RESEARCH ARTICLE

A Study on Association Rule Mining Using ACO Algorithm for Generating Optimized ResultSet

NN Das[1], Anjali Saini[2]

[1] *Computer Science and Engineering, Itm University, Gurgaon, India*
[2] *Computer Science and Engineering, Itm University, Gurgaon, India*
[1] nndas@itmindia.edu; [2] angel.anjali43@gmail.com

Abstract - The Knowledge Discovery in Databases (KDD) field of data mining is concerned with the development of methods, techniques and algorithm which can make sense of the available data. Knowledge Discovery in Database is useful in finding trends, patterns and anomalies in the databases which is helpful to make accurate decisions for the future. Association rule mining finds collections of data attributes that are statistically related to the data available. In this paper, Ant Colony Optimization (ACO) improved association mining is suggested to perform the association mining on medical data set. During the rule generation, the global pruning will be used to eliminate the rules as well as attributes that are not effective. The work is about to generate the optimized and accurate resultset so that different decisions regarding the disease classification can be done.

Keywords -   Association Rule Mining; Ant Colony Optimization; Dataset and Attribute Filteration

I INTRODUCTION

Data mining process includes understanding the business requirements and its needs. While understanding the business requirements both data and business requirements should be understand. Then, using this business requirement it identifies data source and data format in this the data is prepared and modeled for evaluation and then using these data source and data format, it builds data model. This data model is used to build data structure. Then, the mining operation is performed on this data structure. Data mining field comprises of four main disciplines-Statistics(defines tools for measuring significance in the data),Machine learning(provide algorithm to induce knowledge from the data),Artificial intelligence(involve knowledge for encoding and search techniques),Data management and databases(provides an efficient way of accessing and maintaining data). The amount of data especially in the field of web service-based applications, architectures of service-oriented and cloud computing, is increasing every day by day and to extract the valuable information from them different data mining technologies are being used. Data mining is defined as the technology to extract the valuable and useful information from the data and to

analyze the patterns and association among these categorization .The categorization is a supervised form of machine learning. Machine learning comprises of supervised, unsupervised, semi- supervised and reinforced learning. In the supervised form of learning ,the learning is from trained data available.

Association Rule Mining is an important technique in data mining. A major concern today in Association Rule Mining is to improve the algorithmic performances .Association Rule Mining was used to find association between the data in large dataset In the algorithms of association mining, Apriori is the oldest which is offered by Agrawal R in 1993. Apriori is the best one under association mining. It uses a breadth first search technique.

 Ant Colony Optimization :In ACO, ants have the ability to select the shortest path among few possible paths connecting their nest to a food site. The a volatile chemical substance known as pheromone laid on the ground by the ants while walking and affecting in turn their moving decisions according to its local intensity is the mediator of this behavior. In this work, we will define the record-set in the form of network and at each node the rule analysis will be performed under the support and confidence value whereas rule path will be generated by ACO. The values generated from the rule will be represented as the pheromone values. Discovering Association rules in the data set is the important class of data mining and there is a wide need of finding association rules in the current scenarios. Though many algorithms have been proposed in this respect, but finding association rules in the large data set are still one of the cumbersome time consuming task.

II   LITERATURE REVIEW

Several studies are reported in literature for Association Rule Mining and ACO. The work is as follows:

In Year 2002, Adepele Olukunle performed a work," A Fast Algorithm for Mining Association Rules in Medical Image Data"[1]. This paper presents a fast association rule mining algorithm which is suitable for medical image data sets. Author provide a flavour of Presented implementation environment. Author also give an example, how Presented proposed algorithm work to assess its suitability.

 In Year 2006, Carlos Ordonez performed a work," Association Rule Discovery With the Train and Test Approach for Heart Disease Prediction"[2]. Association rules represent a rising technique to improve heart disease prediction. Author introduce an algorithm that uses search constraints to reduce searches for association rules on a training set,the number of rules and finally confirm them on an independent test set. Themedical significance of discovered rules is evaluated with support ,lift and confidence. Association rules are applied on a real data set containing medical records of patients with heart disease. In medical language, association rules relate heart perfusion measurements and risk factors to the degree of disease in four specific arteries.

