International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-2 Number-4 Issue-6 December-2012

Reduction of Negative and Positive Association Rule Mining and Maintain Superiority of Rule Using Modified Genetic Algorithm

Nikhil Jain[1],Vishal Sharma[2],Mahesh Malviya[3]
Department of Computer Science and Engineering Jawaharlal Institute of Technology,Khargone (M.P.) India

Abstract

Association rule mining play important rule in market data analysis and also in medical diagnosis of correlated problem. For the generation of association rule mining various technique are used such as Apriori algorithm, FP-growth and tree based algorithm. Some algorithms are wonder performance but generate negative association rule and also suffered from Superiority measure problem. In this paper we proposed a multi-objective association rule mining based on genetic algorithm and Euclidean distance formula. In this method we find the near distance of rule set using Euclidean distance formula and generate two class higher class and lower class .the validate of class check by distance weight vector. Basically distance weight vector maintain a threshold value of rule itemsets. In whole process we used genetic algorithm for optimization of rule set. Here we set population size is 1000 and selection process validate by distance weight vector. Our proposed algorithm distance weight optimization of association rule mining with genetic algorithm compared with multi-objective association rule optimization using genetic algorithm. Our proposed algorithm is better rule set generation instead of MORA method.

Keywords

***Association Rule Mining, Negative and Positive rules, Superiority, Genetic algorithm*.**

1. Introduction

Association rule mining is a technique to detect the hidden facts in large dataset and draw interferences on how subsets of items influence the presence of other subsets. Association rule mining aims to find strong relation between attributes. All frequent generalized patterns are not very efficient because a portion of the frequent patterns are redundant in the association rule mining. This is why this algorithm produces some redundant rule along with the interesting rule. This drawback can be overcome with the help of genetic algorithm. Since most of the data mining approaches uses the greedy algorithm instead of genetic algorithm. Genetic algorithm is somewhat best as compare to the greedy algorithm because it performs a global search and copes better with the attribute interaction. In genetic algorithm population evolution is simulated. Genetic algorithm is a biological technique which uses chromosome as an element on which solutions (individuals) are manipulated. Generally association rule focuses on finding positive relationship between the data set. Negative association rule is also important in analysis of intelligent data. Negative association rule mining is used where a domain has too many factors. Negative association rule mining works in reverse manner as it decides whether which one is important instead of checking all rules. But problem with the negative association rule is it uses large space and can take more time to generate the rules as compare to the traditional mining association rule. So a better approach called generalized negative association rule is proposed. In the generalized association rule database is scanned once and transaction is transformed into space reduced structure. The association rule mining problem can be categorized in numerical and categorical attributes in a database. The application of association rule mining is on market basket data, whether prediction, multimedia data. The rest of paper is organized as follows. In Section 2 discuss related work of association rule mining. The Section 3 discuss genetic algorithm and distance formula. Section 4 discusses proposed algorithm .section 5 discuss experimental result followed by a conclusion in Section 6.

2. Related Work

Peter P. Wakabi–Waiswa, Venansius Baryamureeba and Karunakaran Sarukesi entitled "Optimized Association Rule Mining with Genetic Algorithms" propose an optimized association rule mining using genetic algorithm [1]. The Association Rule Mining (ARM) approach to data mining was introduced in as a structured mechanism for unearthing hidden facts in large datasets and drawing inferences on how a subset of items influences the presence of another subset. A popular approach to mining optimal rules is to establish a partial ordering of the rules based on a set of metrics [8], [9]. The are many rule interestingness metrics including support, confidence, conviction, lift, Laplace, gain, gini, and the chi square

