ISSN No. 0976-5697

Volume 4, No. 3, March 2013 (Special Issue)

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Reduction of Negative Rules in Association Rule Mining Using Distance Security and Genetic Algorithm

Girish Kumar Ameta
M.Tech. Scholar, Dept. of Computer Science & Engg.
Arya college of Engineering & IT Jaipur, India
kumargirish360@gmail.com

Chhavi Saxena
Assistant Professor, Dept. of Computer Science & Engg. Arya college of Engineering & IT Jaipur, India
Chhavisaxena_81@rediffmail.com

Abstract: The increasing rate of data is a challenging task for mined useful association rule in data mining. The classical association rule mining generate rule with various problem such as pruning pass of transaction database, negative rule generation and superiority of rule set. Time to time various researchers modified classical association rule mining with different approach. But in current scenario association rule mining suffered from superiority rule generation. The problem of superiority is solved by multi-objective association rule mining, but this process suffered continuity of rule generation. In this paper we are proposing a new algorithm distance weight optimization of association rule mining. In this method, we find the near distance of rule set that uses equalize distance formula and generate two class higher class and lower class .The validation of class checked by distance weight vector. Basically distance weight vector maintain a threshold value of rule item sets. 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.

Keywords: Data mining, distance weight optimization, negative association rule mining,

I. INTRODUCTION

This paper describes our new proposed algorithm *distance weight optimization of association rule mining*, implantation and working of algorithm. The proposed algorithm is implemented with the genetic algorithm and, compared with multi-objective association rule optimization using genetic algorithm. This paper also suggests that proposed algorithm is better rule set generator as compared to the MORA method. *Section 2* of this paper describes about introduction of the association rule mining and challenges of finding the interested patterns among the item sets. *Section 3* describes about existing approaches and description of the related work for better association rule mining and methods to meet such challenges of the data mining. *Section 4* describes the proposed algorithm and its usage in the association rule mining.

II. ASSOCIATION RULE MINING

Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. The problem is usually decomposed into two sub problems. One is to find those item sets whose occurrences exceed a predefined threshold in the database; those item sets are called frequent or large item sets. The second problem is to generate association rules from those large item sets with the constraints of minimal confidence[1]. Suppose one of the large item sets is Lk, Lk = {I1, I2… Ik}, association rules with this item sets are generated in the following way: the first rule is {I1, I2… Ik-1} ⇒ {Ik}, by checking the confidence this rule can be determined as interesting or not. Then other rules are generated by deleting the last items in the antecedent and inserting it to the consequent, further the confidences of the new rules are checked to determine the interestingness of them. Those processes iterated until the antecedent becomes empty. Since the second sub problem is quite straight forward, most of the researches focus on the first sub problem.

The first sub-problem can be further divided into two sub-problems: candidate large item sets generation process and frequent item sets generation process. We call those item sets whose support exceed the support threshold as large or frequent item-sets, those item sets that are expected or have the hope to be large or frequent are called candidate item sets. In many cases, the algorithms generate an extremely large number of association rules, often in thousands or even millions. Further, the association rules are sometimes very large[3]. It is nearly impossible for the end users to comprehend or validate such large number of complex association rules, thereby limiting the usefulness of the data mining results. Several strategies have been proposed to reduce the number of association rules, such as generating only “interesting” rules, generating only “no redundant” rules, or generating only those rules satisfying certain other criteria such as coverage, leverage, lift or strength. All methodology and process are not described here. But some related work in the field of association rule of mining by the name authors and their respective title.

A. Improvement on the Constrained Association Rule Mining Algorithm of Separate:

In this title authors describe the constrained technique for optimization of association rule mining as Separate is a desirable algorithm in terms of efficiency and candidate

CONFERENCE PAPER
II International Conference on
“Advance Computing and Creating Entrepreneurs (ACCE2013)”
On 19-20 Feb 2013
Organized by
2nd SIG-WNs, Div IV & Udaipur Chapter , CSI , IEEE Computer Society Chapter India Council , IEEE Student Chapter Geetanjali Institute of Technical Studies, Udaipur, Rajasthan, India

generation. However, Separate is not perfect due to deficiency of its joint function, especially when the length of item set or the number of candidate item sets is large. In this paper, three lemmas are proposed and proved mathematically; and based on these lemmas, a novel early stop function is designed elaborately. The early stop algorithm is capable of breaking the process of loop in the case of dissatisfying the join term, and by this means, performance is improved remarkably. Experiments have demonstrated that the proposed algorithm is more preferable compared with the currently-used join function. To improve the performance of Separate algorithm, a novel Early Stop algorithm is designed elaborately according to three lemmas. It has been validated experimental that Early Stop outperforms Join function in terms of execution time, although there is no any daunting programming effort involved. In the future work, the authors will consider the application of Early Stop in other Apriori-based algorithms.

