Analysis of Various Multiobjective Genetic Approaches in Association Rule Mining

Sonia Sharma
Assistant Professor
Department of Computer Science & Engg.
CT Institute of Engg. Mgt. and Tech.

Vinay Chopra
Assistant Professor
Department of Computer Science & Engg.
DAV Institute of Engg. and Tech.

ABSTRACT
Data mining is used now days by companies with a strong consumer focus. It enables these companies to know the relationships among "internal" factors such as, product positioning, price or staff skills, and "external" factors such as indicators, economic, competition, and customer demographics. The overall aim of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. In this paper, the multi-objective genetic approach for the result Comparison of Pittsburgh and Michigan approach using multi-objective genetic algorithm has been proposed, and it is shown that using Pittsburgh approach is much better than the Michigan approach.

Keywords
Michigan, Pittsburgh, Multi-objective, Genetic Algorithm.

1. 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 K is item-set of m distinct attributes, K={K1, K2, …., Km} and D is database (transaction set), D={T1,T2,….TN}, where T ⊆ I and there are two item-sets X and Y, such that X ⊆ T and Y ⊆ T, then association rule, X=>Y holds where X ⊂ I and Y ⊂ I and X ∩ Y = ø. 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).

2. 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. 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.

2.1 Approaches of MOG [8]
The three Approaches to MOG are:

I. Composite Objective
II. Preemptive Optimization
III. Purely Multi-Objective

I. Composite Objective

i. To assign weights to every function according to some criteria

ii. To max or min objectives receive alternate sign.

iii. To add up the weighted functions to get new composite function

II. Preemptive Optimization

i. To settle the objectives based on their Priority

ii. To improve the first-priority objective

iii. To show new constraint based upon optimum value obtained

III. Pure MOPs

It has of two categories: population based MOG and pareto optimality MOG

i. Population-Based Solutions.

a) Allow for the inquiry of trade-offs in b/w striving objectives.

b) GA are used for solving MO optimizations in their natural and pure form.

ii. Pareto Optimality

a) Multi-Objective Optimization → Exchange between competing objectives

b) Pareto approach → exploring the exchange surface, yielding a set of possible solutions also known as Edge worth-Pareto optimality

2.2 Functioning 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. A multi-objective genetic algorithm model is presented for the finding the interesting association rules from large

datasets. Here it is discussed that the various Operators of Genetic Algorithm i.e Selection, encoding the genetic operators, and the fitness function used in this paper for finding the different result observations.[8]

i. *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 different methods to select the best chromosomes from the given population.

ii. *Encoding [2]* there are two techniques based on how one 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]*

iii. *Crossover [4]* is a genetic operator used to vary the programming of a chromosome or chromosomes from one generation to exact next generation. Crossover is analogous to reproduction and biological crossover, on which genetic algorithms depends. Cross over can be defined aa a process of choosing more than one parent solutions and generating a child from that particular solution.

iv. *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 Genetic algorithms give us the best results with mutation. Mutation happens during evolution according to a user-definable mutation probability. The probability for this should be set as minimum as possible and if it will high, then the search will come into a earliest random search.

3. MOG ENCODING APPROACHES

The two techniques based on how one can encode the rules into the population of individuals are 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.

4. RESULTS & DISCUSSIONS

When both the approach i.e. Pittsburgh approach is and the Michigan approach are compared on various data sets then the following results came from the various authorized datasets when applied. Various results are discussed in the tables below:

When Pittsburgh approach is compared with the Michigan approach then the following results came from the various authorized datasets when applied. Various results are discussed in this paper on the bases of no of rules, number of generations, fitness function, and time. The paper will compare the results of proposed approach with the previous approach on the bases of following two points:

1) Firstly to compare the results on the bases of occurrence of number of rules and on the bases of number of generations on some data set.

2) Secondly to compare the results on the bases of fitness function and time taken by the by the rules of a particular data sets.

