Association Rules Optimization using Particle Swarm Optimization Algorithm with Mutation

Mayank Agrawal, Manuj Mishra, Shiv Pratap Singh Kushwah

Abstract - In data mining, Association rule mining is one of the popular and simple method to find the frequent item sets from a large dataset. While generating frequent item sets from a large dataset using association rule mining, computer takes too much time. This can be improved by using particle swarm optimization algorithm (PSO). PSO algorithm is population based heuristic search technique used for solving different NP-complete problems. The basic drawback with PSO algorithm is getting trapped with local optima. So in this work, particle swarm optimization algorithm with mutation operator is used to generate high quality association rules for finding frequent item sets from large data sets. The mutation operator is used after the update phase of PSO algorithm in this work. In general the rule generated by association rule mining technique do not consider the negative occurrences of attributes in them, but by using PSO algorithm over these rules the system can predict the rules which contains negative attributes.

Keywords: Particle Swarm Optimization (PSO), Mutation, Association rule, Support, Confidence, Frequent item set, Data mining.

I. INTRODUCTION

Now a days, growth of databases increase day-by-day due to the number of requirements or number of customers or end users. In this scenario, data mining [2, 9] plays an important rule for mining data according to the customer's requirements. Association rule mining is one of the popular and well known research methods for discovering interesting relationships between variables in large information repository or databases. Paper [1] describes how data mining and knowledge discovery are related to different fields like machine learning; statistics etc. genetic algorithm [5] based association rule finding is also discussed in paper [3]. For finding association rules, minimum support value plays an important rule. Paper [4] present a genetic algorithm based strategy for discovering association rules without specifying the value of minimum support. Many techniques have been proposed to optimize association rules [6, 7]. Also genetic algorithm based techniques [10, 11] have proposed previously. In this work, Particle swarm optimization algorithm with mutation operator based association rule optimization technique has proposed. The remainder of this paper is organized as follows. Section 2 describes the particle swarm optimization algorithm with mutation. In section 3, Association rule mining is explained. Proposed algorithm is explained in section 4. Experimental results and parameter setup for result comparison are shown in section 5. Finally, section 6 concludes the paper.

Manuscript Received on February 16, 2015.
Mayank Agrawal, Research Scholar, ITM Universe, Gwalior, India.
Manuj Mishra, Asst. Prof., Department of CSE, ITM Universe, Gwalior, India.
Shiv Pratap Singh Kushwah, Asst. Prof., Department of CSE, ITM Universe, Gwalior, India.

II. PARTICLE SWARM OPTIMIZATION ALGORITHM WITH MUTATION

Kennedy and Eberhart introduced in 1995, a new algorithm, called Particle swarm optimization algorithm. This algorithm is based on the social as well as cognitive behavior of swarms, used to solve different kinds of problems related to different fields, especially in engineering and computer science field. Due to the information sharing and simple computation, this algorithm is very popular. In this algorithm, the individuals also called particles are distributed through the multi-dimensional search space where each particle represents a candidate solution to the optimization problem. The fitness value of each solution is based on the performance function of the problem that is being optimized. Here particles movements are affected by two key factors using information from particle-to-particle as well as iteration-to-iteration. The best solution, called pbest, is stored in particles memory as a result of iteration-to-iteration information and shared information among different particles. The best solution visited by any particle, called gbest, is stored in particles memory as a result of particle-to-particle information. These two factors are called social and cognitive components, respectively. After the PSO algorithm's each iteration if better or more dominating solution is found in terms of fitness value then the best solutions are updated for each individual. This process continues until desired solution for the optimization problem that is being solved, is not found. In multi-dimensional search space, the i-th particle position and velocity are represented by the following m-dimensional vectors, Yi = yi1, yi2, …,yim) and Vi = (vi1, vi2,…,vim)T. The i-th particle's best solution that is visited previously, denoted as Pi = (pi1, pi2, …,pim)T. here 'g' is denoted by the best particle index. The i-th particle velocity is updated using the following update equation given by

$$v_{id} = v_{id} + c_1 r_1 (p_{id} - x_{id}) + c_2 r_2 (p_{gd} - x_{id}), \quad (1)$$

and i-th particle position is updated using the equation given below

$$x_{id} = x_{id} + v_{id} \quad (2)$$

Where constants c1 is called cognitive scaling parameter and c2 is called social scaling parameters respectively, r1 and r2 represents random numbers, d represents dimension, i represents the particle index and S represents the size of the swarm. The Vmax, called maximum velocity, used to control the particle swarm global exploration ability.