B. An Improved Multiple Minimum Support Based Approach to Mine Rare Association Rules:

In this title author describe a minimum support of multiple term for optimization of association rule mining. Rare association rules are the association rules containing rare items. Rare items are less frequent items. For extracting rare item sets, the single minimum support (minsup) based approaches like Apriori approach suffer from "rare item problem" dilemma. At high minsup value, rare itemsets are missed, and at low minsup value, the number of frequent item sets explodes. To extract rare item sets, an effort has been made in the literature in which minsup of each item is fixed equal to the percentage of its support. Even though this approach improves the performance over single minsup based approaches, it still suffers from "rare item problem" dilemma. If minsup for the item is fixed by setting the percentage value high, the rare item sets are missed as the minsup for the rare items becomes close to their support, and if minsup for the item is fixed by setting the percentage value low, the number of frequent item sets explodes. In this paper, we propose an improved approach in which minsup s fixed for each item based on the notion of "support difference". The proposed approach assigns appropriate minsup values for frequent as well as rare items based on their item supports and reduces both "rule missing" and "rule explosion" problems. Experimental results on both synthetic and real world datasets show that the proposed approach improves performance over existing approaches by minimizing the explosion of number of frequent item sets involving frequent items and without missing the frequent item sets involving rare items. Most important, the proposed approach ensures that the difference between the support of an item and the corresponding minimum support remains constant for all items including rare items. As a result, it efficiently reduces the explosion of frequent item sets involving frequent items without affecting the extraction of frequent item sets involving rare items. We have evaluated the performance of the proposed approach by conducting experimental results on both synthetic and real world datasets. The results show that, as compared to existing approaches, the proposed approach prunes frequent item sets involving frequent items in a more efficient manner and without missing the frequent item sets involving rare items.

C. Optimized Association Rule Mining with Genetic Algorithms:

The mechanism for unearthing hidden facts in large datasets and drawing inferences on how a subset of items influences the presence of another subset is known as Association Rule Mining (ARM). There is a wide variety of rule interestingness metrics that can be applied in ARM. Due to the wide range of rule quality metrics it is hard to determine which are the most 'interesting' or 'optimal' rules in the dataset. In this paper we propose a multi–objective approach to generating optimal association rules using two new rule quality metrics: syntactic superiority and transactional superiority. These two metrics ensure that dominated but interesting rules are returned to not eliminate from the resulting set of rules. Experimental results show that when we modify the dominance relations new interesting rules emerge implying that when dominance is solely determined through the raw objective values there is a high chance of eliminating interesting rules. Keywords: optimal association rules, genetic algorithms, multi–objective interestingness metrics we have observed that when we modify the dominance relations new rules in large numbers are found. This implies that when dominance 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 and also to significantly improve the response time of the algorithm.

III. DESCRIPTION OF APPROACHES

We proposed a novel algorithm for optimization of association rule mining, the proposed algorithm resolve the problem of negative rule generation and also optimized the process of superiority of rules. Superiority of association rule mining is a great challenge for large dataset. In the generation of superiority of rules association existing algorithm or method generate a series of negative rules, which generated rule affected a performance of association rule mining. In the process of rule generation various multi objective association rule mining algorithm are proposed but all these are not solve superiority problem of association rule mining.

In this paper we proposed distance weight optimizations of association rule mining with genetic algorithm. In this algorithm we used second odder quadratic equation and nearest neighbor classification technique for the selection of set of candidate of superiority of key for generation of rules. In the generation of rule selection of support value of transaction data set is play a important role , for this role we used heuristic search algorithm for better

CONFERENCE PAPER
II International Conference on
"Advance Computing and Creating Entrepreneurs (ACCE2013)"
On 19-20 Feb 2013
Organized by
2nd SIG-WNs, Div IV & Udaipur Chapter , CSI , IEEE Computer Society Chapter India Council ,
IEEE Student Chapter Geetanjali Institute of Technical Studies, Udaipur, Rajasthan, India

searching of support value for generation of optimized association rule.

In the process of novel algorithm for rule optimizations fist we discuss association rule mining, KNN and genetic algorithm and finally we proposed a hybrid method for optimization of association rule mining (DWORAM).

A. KNN:

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 item set. 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), whose 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

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

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.

B. Genetic Algorithm:

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. Here we discuss process of genetic algorithm.

Genetic Algorithm (GA), first introduced by John Holland in the early seventies, is the powerful stochastic algorithm based on the principles of natural selection and natural genetics, which has been quite successfully, applied in machine learning and optimization problems. To solve a Problem, a GA maintains a population of individuals (also called strings or chromosomes) and probabilistically modifies the population by some genetic, operators such as selection, crossover and mutation, with the intent of seeking a near optimal solution to the problem. Coding to Strings in GA[5,6], each individual in a population is usually coded as coded as a fixed-length binary string. The length of the string depends on the domain of the parameters and the required precision. For example, if the domain of the parameter x is [2,5] and the precision requirement is six places after the decimal point, then the domain [2,5] should be divided into 7,000,000 equal size ranges. The implies that the length of the string requires to be 23, for the reason that 4194304=222<7000000<223=8388608 the decoding from a binary string <b22b21…b0> into a real number is straightforward and is completed in two steps.

a. Convert the binary string <b22b21…b0> from the base 10 by

$$x' = \sum_{i=0}^{22} b_i 2^i$$

b. Calculate the corresponding real number x by

$$x = -2.0 + x' \frac{7}{2^{23} - 1}$$

a) Initial Population: The initial process is quite simple. We create a population of individuals, where individual in a population is a binary string with a fixed-length, and every bit of the binary string is initialized randomly.
b) Evaluation: In each generation for which the GA is run, each individual in the population is evaluated against the unknown environment. The fitness values are associated with the values of objective function.
c) Genetic Operators: Genetic operators drive the evolutionary process of a population in GA, after the Darwinian principle of survival of the fittest and naturally occurring genetic operations.