These points are briefly discussed below in the form of various tables and graphs and moreover the decisions are taken on testing the results on the various authorized data sets. In this paper the decisions about the results have been taken by testing the results on the five data sets namely Breast Cancer, Contact lenses, Weather, Vote, Zoo. The result values of these data sets are taken and analyzed and discussed, on the bases of which various decisions are taken.

The first comparison of the result analysis is discussed in the table. The comparison of the proposed approach i.e Pittsburgh approach with the previous approach i.e Michigan approach on the 5 data sets and we see that the number of rules of the newly proposed approach increases as compared to the previous approach which gives the more accurate rules and moreover also the number of generations increases and due to this as generation increases with that rules also increases and hence the most valuable data has been taken from the data sets and due to this the pruning of the data also reduces and hence the more refined form of the data from the given data set. In the table below it is very much clearly shown that the there is a great increment of in the number of rule set and in the total number of generation and hence due to this pruning of dataset also reduce which leads to major improvement in the results of the newly proposed approach.

Table 1: Comparison on bases of number of generations and number of rules

Data sets	Results with Michigan approach		Results with Pittsburgh approach	
	Number of Rules	Number of genera-tions	Number of rules	Number of genera-tions
Breast Cancer	170	6	246	10
Contact lenses	145	3	204	10
Weather	118	3	194	10
Vote	156	10	250	10
Zoo	176	10	248	10

The figure1 below discusses the result comparison of proposed and previous approach clearly in the form of a graph, in order to compute the results in the form of graph provides the values on both the axis of the graph. On the Y-axis of the graph the number of rules and number of generations are taken and on the X-axis various datasets are taken. From the graph it is very much clear that the number of rule and the number of generations of the Pittsburgh approach are more as compared to the Michigan approach, so it can be said that the Pittsburgh approach is much better than the Michigan one because as the number of rules and the number of generations are more than the lesser is the value of the

pruning. In the data mining task as much as the pruning of data set is less more valuable results are achieved. Hence it can be said that the Pittsburgh approach is much better than the Michigan one.

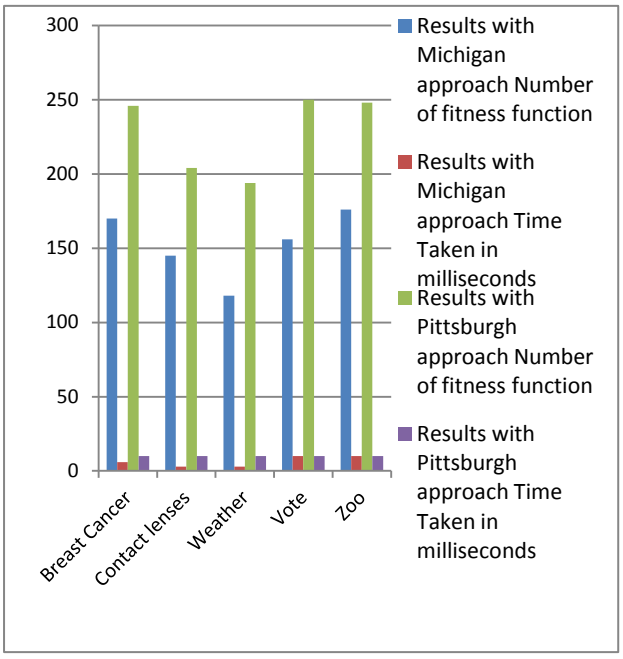

Fig1: Result comparison on bases of number of generations and number of rules

The second comparison of the result analysis is discussed in the following table in which comparison of the fitness function of the rule and time taken by the rule on both Pittsburgh and Michigan approach on the 5 data sets. In order to achieve the results in the form of graph provides the values on both the X-axis and Y-axis of the graph, along the X-axis various data sets are taken and along the Y-axis fitness function and time taken by the rule is taken. Its seen clearly in the graph that fitness function of the rule increases as a result of which much better results are achieved but the time taken by the Pittsburgh approach is more as compared to the Michigan this is because that in the Pittsburgh approach the set of rules are there in a chromosome due to which more time is taken but in previous approach each rule is considered as a single chromosome.