value. In this paper authors have proposed Mining Optimized Association Rules Algorithm (MOAR) which maintains two populations: the internal population, and a Pareto−store. This algorithm incorporates partial ordering in the search mechanism for rules. The algorithm iteratively evaluates a solution by first decoding the bit string into a rule which is compared to the data in the database. For all records that cover a given consequent the values are compared against the consequent to calculate the supports of the antecedent and the consequent. After the dataset scan, the measures of strength used as the objectives are calculated for each rule. This algorithm adopts the Pareto dominance approach pro-posed in for maintaining multiple stable niches. This ensures that the individuals in the Pareto store are uniformly distributed near the Pareto−optimal front. To determine the dominance of an individual all individuals in the internal population and the Pareto store are evaluated. Sahar M. Ghanem, Mona A. Mohamed and Magdy H. Nagi entitled "EDP-ORD: Efficient Distributed/Parallel Optimal Rule Discovery" a problem of yielding too many rules which are infeasible when the minimum support is low, in association rule discovery algorithms [2]. Association rules discovery is the task of inferring rules, which states that certain values occur with other values above a certain frequency in data with some pre-specified certainty. Association rule discovery generates all rules satisfying some constraints, but yields too many rules and is infeasible when the minimum support is small. Optimized rule discovery [ORD] is an efficient alternative that generates a smaller set of rules by pruning redundant rules. ORD [1] is an optimal class association rule mining algorithm that prunes the candidate rules according to interestingness metric and is independent of the data presentation. Sandeep Singh Rawat and Lakshmi Rajamani entitled "Probability Apriori based Approach to Mine Rare Association Rules" proposes difficulties occurred in setting the rare association rules to handle unpredictable items [3]. Authors propose multiple minsup based apriori-like approach called Probability Apriori Multiple Minimum Support (PAMMS) to efficiently discover rare association rules. The problem of mining association rules is to discover all rules that satisfy minsup and minconf constraints. An itemset that satisfies minsup constraint is called frequent itemset or frequent pattern. Senduru Srinivasulu and P.Sakthivel entitled "Extracting Spatial Semantics in Association Rules for Weather Forecasting Image" a new approach to apply association rule on weather forecasting image is proposed [4]. The objective of entitled paper is to predict the temperature, relative humidity, rainfall, wind speed and atmospheric pressure applying association rule on weather forecasting image stored in the databases. Li-Min Tsai, Shu-Jing Lin and Don-Lin Yang entitled "Efficient Mining of Generalized Negative Association Rules" proposes the importance of negative association rule and a method to improve the negative association rule [5]. The negative association rule mining can be applied to a domain that has too many types of factors. Negative association rules can help users quickly decide which ones are important instead of checking too many rules. Algorithms for discovering negative association rules are not widely discussed. The discovery procedure of these algorithms can be decomposed into three stages: (1) find a set of positive rules; (2) generate negative rules based on existing positive rules and domain knowledge; (3) prune the redundant rules.

3. Genetic Algorithm and Distance Formula

In the process of optimization of algorithm of association rule mining we used KNN method for classification of superior support count and confidence value of itemset. KNN is a very famous algorithm for data classification. Here we describe process of KNN methodology for classification of support and confidence. Suppose each sample in our data set has n attributes which we combine to form an n-dimensional vector: x = (x1, x2. . . xn). These n attributes are considered to be the independent variables. Each sample also has another attribute, denoted by y (the dependent variable), who's value depends on the other n attributes x [11]. We assume that y is a categorical variable, and there is a scalar function, f, which assigns a class, y = f(x) to every such vectors. We do not know anything about f (otherwise there is no need for data mining) except that we assume that it is smooth in some sense. We suppose that a set of T such vectors are given together with their corresponding classes: x (i), y (i) for i = 1, 2 . . . T. This set is referred to as the training set. The problem we want to solve is the following. Supposed we are given a new sample where x = u. We want to find the class that this sample belongs. If we knew the function f, we would simply compute v = f (u) to know how to classify this new sample, but of course we do not know anything about f except that it is sufficiently smooth. The idea in k-Nearest Neighbor methods is to identify k samples in the training set whose independent variables x are similar

to u, and to use these k samples to classify this new sample into a class, v. If all we are prepared to assume is that f is a smooth function, a reasonable idea is to look for samples in our training data that are near it (in terms of the independent variables) and then to compute v from the values of y for these samples. When we talk about neighbors we are implying that there is a distance or dissimilarity measure that we can compute between samples based on the independent variables. For the moment we will concern ourselves to the most popular measure of distance: Euclidean distance. The Euclidean distance between the points x and u is

$$d(\mathbf{x}, \mathbf{u}) = \sqrt{\sum_{i=1}^{n} (x_i - u_i)^2}. \quad \text{.....(1)}$$

The simplest case is k = 1 where we find the sample in the training set that is closest (the nearest neighbor) to u and set v = y where y is the class of the nearest neighboring sample. It is a remarkable fact that this simple, intuitive idea of using a single nearest neighbor to classify samples can be very powerful when we have a large number of samples in our training set [11]. It is Possible to prove that if we have a large amount of data and used an arbitrarily sophisticated classification rule, we would be able to reduce the misclassification error at best to half that of the simple 1-NN rule. For k-NN we extend the idea of 1-NN as follows. Find the nearest k neighbors of u and then use a majority decision rule to classify the new sample. The advantage is that higher values of k provide smoothing that reduces the risk of over-fitting due to noise in the training data. In typical applications k is in units or tens rather than in hundreds or thousands. Notice that if k = n, the number of samples in the training data set, we are merely predicting the class that has the majority in the training data for all samples irrespective of u. This is clearly a case of over-smoothing unless there is no information at all in the independent variables about the dependent variable. For the process of separation of class of candidate key for generation of association rule mining by KNN classification ,this classification whole class in two section ,in one section we classified only higher support vale and another section of class contain lower value of class. The process of searching of data according to given support of transaction table we used genetic algorithm for better searching of classified class and finally generated optimized rule.