The most widely used genetic operators are reproduction, crossover and mutation. To perform genetic operators, one must select individuals in the population to be operated on .The selection strategy is chiefly based on the fitness level of the individuals actually presented in the population. There are many different selection strategies based on fitness. The most popular is the fitness proportionate selection. After a new population is formed by selection process, some members of the new populations undergo transformations by means of genetic operators to form new solutions (a recombination step). Because of intuitive similarities, we only employ during the recombination phase of the GA three basic operators: reproduction, crossover and mutation, which are controlled by the parameter pr, pc and pm (reproduction probability, crossover probability and Mutation probability), respectively. Let us illustrate these three genetic operators. As an individual is selected, reproduction operators only copy it from the current population into the new population (i.e., the new generation) without alternation. The crossover operator starts with two selected individuals and then the crossover point (an integer between 1 and L-1, where L is the length of strings) is selected randomly. Assuming the two parental individuals are x1 and x2, and the crossover point is 5 (L=20). If

X1= (01001|101100001000101)
X2= (11010|011100000010000)

Then the two resulting offspring are

X'1= (01001|011100000010000)
X'2= (11010|101100001000101)

The third genetic operator, mutation, introduces random changes in structures in the population, and it may occasionally have beneficial results: escaping from a local optimum. In our GA, mutation is just to negate every bit of the strings, i.e., changes a 1 to 0 and vice versa, with probability pm.

IV. PROPOSED DWOARM METHOD

a. (Distance weight optimization of association rule mining with genetic algorithm): The proposed algorithm is a combination of support weight value and near distance of superior candidate key .Support weight key is a vector value given by the transaction data set. 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. Dividing. After finding a class of lower and higher of given support value, compare the value of distance wet 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):

a. Select data set
b. Put value of support and confidence
c. Start scaning of transaction table
d. Count frequent items
e. Generate frequent itemsets
f. Check the transaction set of data is null
g. Put the value of support as weight
h. Compute the distance with equilidaen distance formula
i. Generate distance vector value for selection process
j. Initialized a population set (t=1)
k. Compare the value of distance vector with population set
l. If value of support greater than vector value
m. Processed for encoded of data
n. Encoding format is binary
o. After encoding offspring are performed
p. Set the value of probability for mutation and the value of probability is 0.006.
q. Set of rule is generated.
r. Check superioty of rule.
s. If rule is not superior go to selection process
t. Else optimized rule is generated.
u. Exit

V. CONCLUSION

In this dissertation 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

CONFERENCE PAPER
II International Conference on
"Advance Computing and Creating Entrepreneurs (ACCE2013)"
On 19-20 Feb 2013
Organized by
2nd SIG-WNs, Div IV & Udaipur Chapter , CSI , IEEE Computer Society Chapter India Council ,
IEEE Student Chapter Geetanjali Institute of Technical Studies, Udaipur, Rajasthan, India

proportional impact on the reduction of exchanged messages.

VI. REFERENCES

[1] By Rakesh Agrawal Tomasz Imielinski Arun Swami Mining Association Rules between Sets of Items in Large Databases ACM SIGMOD Conference Washington DC, USA, May 1993.

[2] By Rakesh Agrawal Ramakrishnan Srikant_ Fast Algorithms for Mining Association RulesVLDB ConferenceSantiago, Chile, 1994.

[3] By Ramakrishnan Srikant* Rakesh Agrawal Mining Generalized Association Rules VLDB Conference Zurich, Swizerland, 1995.

[4] By N. Chaiyaratаiia and A. M. S. Zalzala Recent Developments in Evolutionary and Genetic Algorithms: Theory and Applications Innovations and Applications, 2-4 September 1997, Conference Publication NO. 4 4 6 ,IEEE , 1997.

[5] By Q. C. Meng , T.J. Feng I , *2.* Chen I , C.J. Zhou , J.H. Bo2 Genetic Algorithms Encoding Study and A Sufficient Convergence Condition of GAS 0-7803-5731-0/9)sk$l0.00 0 IEEE 1999.

[6] By Pengfei Guo Xuezhi Wang Yingshi Han The Enhanced Genetic Algorithms for the Optimization Design 978-1-4244-6498-2/10/$26.00 © 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 $20.00 © 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/$20.00 © 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 $25.00 © 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/$26.00 ©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 $26.00 © 2010 IEEE

[13] By Senduru Srinivasulu P.Sakthivel Extracting Spatial Semantics in Association Rules for Weather Forecasting Image 978-1-4244-9008-0/10/$26.00 © 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/$25.00 © 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$$10.00109 99 IEEE.