Further, for better controlling the exploration and exploitation, a new concept of inertia weight is introduced in 1998 [2]. After adding the inertia weight concept in the PSO algorithm, the new velocity update equation becomes:

$$v_{id} = w * v_{id} + c_1 r_1 (p_{id} - x_{id}) + c_2 r_2 (p_{gd} - x_{id}) \quad (3)$$

In [3], the optimal strategy is to set the initial value of inertia weight w equal to 0.9 and reduce it linearly to the value of 0.4, allowing initial exploration followed by acceleration toward an improved global optimum. In paper [9, 11], one more phase in the form of mutation operator of genetic algorithm is added to original particle swarm optimization algorithm described above. In standard PSO algorithm, there are only 2 phases that described the overall working of this algorithm, but here one additional phase in the form of Mutation operator is added after the update phase. Now modified particle swarm optimization algorithm has three phases: initialization phase, particle update phase and mutation phase. With the help of mutation operator, there may be a possibility to change the local best position and the algorithm may not be trapped into local optima. In this work, the mutation phase is implemented on the probabilistic way in each iteration for searching food source during the life process of PSO optimization technique..
The overall algorithm is described in following steps:

1. Initialization Phase

REPEAT
REPEAT
Initialize particle randomly with sequence of tasks
UNTIL (Dimension size)
Calculate fitness of that particle
Calculate global best
UNTIL (Swarm Size)

REPEAT

2. Particle Update Phase

REPEAT
REPEAT
Particle positions update using PSO position update equation
UNTIL (Problem Dimension)
Calculate updated particle fitness
Update historical information, if needed, for global best
UNTIL (Swarm Size)

3. Mutation Operator Phase

If mutation criteria met then
Select random particles from current swarm for mutation operation
Apply mutation operation to randomly selected particle
Selected random particle position updated as a result of mutation.
Calculate updated particle fitness
Update the historical information, if needed, of global best
end if
UNTIL (Stopping Criteria not met)

III. ASSOCIATION RULE MINING

The main aim of association rule mining is to extract frequent item sets, correlation and association among different set of items in the transactional database, relational databases or other information repository. Association rule mining algorithm finds association rules in the form of:
IF AB and CD then HELLO
IF UV and XY then BYE
Here AB, CD, UV and XY are different objects out of which if any person takes AB and CD then due to high propbility, he will take HELLO. Similarly if he will choose UV and XY then he will choose BYE. In general, expressions which are in the form of A=>B, called association rules where A represents antecedent and B represents consequent. Association rules represent how many times B has occurred if A has already occurred depending on the chosen support and confidence value. Here support is nothing but the probability of items or item sets in the given database (like transactional or other) and confidence represents conditional probability.

Apriori Algorithm:

In general, Apriori algorithm [8] works on two phases – first phase is to choose minimum support value which is applied in the database to find frequent item sets while in second phase, these item sets and the minimum confidence constraints are used to generate rules.
The pseudo code for the Apriori algorithm are given as follows -
Step 1: let C_n be the candidate item set of size n.
Step 2: let F_n be the frequent item set of size n.
Step 3: F_1 = {Frequent items}
Step 4: REPEAT
Step 5: C_{n+1} = Candidates generated from F_k ;
Step 6: REPEAT for each transaction t in database
Step 7: increment the count of all candidates in C_{n+1} that are contained in t.
Step 8 : F_{k+1} = Candidates in C_{n+1} with minimum support.
Step 9: UNTIL (F_n not equal to ϕ)
Step 10: return $U_n F_n$

IV. PROPOSED METHODOLOGY

This section represents proposed methodology. Here PSO with mutation algorithm is applied over the rules gathered from apriori algorithm, to find frequent item sets. In order to use the PSO algorithm with mutation, the following points must be addressed: initial population, fitness value, employed, onlooker, mutation and scout bees. Here Initial population is generated using randomly generated transactions. To calculate the fitness value of an individual, following fitness value is used-

$f_i = 1 / (1 + f_i)$ if $f_i >= 0$

$1 + abs(f_i)$ otherwise

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Other points are same like standard PSO algorithm, discussed above.

