International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-2 Issue-4, September 2013

Association Rule Mining: A Multi-objective Genetic Algorithm Approach Using Pittsburgh Technique

Sonia Sharma, Vinay Chopra

Abstract ***Association rules [4] usually found out the relationship between different data entities in given data set and moreover it is very much important task of data mining. Basically, association rule mining is a multi-objective problem, instead of a single objective problem. A multi-objective genetic algorithm approach using Pittsburgh technique is introduced in this paper for discovering the interesting association rules with multiple criteria i.e. support, confidence and simplicity and complexity With Genetic Algorithm. In this paper we have discussed the results on various datasets and show effectiveness of the new proposed algorithm.***

Index Terms- Data Mining, Genetic Algorithm, Optimization, Association Rule, Measure, Apriori, Genetic Operators, Interestingness, Frequent Item-set

I. INTRODUCTION

Association Rule mining is important task of data mining that finds the probability of co-occurrence of items in a collection. The major goal is to extract interdependence associations' structures among the item sets in the transaction databases or other data repositories. The formally the association rule mining problem was firstly stated in by Agrawal. Let I is item-set of m distinct attributes, $I=\{I_1, I_2, \ldots, I_m\}$ and D is database (transaction set), $D=\{T_1,T_2,\ldots T_N\}$, where $T \subseteq I$ and there are two item-sets X and Y, such that $X \subseteq T$ and $Y \subseteq T$, then association rule, X=>Y holds where $X \subset I$ and $Y \subset I$ and $X \cap Y = \emptyset$. X is called antecedent while Y is called consequent; the rule means X tends to Y. The two basic measures for association rules are namely support (sup) and confidence (conf). These two thresholds are called minimal support and minimal confidence respectively. Thus the two basic parameters for the Association Rule Mining (ARM) are: support (sup) and confidence (conf).

II. MULTI-OBJECTIVE OPTIMIZATION

Multi-objective optimization is also known as vector, multiple-criteria, multi-attribute optimization or Pareto optimization [2]. It is an area of multiple decision making based on multiple criteria that is concerned with mathematical optimization problems having one or more objective function to be optimized concurrently.

Revised Manuscript Received on 30 September 2013.
* Correspondence Author
PG Scholar Ms. Sonia Sharma, M.tech Computer Science Punjab Technical University DAVIET Jalandhar, India.
Mr. Vinay Chopra, M.Tech Computer Science Punjab Technical University DAVIET Jalandhar, India.

Multi-objective optimization has been applied on many fields like, engineering, economics, science and logistics where optimal decisions are taken in the presence of trade-offs between more than one conflicting objectives. While maximizing the strength of a particular component Minimizing the weight of other and maximizing the performance while various audits being taken . Optimization problems involving two and three objectives based on the concept on which the multi-objective optimization is being applied.

A. Approaches of MOG

Three Approaches to MOG are:
1) Preemptive Optimization:
2) Composite Objective
3) Purely Multi-Objective:

1) Preemptive Optimization

(i) To arrange the objectives based on their Priority or problem-specification.
(ii) To improve the highest-priority objective
(iii) To present new constraint based on optimum Value obtained

2) Composite Objective
(i) To allocate weights to each function according to some criteria
(ii) To maximize and minimize objectives receive opposite signs
(iii) To sum up the weighted functions to make new composite function
(iv) To solve as a regular single-objective optimization problem

3) Pure MOPs
It is of two types: population based and pareto optimality
(i) Population-Based Solutions.
a) Allow for the inquiry of trade-offs between striving objectives.
b) Genetic algorithms are used for solving Multi-objective optimizations in their pure and natural form.
c) Such techniques depends upon the the concept of Pareto optimality
(ii)Pareto Optimality
a) MOP → Exchange between competing objectives
b) Pareto approach → exploring the exchange surface, yielding a set of possible solutions also known as Edge worth-Pareto optimality

B. Multi-objective Genetic Framework

In this subsection, working of multi-objective genetic algorithm for the association rule mining based on apriori algorithm in data mining flow-chart depicts the main phases of Multi-objective

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

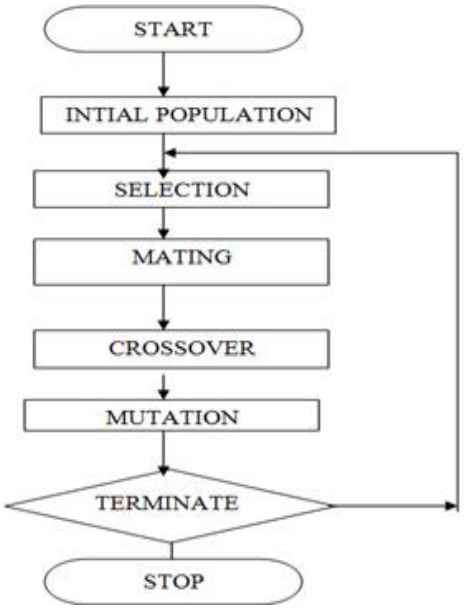

Figure 1: MOG: Framework

C. Working of Multi-objective Genetic Algorithm

Multi-objective genetic algorithms (GA) mimic the biological processes underlying classic Darwinian evolution in order to find solutions to optimization or classification problems. Its implementations utilize a population of candidate solutions (or chromosomes). Each chromosome in the current generation is evaluated using a fitness function and ranked. From the ranking candidates are selected from which the next generation is created. The process repeats until either the number of iterations is exceeded or an acceptable solution is found.

