International Journal of Current Engineering and Technology
ISSN 2277 - 4106

Available at http://inpressco.com/category/ijcet

Review Article

Optimization of Association Rule Mining Process using Apriori and Ant Colony Optimization Algorithm

Shruti Aggarwal[a] and Babita Rani[a*]

[a]SGGSWU, Fatehgarh Sahib

Accepted 31 May 2013, Available online 1June 2013, **Vol.3, No.2 (June 2013)**

Abstract

Association Rule Mining is the essential part of data mining process. Apriori Algorithm is the popular algorithm of association rule mining. Apriori Algorithm that generates all significant association rules between items in the database. On the basis of the association rule mining and Apriori algorithm, an improved algorithm based on the Ant Colony Optimization algorithm will be proposed. We can optimize the result generated by Apriori algorithm using Ant Colony Optimization Algorithm by introducing Probabilistic Scheme. The algorithm improves result produces by Apriori algorithm. Ant Colony Optimization (ACO) is a metaheuristic inspired by the real behavior of ant colonies. In our research we will try to reduce the scanning of the databases by optimizing the frequent dataset scheme by Apriori. We will try to prune the weakest dataset rules only by fetching good rule set from neglected rules. Based on the threshold value, we will try to produce better association rule set.

Keywords*: Data Mining, Association Rule, Apriori Algorithm, ACO*

1. Introduction

Data Mining is a logical process that is used to search the relevant data from the large amounts of information or data. The goal of this technique is to find patterns that were previously unknown. Once you have found these patterns, you can use them to solve a number of problems. Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information (Arabinda Nanda *et al*, 2010).

The construction of a data warehouse, which involves data cleaning and data integration, can be viewed as an important pre-processing step for data mining. Data mining uses the data warehouse as the source of information for knowledge data discovery (KDD) (Joyce Jackson *et al*, 2002).

The abundance of data, coupled with the need for powerful data analysis tools, has been described as a *data rich but information poor* situation.

2. Association Rules

Association Rule Analysis is the task of discovering association rules that occur frequently in a given data set. A typical example of association rule mining application is the market basket analysis. In this process, the behaviour of the customers is studied when buying different products in a shopping store. The discovery of interesting patterns in this collection of data can lead to important marketing and management strategic decisions. For instance, if a customer buys bread, what is the probability that he/she buys milk as well? Depending on the probability of such an association, marketing personnel can develop better planning of the shelf space in the store or can base their discount strategies on such associations/correlations found in the data (Maria-Luiza Antonie et al.).

In general, the association rule is an expression of the form X=>Y, where X is antecedent and Y is consequent. Association rule shows how many times Y has occurred if X has already occurred depending on the support and confidence value. Many algorithms for generating association rules were presented over time. Some well-known algorithms are Apriori, FP-Growth, Ant Colony Optimization and Genetic algorithm (K.Vanitha et al, 2011).

Each association rule has two quality measurements, *Support(S)* and *Confidence(C)* (Yan et al., 2009).
Support of a rule $A \Rightarrow B$ is the probability of the itemset {A, B}. This gives an idea of how often the rule is relevant:

$$\text{Support}\ (A \Rightarrow B) = P\ (\{A, B\}) \quad (1)$$

Confidence of a rule $A \Rightarrow B$ is the conditional probability of B given A. This gives a measure of how accurate the rule is.

Shruti Aggarwal is working as Asst. Prof. and **Babita Rani** is a Research Scholar, *Corresponding author: **Babita Rani**

Confidence ($A \Rightarrow B$) = P ($B|A$) = Support ({A, B}) / Support (A) (2)

2.1 Basic Concepts

Let $I = \{I1, I2, \ldots, I_m\}$ be a set of m distinct attributes, T be transaction that contains a set of items such that $T \subseteq I$, D be a database with different transaction records Ts. An association rule is an implication in the form of $X \Rightarrow Y$, where $X, Y \subset I$ are sets of items called itemsets, and $X \cap Y = \emptyset$. X is called antecedent while Y is called consequent, the rule means X implies Y (Sotiris Kotsiantis et al, 2006).