In Year 2009, Gaurav N. Pradhan performed a work," ASSOCIATION RULE MINING IN MULTIPLE, MULTIDIMENSIONAL TIME SERIES MEDICAL DATA"[4]. In this paper, Author deal with real-life time series data of muscular activities of human participants obtained from multiple Electromyogram (EMG) sensors and discover patterns in these EMG data streams. Each EMG data stream is correlate with quantitative attributes such as energy of the signal and onset time which are required to be mined along with EMG time series patterns. Author propose a two-stage approach for this purpose: in the first stage, Presented emphasis is on discovering frequent patterns in multiple time series by doing sequential mining across time slices. Presented evaluation with large sets of time series data from multiple EMG sensors demonstrate that Presented two-stage approach speeds up the process of finding association rules in such multidimensional environment as compared to other methods and scales up linearly in terms of number of time series involved. Presented approach is generic and applicable to any multiple time series dataset format.

Mr.K.Ravikumar performed a work," ACO based spatial Data Mining for Traffic Risk Analysis"[5]. There was a first study aiming identifying and at predicting the accident risk of the roads. It used a decision tree that learns from the inventoried accident data and the description of the corresponding road sections. The existing work provided a pragmatic approach to multilayer geo-data mining. The process behind was to prepare input data by joining each layer table using a given spatial

criterion, then applying a standard method to build' a decision tree. Presented method has higher efficiency in performance of the discovery process and in the quality of trend patterns discovered compared to other existing approaches using non-intelligent decision tree heuristics.

In Year 2011, Pooia Lalbakhsh performed a work," Focusing on Rule Quality and Pheromone Evaporation to Improve ACO Rule Mining"[8]. In this paper an revised version of Ant-Miner algorithm is introduced and compared to the previously proposed ant-based rule mining algorithms. Presented algorithm modifies the rule pruning process and introduces a dynamic pheromone evaporation approach. The algorithm was run on five standard datasets and the average accuracy rate and numbers of discovered rules were analyzed as two important performance metrics of rule mining.

In Year 2011, Ghada Almodaifer performed a work," Discovering Medical Association Rules from Medical Datasets"[9]. In this paper, Author aim to find interesting medical association rules from medical datasets for prediction purposes. Author provide an association rule mining system that discovers constrained association rules in medical records that includes mathematical, categorical and image features.

In Year 2013, Divya Bhugra performed a work," Association Rule Analysis Using Biogeography Based Optimization"[12]. In this paper, Author have tried to make perfect the rules generated by Association Rule Mining using Biogeography Based Optimization(BBO).BBO has a way of sharing information between solutions depending on the migration mechanisms .The motivation of this paper is to use the feature of BBO for finding more accurate results.

 In Year 2013, K.Rameshkumar performed a work," Relevant Association Rule Mining from Medical Dataset Using New Irrelevant Rule Elimination Technique"[13]. This paper proposes the n-cross validation technique to reduce association rules which are irrelevant to the transaction set. The expected approach used partition based approaches are supported to association rule validation.

III PROPOSED WORK

 Complete work of proposed approach is divided in number of sub stages and each stage is defined in terms of some Algorithmic approach. The complete work is presented a series of comperation represented by individual algorithms.

a .   Data Collection

The first work of the approach is to collect the data to implement the whole concept. We are very clear about the database selection. In this work, medical dataset will be used. The data will be able present the disease analysis in a patient under different classes. The dataset can be diabetic dataset, heart disease dataset etc.

b.   Filtered Database Creation

Once the data is derived respective to the actual required database the next work is to perform the rule mining on this database .To perform the rule based mining the ACO improved Association mining is defined. ACO based algorithm searches for itemsets during its initial database pass and uses its result as the seed for discovering other large datasets during subsequent passes. Large or frequent itemsets are those which having the rules with support level above minimum   and those below are called small itemsets. The algorithm is based on the large itemset property which states: Any subset of a large itemset is large and if an itemset is not large and then none of its supersets are large.

c.   Knowledge Hiding

An association mining will be performed to find the most valuable items set from the whole database. Now on the basis of this algorithm we will find the filtered data value. Now according to the frequent pattern of the user the data will be divided in partitioned and data will be presented in the required form.