Details of implementation are varied but a generic view of a GA [1] would include:

1. Initialization of the initial population (either randomly or from a best guess or previous partial solution)
2. Evaluation of the fitness functions on each chromosome to determine ranking.
3. Application of the selection method on the population to determine mating rights.
4. Application of the multi-objective genetic operators on the chromosomes selected for mating.
5. Return to Step #2. .

III. PROPOSED APPROACH

In this particular section a multi-objective genetic algorithm model is presented for the finding the interesting association rules from large datasets. Here we discuss the various Operators of Genetic Algorithm i.e Selection, encoding the genetic operators, and the fitness function which is used in this paper for finding out the various results.

Selection In this the Chromosomes are selected from the given population to be parents for crossover process. The problem is how to select these chromosomes from the given population. The best Chromosomes survive and create new offspring and those which are not the best they dies this is according to the Darwin's evolution theory There are many methods how to select the best chromosomes from the given population for example Boltzmann selection, tournament selection, roulette wheel selection, rank selection, steady state selection etc.

Encoding [2] there are two techniques based on how we can encode the rules into the population of individuals namely Michigan technique and Pittsburgh technique. In the Michigan technique each and every rule is encoded into an individual, but in the Pittsburgh technique set of rules is encoded into a chromosome. In this paper Pittsburgh technique is adopted i.e the set of rules are encoded into the individual chromosome.

Genetic Operators [5] various genetic operators are Selection as discussed earlier, Crossover and then Mutation

Crossover [4] is a genetic operator used to vary the programming of a chromosome or chromosomes from one generation to immediate next generation. Crossover is analogous to reproduction and biological crossover, on which genetic algorithms are dependent. Cross over is a process of acquiring more than one parent solutions and generating a child from that solution.

Mutation [5] is a genetic operator used to maintain genetic variety from one generation of a population of genetic algorithm chromosomes to the immediate next. Mutation changes one or more gene values in a chromosome from its starting state. In mutation, the solution may change absolutely from the previous solution. Hence GA can give better results with mutation. Mutation happens during evolution according to a user-definable mutation probability. This probability should be set minimum. If it is set too high, the search will turn into a earliest random search.

Fitness Function The fitness function is a measure of how well a candidate solves the problem. Implementations vary in the choice and practice of the selection method; suffice to say that the purpose of the selection method is to choose candidates whose multi-objective genetic mix will tend to lead to improve candidate solutions in the next generation. Examples of common selection methods include random, elitist, roulette wheel, tournament, etc. Multi-objective genetic operators provide mixing of chromosome portions from the parent or parents to form the offspring of the next generation. Examples of multi-objective genetic operators include crossover, mutation, inversion, etc. In this paper we are using the fitness function with the combination of Support, Confidence, Simplicity and Complexity.

Support for an association rule is defined as the percentage or fraction of transactions in N that contain X ∪ Y. Support is calculated by the following formula:

$$\sup(X \cup Y) = \frac{\text{count}(X \cup Y)}{\text{count}(N)}$$

Where XUY is the number of transactions containing both in X and Y and N is the total number of transactions in the database.

Confidence for an association rule is defined as the percentage or fraction of the number of transactions in N that contain X also contains Y. Confidence is calculated by the formula:

$$\text{conf}(X => Y) = \frac{\sup(X \cup Y)}{\sup(X)}$$

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

The simplicity of the rule can be defined by the total number of attributes on the antecedent part of the rule i.e on the left hand side of the rule and tries to calculate the understandability of the rule. Simplicity can be calculated by the formula:

Simp=((Math.log(1 + Y))/(Math.log(1 + XUY)))

The Complexity of the rule is also introduced to more simplifying the rule or to make the complex rule simpler and more understandable if on the consequent part the number of attributes is more. The complexity of the rule can be calculated by the formula:

comp = 1 / ((Math.log(1 + Z)) + (Math.log(1 + Y)) / (Math.log(1 + XUY)));

So from all these factors we can calculate the fitness function as:

$$F(x)=\frac{(w1*sup + w2*conf+w3*simp+w4*comp)}{(w1+w2+w3+w4)} \qquad (1)$$

Here w1, w2, w3, w4 are the weights that are defined by the user according to priority.

IV. RESULTS AND DISCUSSIONS

When this newly introduced approach i.e. Pittsburgh approach is compared with the Michigan approach then we got the following results which came from the various authorized datasets when applied. Various results are discussed in the tables below on the bases of no of rules, fitness function, and time.