In general, a set of items (such as the antecedent or the consequent of a rule) is called an itemset. The number of items in an itemset is called the length of an itemset. Itemsets of some length k are referred to as k-itemsets. Generally, an association rules mining algorithm contains the following steps:

1. The set of candidate k-itemsets is generated by 1-extensions of the large (k -1)- itemsets generated in the previous iteration.

2. Supports for the candidate k-itemsets are generated by a pass over the database.

3. Itemsets that do not have the minimum support are discarded and the remaining itemsets are called large k-itemsets.

This process is repeated until no more large itemsets are found (Sotiris Kotsiantis et al, 2006).

3. Apriori Algorithm

Apriori Algorithm is the originality algorithm of Boolean association rules of mining frequent item sets, developed by R. Agrawal and R. Srikant in 1994. Apriori Algorithm employs the bottom up, level-wise search method, it include all the frequent item sets (Jiao Yabing, 2013).

Apriori employs an iterative approach known as a level-wise search, where k-itemsets are used to explore (k+1)-itemsets. Apriori algorithm uses a breadth first search approach, first finding all frequent 1-itemsets and then discovering 2-itemsets and so on, until no more frequent K-itemsets can be found. To improve the efficiency of the level-wise generation of frequent itemsets, an important property called the Apriori property (All nonempty subsets of a frequent itemset must also be frequent) (Markus Hegland, 2005).

Apriori algorithm has two steps that are performed at each iteration, so this algorithm also called the iteration approach. The steps are:

a. Join Step: In this, join the itemsets with itself for generate the new itemsets. Itemsets are joinable if and only if there are one or more than one items are common.

For Example: There are three itemsets such as {1, 2}, {2, 3} and {3, 5}.

Table 3.1: Show the Candidate 2-itemsets

List Of Itemsets	Min_sup
{1,2}	1
{2,3}	3
{3,5}	4

Itemsets {1, 2} and {2, 3} are joinable because item 2 is common. Similarly itemsets {2, 3} and {3, 5} are joinable but itemsets {1, 2} and {3, 5} are not joinable because there are no item is common.

Table 3.2: Show the Candidate 2-itemsets

List Of Itemsets	Min_sup
{1,2}	1
{2,3}	3
{3,5}	4
{1,3}	2
{2,5}	2

b. Prune Step: If support of any itemset is less than minimum support then prune the items from the itemset and Apriori property is used for prune the items from the itemset.

For Example: Suppose minimum support (min_sup) = 2.

Table 3.3: Show the Candidate 2-itemsets

List Of Itemsets	Min_sup
{1,2}	1
{2,3}	3
{3,5}	4
{1,3}	2
{2,5}	2

Prune the {1, 2} itemset because its support value is less the min_sup.

4. Ant Colony Optimization

Ant Colony Optimization (ACO) (Vittorio Maniezzo et al.) is a paradigm for designing metaheuristic algorithms for combinatorial optimization problems. This algorithm was first proposed by M. Dorigo, 1992. Ant Colony Algorithm is a multi-agent approach for solving difficult combinatorial optimization problems like Traveling Salesman, vehicle routing, sequential ordering, graph