The steps of proposed algorithm for generating optimal association rules via PSO with mutation are as follows-

Step 1: Start

Step 2: Load dataset

Step 3: Find frequent item sets using apriori algorithm. Suppose F is the set of all frequent item sets generated by apriori algorithm and X is the output set, containing all generated association rules, initialized to zero.

Step 4: Set the termination condition for the PSO with mutation algorithm.

Step 5: Depict each item sets of Z and apply PSO with mutation algorithm on selected members to generate association rules.

Step 6: Evaluate fitness value of each rule.

Step 7: If the fitness function satisfied the desired criteria then add these rules in output set.

Step 8: if the desired number of generations not completed then goto step 3

Step 9: Stop

Block diagram of proposed work:

Block diagram of the proposed algorithm for optimizing association rules are given below and shown in figure 1.

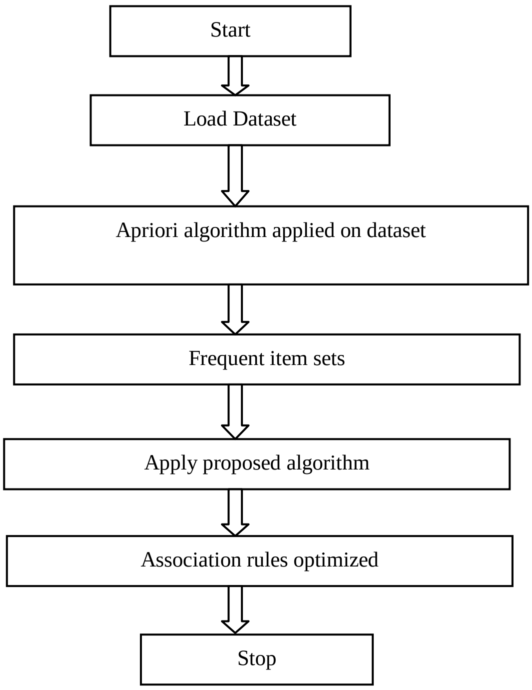

Fig. 1 shows the block diagram of proposed algorithm

V. EXPERIMENTAL RESULTS & PARAMETER SETUP

1. Data Sets :

To check the performance of the proposed work, different datasets are selected from UCI machine learning repository. Currently, 187 datasets are maintained by UCI machine learning research group. Out of these datasets, three popular datasets of Voting, Iris and Wine are selected for our experiments.

Details of these datasets are given below –

- Voting dataset – Features =16
  Instances = 435
  Class = 2
- Wine dataset – Features =13
  Instances = 178
  Class = 3
- Iris dataset – Features =04
  Instances = 150
  Class = 3

2. Parameter Settings:

There are four main control parameter which are used to test the performance of proposed algorithm. First control parameter is the number of food sources which is equal to 20 and also it is equal to the number of employed bees and onlooker bees. Second control parameter is the maximum cycle number (MCN) which is equal to 2000 in this experiment. Third and the fourth control parameter are the mutation probability which is equal to 0.1 and the limit value. After the final rule has generated, two control parameter for class prediction are quality weight (α) and coverage weight (β), both are initialized to 0.5 in this experiment. Proposed work is compared with KNN algorithm and standard PSO algorithm. Table 1 shows the performance classification accuracy. Figure 2 shows the graphical comparison between different algorithms.