Table 2: Comparison on bases of fitness function and time taken

Data sets	**Results with Michigan approach**		**Results with Pittsburgh approach**	
	Fitness function	**Time Taken**	**Fitness function**	**Time Taken**
Breast Cancer	0.8247	14447	0.8442	4140
Contact lenses	0.7882	2341	0.8762	7316
Weather	0.7568	1374	0.8442	4244
Vote	0.8280	2658	0.9180	5452
Zoo	0.8286	1398	0.9299	2737

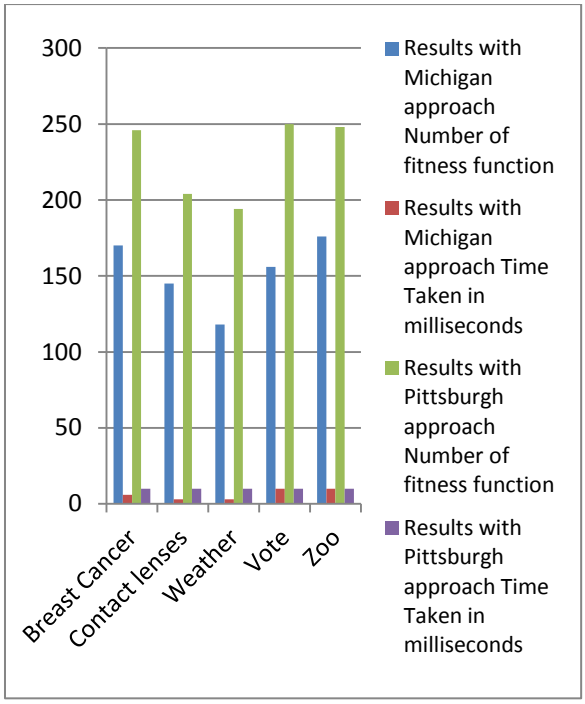

Fig2: Result comparison on bases of number of generations and number of rules

Thus on analyzing the both comparisons it came into the decision that the overall performance of the Pittsburgh approach is much better than the Michigan approach.

5. CONCLUSION & FUTURE SCOPE

By discussing the results in the above section, it has been noticed that the newly proposed approach i.e Pittsburgh approach is much better than the Michigan approach. In the future work, one can implement the results by using fuzzy approach or by using the neuro fuzzy technique

6. REFERENCES

[1] Indra k and kanmanis, March 2012 " Performance analysis of genetic algorithm for mining association rules" *International Journal Of Computer Science*, issues Vol. 9, issue 2, pp: 318-376.

[2] Rajul Anand, Abhishek Vaid, Pramod Kumar Singh 2009 "Association Rule Mining Using Multi-Objective evolutionary algorithm: Strengths and challenges" *IEEE Conference*, , pp:385-389.

[3] Rupali Haldulakar and Prof. Jitender Aggarwal March 2011 "optimization and association rule mining through genetic algorithm*" International Journal Of Computer Sciences And Engineering* Vol. 3, No. 3, , pp: 1252-1259.

[4] Jian Hu and Xing Yang Li2007 "Association rule mining using multi-objective co evolutionary algorithm" *Ieee International Conference On Computational Intelligence And Security Workshop*, , pp: 405-408.

[5] Basheer Mohamad, February 2013 "Discovering interesting association rules a multiobjective genetic algorithm" *International Journal Of Applied Information System*, Vol. 5 No. 3, pp: 47-52.

[6] Sanat Jain, Swati Ka April 2012bra "Mining and optimization of association rules using effective algorithm" *International Journal Of Emerging Technology And Advanced Engineering*, Vol. 2 issue 4.

[7] J.Malar Vizhi and Dr. T.Bhuvaneswari Jan 2012 " Data quality measurement on categorical data using genetic algorithm" *International Journal And Determining And Knowledge Management Process*, Vol. 2, 01, pp: 33-42,

[8] Sonia Sharma, vinay chopra September 2013 " Association Rule Mining: A Multi-objective Genetic Algorithm Approach Using Pittsburgh Technique" International Journal of Recent Technology and Engineering ,Vol 2 Issue 4.