The basic work flow is listed below:

```
                    ┌─────────────┐
                    │    Start    │
                    └─────────────┘
                           │
                           ▼
                    ┌─────────────────┐
                    │  Input Dataset  │
                    └─────────────────┘
                           │
                           ▼
                    ┌──────────────────────┐
                    │ Perform Dataset      │
                    │ Analysis             │
                    └──────────────────────┘
                           │
                           ▼
                    ┌──────────────────────────┐
                    │ Perform the Attribute    │
                    │ Filteration              │
                    └──────────────────────────┘
                           │
                           ▼
            ┌──────────────────────────────────────┐
            │ Generate initial ruleset based on ACO │
            │ based Analysis under objective function│
            └──────────────────────────────────────┘
                           │
                           ▼
                    ┌──────────────────────────┐
                    │ Perform association      │
                    │ mining                   │
                    └──────────────────────────┘
                           │
                           ▼
            ┌──────────────────────────────────────┐
            │ Perform pruning based on confidence   │
            │ and support values                    │
            └──────────────────────────────────────┘
                           │
                           ▼
                    ┌──────────────────────┐
                    │ Perform Analysis     │
                    └──────────────────────┘
                           │
                           ▼
                    ┌──────────────────────┐
                    │ Get Result rules     │
                    └──────────────────────┘
                           │
                           ▼
                    ┌─────────────┐
                    │     End     │
                    └─────────────┘
```
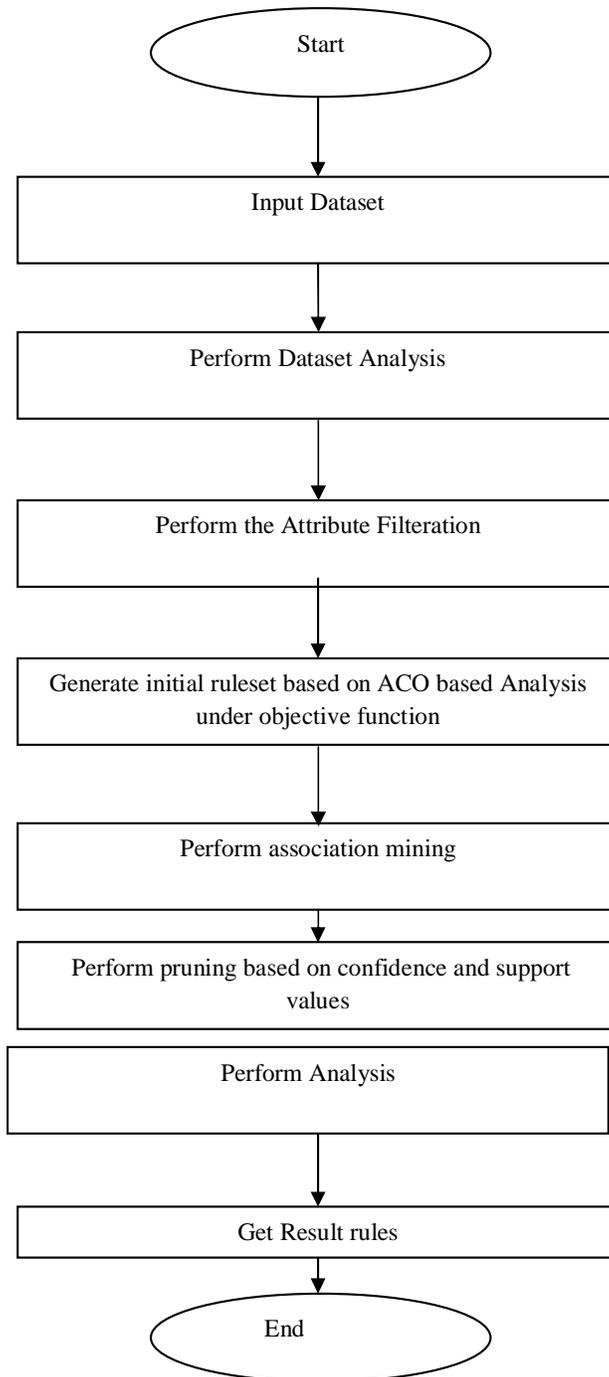
ANT COLONY OPTIMIZATION (ACO): Ant Colony Optimization (ACO) was introduced in the early 1990s.It is based on the foraging behavior of ants, which have the ability to select the shortest path among few possible paths connecting their nest to a food ground or site. A volatile chemical substance lay on the ground by the ants while walking and affecting in turn their moving decisions according to its local intensity is called pheromone which is the mediator of this behavior.
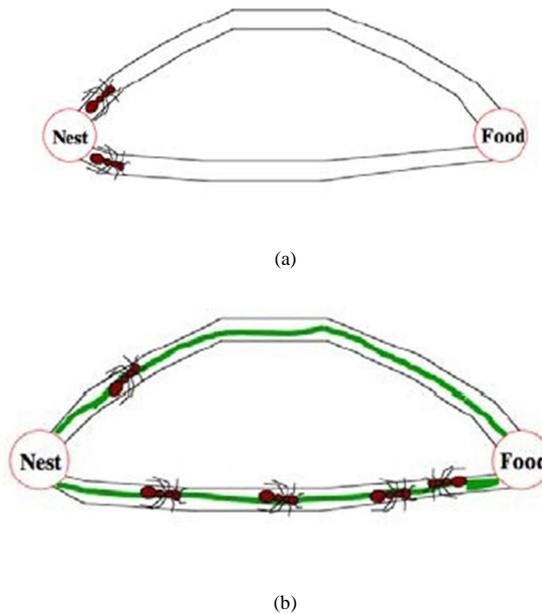
(a)