Table 1: Result comparison on bases of number of rule

Name of Dataset	Results with Michigan approach	Results with Pittsburgh approach
Breast Cancer	6	10
Contact lenses	3	10
Weather	3	10
Vote	9	10
Zoo	10	10

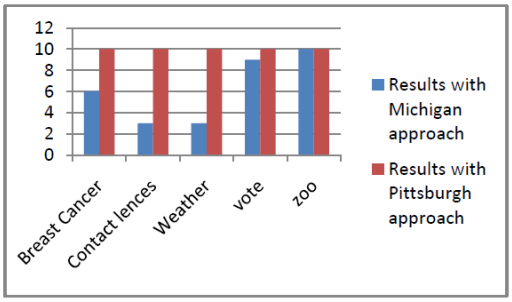

Figure2: comparision on bases of number of rules

Table 2: Result comparison on bases of fitness function

Name of Dataset	Results with Michigan approach	Results with Pittsburgh approach
Breast Cancer	0.8247	0.8442
Contact lences	0.7882	0.8762
Weather	0.7782	0.8442
Vote	0.82935	0.9180
Zoo	0.8286	0.9299

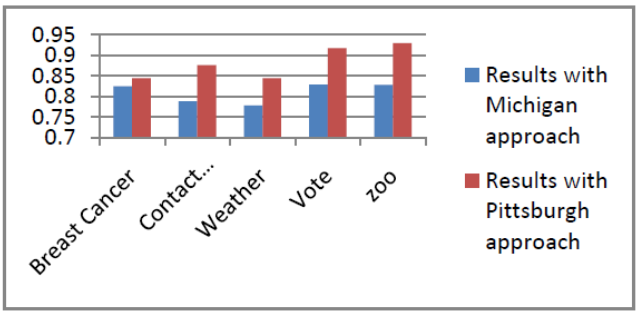

Figure3: comparision on bases of fitness function

Table 3: Result comparison on bases of time taken in milliseconds

Name of Dataset	Results with Michigan approach	Results with Pittsburgh approach
Breast Cancer	2163	61651
Contact lenses	6630	11934
Weather	2418	6411
Vote	38313	82571
Zoo	21044	42011

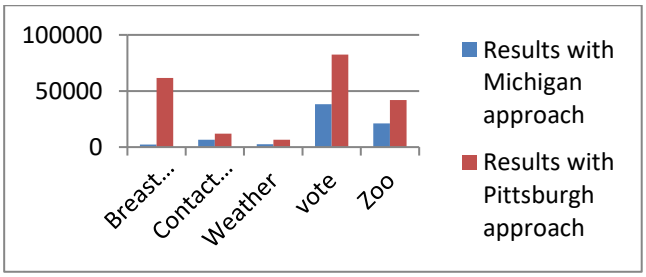

Figure4: comparison on bases time taken in milliseconds

V. CONCLUSION

On seeing the various results on various datasets we came into the conclusion that in the Pittsburgh approach the value of fitness function is much better than the value calculated from the Michigan approach, as the fitness function lies between 0-1 and more accurate the result is as much as the value is near to the 1. Also the no of generations increases in this approach which also leads to find out the better rules, but the time taken in the Pittsburgh approach more than that of Michigan because in the Pittsburgh we take a set of rule as one chromosome and that set may contain any no. of riles instead of a single rule.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

REFERENCES

1. Indra k and kanmanis " Performance analysis of genetic algorithm for mining association rules" *International Journal Of Computer Science*, issues Vol. 9, issue 2, March 2012, pp: 318-376.
2. Rajul Anand, Abhishek Vaid, Pramod Kumar Singh "Association Rule Mining Using Multi-Objective evolutionary algorithm: Strengths and challenges" *IEEE Conference*, 2009, pp:385-389.
3. Rupali Haldulakar and Prof. Jitender Aggarwal "optimization and association rule mining through genetic algorithm*" International Journal Of Computer Sciences And Engineering* Vol. 3, No. 3, March 2011, pp: 1252-1259.
4. Jian Hu and Xing Yang Li "Association rule mining using multi-objective co evolutionary algorithm" *Ieee International Conference On Computational Intelligence And Security Workshop*, 2007, pp: 405-408.
5. Basheer Mohamad "Discovering interesting association rules a multiobjective genetic algorithm" *International Journal Of Applied Information System*, Vol. 5 No. 3, February 2013, pp: 47-52.
6. Sanat Jain, Swati Kabra "Mining and optimization of association rules using effective algorithm" *International Journal Of Emerging Technology And Advanced Engineering*, Vol. 2 issue 4 April 2012.
7. J.Malar Vizhi and Dr. T.Bhuvaneswari " Data quality measurement on categorical data using genetic algorithm" *International Journal And Determining And Knowledge Management Process*, Vol. 2, 01 Jan 2012, pp: 33-42,.

AUTHOR PROFILE

Ms. Sonia Sharma PG Scholar, Pursuing M.Tech in Computer Science under the DAV Institute of Engineering and Technology, Jalandhar, India. She has completed her under graduate course B.Tech in computer science under the Department Of Computer Science from Lovely Institute of Technology. Her areas of interest are Data mining and optimization techniques and algorithms

Mr.Vinay Chopra has 5 years experience as the Assistant. Professor in the Department of Computer Science of DAV Institute of Engineering and Technology, Jalandhar, India. He has done his masters in Software Engineering and pursuing his P.hd. He has life membership of Punjab Academy of sciences.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)