4. Proposed Algorithm

The proposed algorithm is a combination of support weight value and near distance of superior candidate key and parity based selection of rule based on group value of rule .Support weight key is a vector value given by the transaction data set and plays a role of rule selection on the base of genetic parity Order. The support value passes as a vector for finding a near distance between superior candidate key. After finding a superior candidate key the nearest distance divide into two classes, one class take a higher odder value and another class gain lower value for rule generation process. The process of selection of class also reduces the passes of data set. After finding a class of lower and higher of given support value compares the value of distance weight vector. Here distance weight vector work as a fitness function for selection process of genetic algorithm. Here we present steps of process of algorithm step by step and finally draw a flow chart of complete process.

Steps of algorithm (DWOARM)

1. Select data set
2. Put value of support and confidence
3. Start scanning of transaction table
4. Count frequent items
5. Generate frequent itemsets
6. Check the transaction set of data is null
7. Put the value of support as weight
8. Compute the distance with Euclidean distance formula
9. Generate distance vector value for selection process
10. Assign random parity of each group of selected vector
11. Initialized a population set (t=1)
12. Compare the value of distance vector with population set
13. Selected value of parity arrange by distance weight factor.
14. If value of support greater than vector value of parity order.
15. Processed for encoded of data
16. Encoding format is binary
17. After encoding offspring are performed
18. Set the value of probability for mutation and the value of probability is 0.006.
19. Set of rule is generated.
20. Check superiority of rule
21. If rule is not superior go to selection process
22. Else optimized rule is generated.
23. Exit

Now we explain complete process of algorithm shows block diagram of proposed algorithm using genetic algorithm.

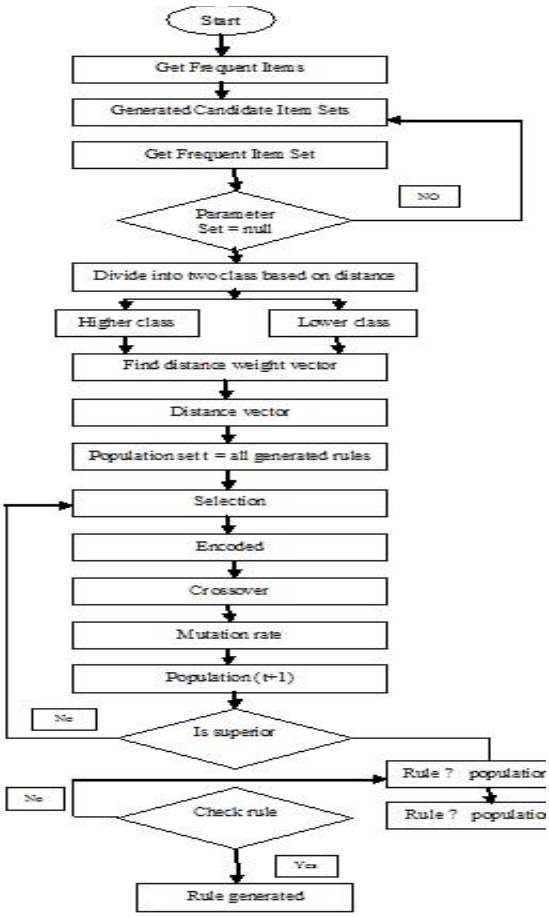

Figure.1: shows that proposed block model of algorithm.

5. Simulation Results

To investigate the effectiveness of the proposed method implement in Matlab 7.8.0 and testing of result we used wine data set and chess data set, that data set provided by UCI machine laboratory and apply our proposed algorithm and pervious algorithm with support and confidence. Our proposed algorithm distance weight optimization of association rule mining with genetic algorithm compared with multi-objective association rule optimization using genetic algorithm. Our proposed algorithm is better rule set generation instead of MORA method.

Table 1: shows that given value of support and confidence for uci data set.