(b)

Figure: Principle of Ant Colony Optimization

As shown in figure at the beginning, no pheromone is put or place on the branches and the ants do not have any bit of information about the branches length. However, since one branch or path is shorter than the other, the shorter branch receives or accepts the chemical substance i.e pheromone at a higher rate than the longer branch . As ants can smell this pheromone, and their probabilistic decisions or conclusion are based in favor of paths or branch marked with higher amount of pheromone.

Data set analysis and Attribute Filteration :

In data dataset analysis ,the data is analysis on the basis of how the data is, what the data contains ,missing values etc.In attribute filteration, if most of the values in dataset are missing ,contains invalid data or some question mark values then these attributes are removed from data set.

REFERENCES

[1] Adepele Olukunle," *A Fast Algorithm for Mining Association Rules in Medical Image Data*", Proceedings of 2002 IEEE Canadian Conference on Electrical & Computer Engineering 0-7803-7514-9/02@ 2002 IEEE

[2] Carlos Ordonez," *Association Rule Discovery With the Train and Test Approach for Heart Disease Prediction*", IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 10, NO. 2, APRIL 2006, 1089-7771 © 2006 IEEE

[3] Chunxue Shi*," Path Planning for Deep Sea Mining Robot Based on ACO-PSO Hybrid Algorithm*", 2008 International Conference on Intelligent Computation Technology and Automation

[4] Gaurav N. Pradhan," *ASSOCIATION RULE MINING IN MULTIPLE, MULTIDIMENSIONAL TIME SERIES MEDICAL DATA*", ICME 2009 978-1-4244-4291-1/09©2009 IEEE

[5] Mr.K.Ravikumar*," ACO based spatial Data Mining   for Traffic Risk Analysis*".

[6] Wei Wang," *Mining Association rules in Medical Data Based on Concept Lattice*", Proceedings of the 8th World Congress on Intelligent Control and Automation July 6-9 2010, Jinan, China 978-1-4244-6712-9/10©2010 IEEE

[7] Mostafa Fathi Ganji," *Parallel Fuzzy Rule Learning Using an ACO-Based Algorithm for Medical Data Mining*", 978-1-4244-6439-5/10©2010 IEEE

[8] Pooia Lalbakhsh," *Focusing on Rule Quality and Pheromone Evaporation to Improve ACO Rule Mining*", 2011 IEEE Symposium on Computers & Informatics 978-1-61284-691-0/11©2011 IEEE

[9] Ghada Almodaifer," *Discovering Medical Association Rules from Medical Datasets*", 978-1- 61284-704-7/11 ©2011 IEEE

[10] Qiaoling Duan," *Mining Indirect Association Rules in Multi-database*", 2012 3rd International Conference on System Science, Engineering Design and Manufacturing Informatization 978-1-4673-0915-8/12©2012 IEEE

[11] P. Kasemthaweesab," *Association Analysis of Diabetes Mellitus (DM) With Complication States Based on Association Rules"*, 978-1-4577-2119-9/12@ 2011 IEEE

[12] Divya Bhugra," *Association Rule Analysis Using Biogeography Based Optimization*", 2013 International Conference on Computer Communication and Informatics (ICCCI -2013), Jan. 09 – 11, 2013, Coimbatore, INDIA 978-1-4673-2907-1/13 ©2013 IEEE

[13] K.Rameshkumar*,"Relevant Association Rule Mining from Medical Dataset Using Irrelevant Rule Elimination Technique"*