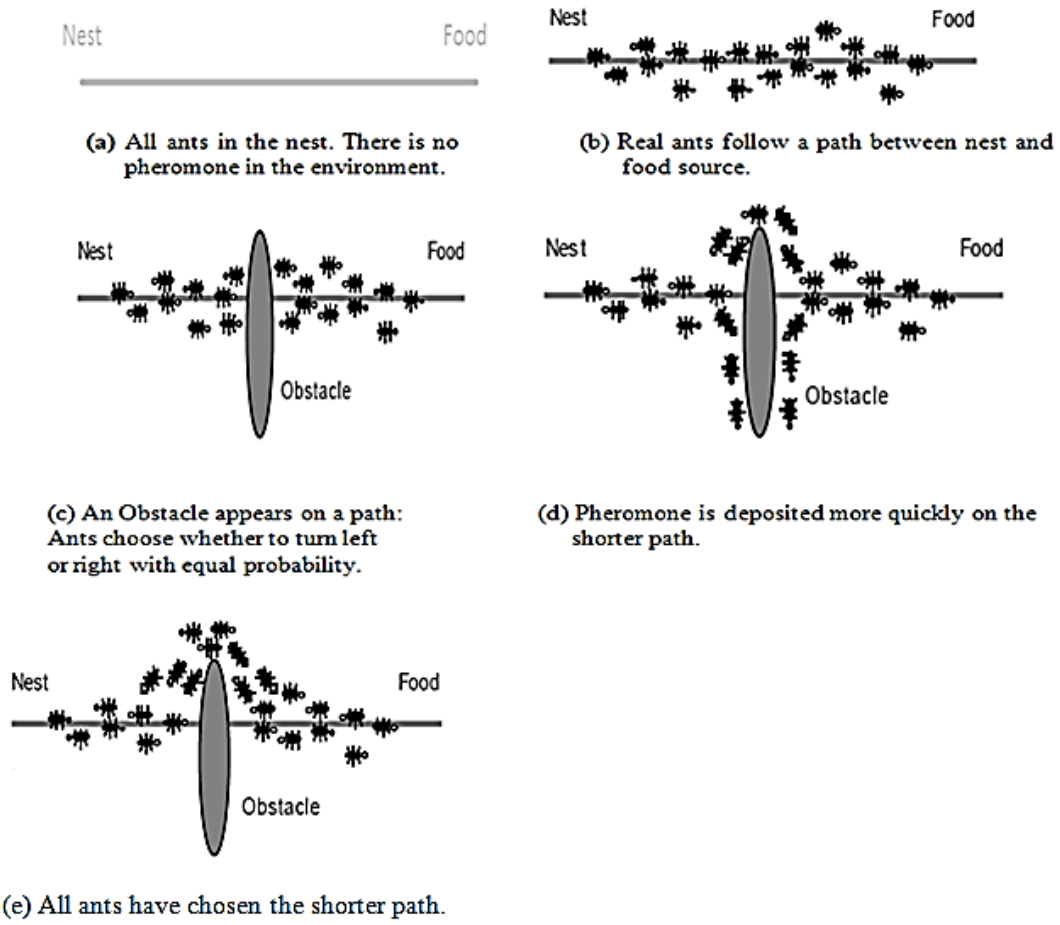

Fig 4.1 An experimental setting that demonstrates the shortest path finding capability of ant colonies. Between the ants' nest and the only food source exist two paths of different lengths.

coloring, routing in communications networks (Theresa Barker, 2005).

The ACO system contains two rules:
1. Local pheromone update rule, which applied whilst constructing solutions.
2. Global pheromone updating rule, which applied after all ants construct a solution.

Furthermore, an ACO algorithm includes two more mechanisms: trail evaporation and daemon actions. Trail evaporation decreases all trail values over time, in order to avoid unlimited accumulation of trails over some component. Daemon actions can be used to implement centralized actions which cannot be performed by single ants, such as the invocation of a local optimization procedure, or the update of global information to be used to decide whether to bias the search process from a non-local perspective (M. Dorigo et al, 2006)(M. Dorigo et al, 1999).

Ant Colony Optimization (ACO) is inspired by shortest path searching behavior of various ant species. Ants are social insects that live in colonies and because of their mutual interaction they are capable of performing difficult tasks. A very interesting aspect of behavior of ant species is their ability to find shortest path between the ants' nest and the food sources (Mr.Avdhesh Mann et al, 2012). Ants (blind) go through the food while laying down pheromone trails (Vittorio Maniezzo et al, 2004). Shortest path is discovered via pheromone trails

- Each ant moves at random
- Pheromone is deposited on path on which ant moves
- More the pheromone trails better the path (positive feedback sys)
- Ants follow the intense pheromone trails.

Example in Fig 4.1 Shows that ants moves from their nest to food.