ATTRIBUTES	A	B	C	D	E	F	G
MINIMUM SUPPORT	2	4	6	7	8	9	10
MINIMUM OONFIDENCE	.1	.2	.3	.4	.5	.6	.7
EXECUTION TIME	4.634764	4.683445	2.115712	2.132612	2.082319	2.103768	0.516367
NO OF RULES	3262	3262	1932	1932	1932	1932	602

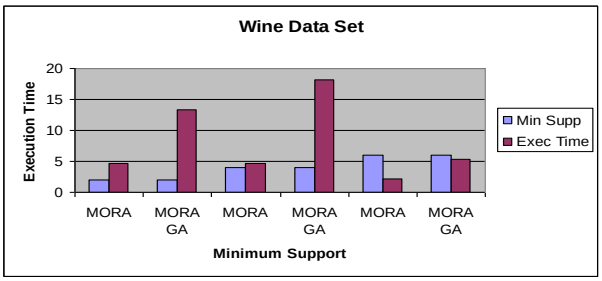

Figure.2: Shows the comparative value of minimum support and execution time of MORA and MORA GA algorithm for the processor extraction of rule. MORA GA takes more time in comparison of sample MORA algorithm.

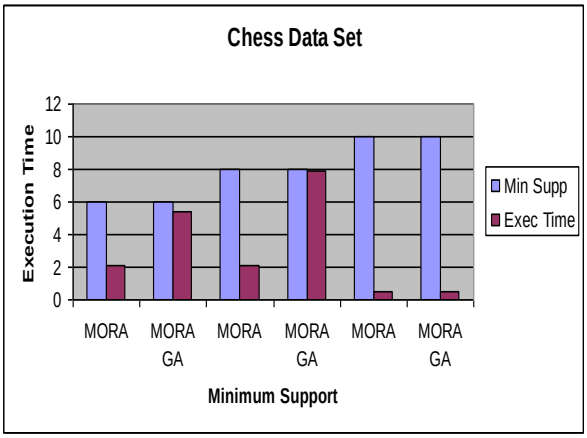

Figure.3: Shows the comparative value of minimum support and execution time of MORA and MORA GA algorithm for the processor extraction of rule. MORA GA takes more time in comparison of sample MORA algorithm.

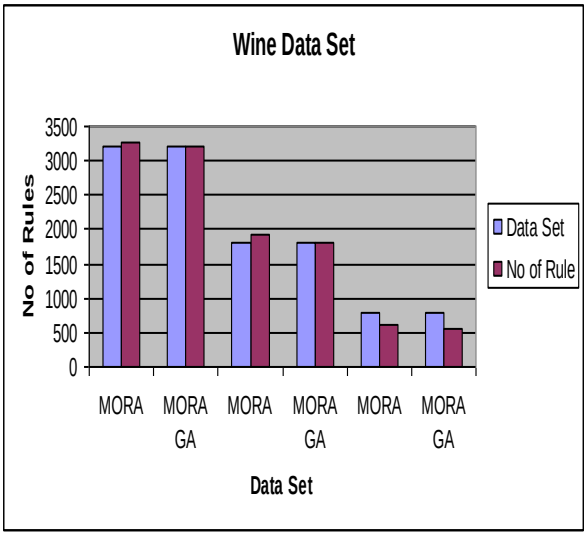

Figure. 4: show the comparative value of no of rules and data set of MORA and MORA GA algorithm for processor extraction of rule. MORA GA produces less no of rules than MORA algorithm.

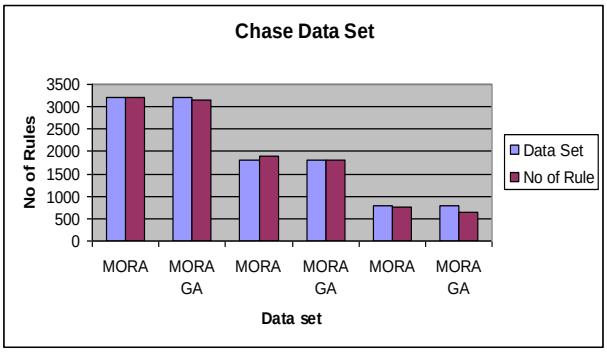

Figure.5: shows the comparative value of no of rules and data set of MORA and MORA GA algorithm for processor extraction of rule. MORA GA produces less no of rules than MORA algorithm.