Table 1: performance accuracy of classification

Datasets	***KNN (%)***	***ABC (%)***	***Proposed Work (%)***
Voting	95.10	97.21	97.47
Iris	94.08	96.44	97.89
Wine	96.22	98.13	98.75

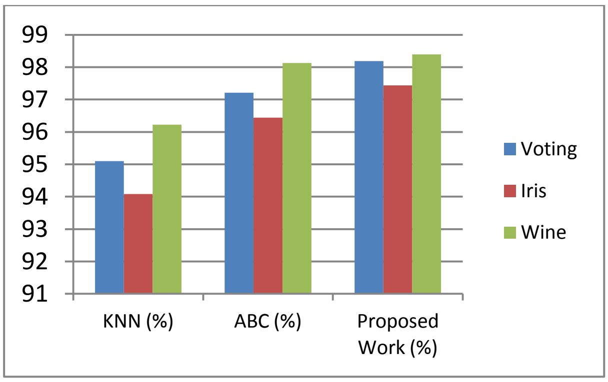

Figure 2 shows the performance of proposed work

VI. CONCLUSION

Now days, the size of the databases are increased day–by-day. To find frequent item sets, there is a need of association rule mining. In this work, generated association rules using apriori algorithm are optimized using particle swarm optimization with mutation algorithm, where mutation operator is used after the particle update phase of PSO algorithm.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

To check the performance of proposed work, three datasets of voting, wine and iris are used, collected from UCI machine learning repository. Experimental results show that the performance of the proposed work with previously proposed works. Future work is to use the proposed work with different databases.

REFERENCES

1. U. Fayyad and R. Uthurusamy, “Data Mining and Knowledge Discovery in Databases”, Communications of the ACM, vol. 39, no. 11, 1996, pp.24–34.
2. J. Han and M. Kamber, “Data Mining Concepts and Techniques”, Morgan Kaufmann, 2006.
3. W. Soto and A. Olaya-Benavides, "A Genetic Algorithm for Discovery of Association Rules." In Computer Science Society (SCCC), 2011, pp. 289-293.
4. X. Yan, C. Zhang and S. Zhang, “Genetic Algorithm- Based Strategy for Identifying Association Rules without Specifying Actual Minimum Support”, Expert Systems with Applications, vol. 36, 2009, pp. 3066–3076.
5. S. N. Sivanandamand and S. N. Deepa, “Introduction to Genetic Algorithms”, Springer-Verlag Berlin Heidelberg, 2008.
6. M. Anandhavalli and S. Kumar Sudhanshu, A. Kumar and M.K. Ghose, “Optimized Association Rule Mining Using Genetic Algorithm”, Advances in Information Mining, vol. 1, issue 2, 2009, pp. 01-04.
7. P. Wakabi-Waiswa and V. Baryamureeba, "Mining High Quality Association Rules using Genetic Algorithms", In Proceedings of the twenty second Midwest Artificial Intelligence and Cognitive Science Conference, 2009, pp. 73-78.
8. Markus Hegland, “The Apriori Algorithm – a Tutorial”, CMA, Australian National University, WSPC/Lecture Notes Series, 22-27. March 30, 2005.
9. Pujari A.K., “Data Mining Techniques”, Universities Press, 2001.
10. S. Ghosh, S. Biswas, D. Sarkar and P.P. Sarkar, “ Mining Frequent item sets using Genetic Algorithm”, international journal of artificial intelligence & applications (IJAIA), Vol. 1, No. 4, October 2010.
11. Mohit K. Gupta and Geeta Sikka, “ Association Rules Extraction using Multi-objective Feature of Genetic Algorihtm”, Proceedings of the world congresson engineering & computer science 2013 Vol II, WCECS 2013, 23-25 October- 2013, San Francisco, USA.
12. Amit Singh, Neetesh Gupta and Amit Sinhal, “Artificial Bee Colony Algorithm with Uniform Mutation”, Proceeding of the international conference of soft computing for problem solving (SocPros 2011), Vol. 130, pp 503-511.

AUTHORS PROFILE

Mr. Mayank Agrawal, is perusing M.Tech from ITM College Gwalior. His area of research is Data Mining and Soft Computing.

Prof. Manuj Mishra, is an Assistant professor of ITM college Gwalior in computer science department. He has published number of research papers in various research fields.

Prof. Shiv Pratap Singh Kushwah, is an Assistant professor of ITM college Gwalior in computer science department. He has published number of research papers in various research fields.