameer Kumar Vishnoi, VivekBadhe " Association Rule Mining for Profit Patterns Using Genetic Algorithms" in International Journal of Emerging Technology and Advanced Engineering, Volume 4, Issue 5, May 2014

[14]. Bilal Alatas, Erhan Akin "An Efficient Genetic Algorithm for Automated Mining of Both Positive And Negative Quantitative Association Rules" in Soft Computing 10: pp 230-237 (2006)

[15]. AnsafsallebAouissi, ChristelVrain, Cyril Nortet "Quantminer: A Genetic Algorithm For Mining Quantitative Association Rules" in IJCAI, pp 1035-1040, 2007

[16]. D. Martin, A. Rosete, J.AlcalaFdez, F.Herrera"QAR-CIP-NSGA-II: A New Multi-ObjectiveEvolutionary Algorithm to Mine Quantitative Association Rules" in Information Sciences, 2013

[17]. KannikaNiraiVaani M, E Ramaraj "An Integrated Approach to Derive Effective Rules From Association Rules From Association Rules Mining Using Genetic Algorithm" in Proceeding Of the 2013 International Conference On Pattern Recognition, Informatics and Mobile Engineering(PRIME) Feb 21-22, 2013

[18]. D.Martin, AlejandroRosete, Jesus Alcala Fdez " A New Multiobjective Evolutionary Algorithm for Mining a Reduced Set Of Interesting Posititve and Negative Quantitative Association Rules in IEEE Transaction on Evolutionary Computation, Volume. 18, No. 1, February 2014

[19]. B. Minaei-Bidgoli, R. Barmaki, M. Nasiri "Mining Association Rules Via Multi-objective Genetic Algorithms" in Information Sciences 233(2013) pp 15-18.