6. Conclusion and Future Work

We proposed a novel method for optimization of association rule mining. Our propped algorithm is combination of distance function and genetic algorithm. We have observed that when we modify the distance weight new rules in large numbers are found. This implies that when weight is solely determined through support and confidence, there is a high chance of eliminating interesting rules. With more rules emerging it implies there should be a mechanism for managing their large numbers. The large generated rule is optimized with genetic algorithm. We theoretically proofed a relation between locally large and globally large patterns that is used for local pruning at each site to reduce the searched candidates. We derived a locally large threshold using a globally set minimum recall threshold. Local pruning achieves a reduction in the number of searched candidates and this reduction has a proportional impact on the reduction of exchanged messages. In the process of distance weight calculation large number of rule set generated that rule set take huge amount of time in compression of MOARGM algorithm. In future we minimize the time complexity of our method.

References

[1] Peter P. Wakabi–Waiswa, Venansius Baryamureeba and Karunakaran Sarukesi "Optimized Association Rule Mining with Genetic Algorithms" in Seventh International Conference on Natural Computation, 2011.

[2] Sahar M. Ghanem, Mona A. Mohamed and Magdy H. Nagi "EDP-ORD: Efficient Distributed / Parallel Optimal Rule Discovery" in IEEE Transaction, 2011.

[3] Sandeep Singh Rawat and Lakshmi Rajamani "Probability Apriori based Approach to Mine Rare Association Rules" in 3rd Conference on Data Mining and Optimization (DMO), 2011.

[4] Senduru Srinivasulu and P.Sakthivel "Extracting Spatial Semantics in Association Rules for Weather Forecasting Image" in IEEE Transaction, 2010.

[5] Li-Min Tsai, Shu-Jing Lin and Don-Lin Yang entitled "Efficient Mining of Generalized Negative Association Rules" in IEEE International Conference on Granular Computing, 2010.

[6] By Pengfei Guo Xuezhi Wang Yingshi Han The Enhanced Genetic Algorithms for the Optimization Design 978-1-4244-6498-2/10 IEEE 2010.

[7] By Masaya Yoshikawa and Hidekazu Terai A Hybrid Ant Colony Optimization Technique for Job-Shop Scheduling Problems Software Engineering Research, Management and Applications (SERA'06) 0-7695-2656-X/06 2006.

[8] By Chi-Ren Shyu1,2, Matt Klaric1,2, Grant Scott1,2, and Wannapa Kay Mahamaneerat1 Knowledge Discovery by Mining Association Rules and Temporal-Spatial Information from Large-Scale Geospatial Image Databases 0-7803-9510-7/06 IEEE 2006.

[9] By LI Tong-yan, LI Xing-ming New Criterion for Mining Strong Association Rules in Unbalanced Events Intelligent Information Hiding and Multimedia Signal Processing 978-0-7695-3278-3/08 $25.00 © IEEE 2008.

[10] By Zhibo Chen, Carlos Ordonez, Kai Zhao Comparing Reliability of Association Rules and OLAP Statistical Tests Data Mining Workshops 978-0-7695-3503-6/08 IEEE 2008.

[11] By Lijuan Zhou Linshuang Wang Xuebin Ge Qian Shi A Clustering-Based KNN Improved Algorithm CLKNN for Text Classification Informatics in Control, Automation and Robotics 978-1-4244-5194-4/10 IEEE 2010 .

[12] By XING Xue CHEN Yao WANG Yan-en Study on Mining Theories of Association Rules and Its Application Information Technology and Ocean Engineering 978-0-7695-3942-3/10 2010 IEEE.

[13] By Senduru Srinivasulu P.Sakthivel Extracting Spatial Semantics in Association Rules for Weather Forecasting Image 978-1-4244-9008-0/10 IEEE 2010.

[14] By TIAN He, XU Jing, LIAN Kunmei, ZHANG Ying Research on Strong-association Rule Based Web Application Vulnerability Detection 978-1-4244-4520-2/09 IEEE 2009**.**

[15] By Dieferson Luis Alves de Araujo' , Heitor S. Lopes', Alex A**.** Freitas2 A Parallel Genetic Algorithm for Rule Discovery in Large Databases 0-7803-5731-0/99 IEEE.

[16] By Xiaofeng Yuan, Hualong Xu, and Shuhong Chen "Improvement on the Constrained Association Rule Mining Algorithm of Separate" 1-4244-0682-X/06 IEEE 2006.

Nikhil Jain(1987,Indore) received bachelor of engineering in Computer Science engineering from LNCT Indore under RGPV Bhopal in 2009.currently doing Master of Engineering(Software engineering) from Jawaharlal Institute of Technology Borawan (Khargone) under RGPV Bhopal M.P. India.