5. Related Work

Badri Prasad Patel, Nitesh Gupta, Rajneesh K. Karn and Y.K. Rana show that the quality of rule sets from the Apriori algorithm for association rule mining can be improved by using Ant Colony Optimization (ACO) (Badri Prasad Patel et al, 2011). In this Authors used the Abalone dataset from UCI machine learning repository. The dataset has 4177 samples. It is composed of a discrete attribute and 8 continuous attributes. The improved

algorithm discovers frequent itemsets and shows much greater efficiency than the Apriori algorithm. Author's algorithm takes both efficiency and accuracy into account and it is proved and validated by experiment, so that they can mine association information from massive data better. The Apriori algorithm scans the database many times. When the database storing large number of data, time scanning the data is very long, so efficiency is very low. Increase the length of frequent itemsets, significant increase in computing time. The Apriori algorithm will produce overfull candidates of frequent itemsets, so the algorithm needs scan database frequently when finding frequent itemsets. And it will take more resource and time to accomplish one scanning. So it must be inefficient. Therefore authors have proposed an algorithm based on ACO to optimize the association rule generated by using Apriori algorithm.

Huan Wu, Zhigang Lu, Lin Pan, Rongsheng Xu, Wenbao Jiang in 2009 elaborates that because of the rapid growth in worldwide information, efficiency of association rules mining (ARM) has been concerned for several years. In this paper, based on the original Apriori algorithm, an improved algorithm Improved Apriori Algorithm (IAA) is proposed. IAA adopts a new count-based method to prune candidate itemsets and uses generation record to reduce total data scan amount. Experiments demonstrate that proposed algorithm outperforms the original Apriori and some other existing ARM methods (Huan Wu et al, 2009).

In this paper, an improved Apriori-based algorithm IAA is proposed. Through pruning candidate itemsets by a new count-based method and decreasing the mount of scan data by candidate generation record, this algorithm can reduce the redundant operation while generating frequent itemsets and association rules in the database. Validated by the experiments, the improvement is notable. This work is part of our Distributed Network Behavior Analysis System, though authors have considered C-R problem in proposed algorithm, for specific dataset, more work is still needed. They also need further research to implement this algorithm in their distributed system.

6. Proposed Work

This work presents an ACO algorithm for the specific problem of minimizing the number of association rules. Apriori algorithm uses a data set and uses a user interested support and confidence value then produces the association rule set. The rules mined through association rule mining algorithms are used for decision making. The good quality of rules mined help in better decision making. Association rule mining process needs to be optimizing so that good quality rules are mined.

This paper introduces probabilistic candidate concept for possibilities of finding better rule set in weak or prune probe rules. Quality of rule sets from the Apriori algorithm for association rule mining can be improved by using Improved Ant Colony Optimization. In our research we will try to reduce the number of scanning of database and will find the quality rule sets for association rule mining.

Conclusion

We have proposed the ACO algorithm for optimization association rule generated through Apriori algorithm. This work describes a method for the problem of association rule mining. An ant colony optimization (ACO) algorithm is proposed in order to minimize number of association rules. Our research will find better association rule sets. By implementing the Probabilistic section filtering, we can find more itemset for rule generation. Our Experiment will reduce dataset scanning and hence the accuracy in database processes.

References

Arabinda Nanda, Saroj Kumar Rout (10th may, 2010), Data Mining & Knowledge Discovery in Databases: An AI Perspective, *Proceedings of national Seminar on Future Trends in Data Mining (NSFTDM-2010)*, Organenised by Department of Computer Science, Gandhi Engineering college, Bhubaneswar.

Joyce Jackson (2002), Data Mining: A Conceptual Overviewdata Mining: A Conceptual Overview, *Communications of the Association for Information Systems*, Vol. 8, pp. 267-296.

Maria-Luiza Antonie, Osmar R. Zaiane, Mining Positive and Negative Association Rules:An Approach for Confined Rules, http:// webdocs.cs.ualberta.ca/~ zaiane/postscript/pkdd04.pdf.

K.Vanitha, R.Santhi (June 2011), Evaluating The Performance Of Association Rule Mining Algorithms, *Journal of Global Research in Computer Science*, Vol. 2, No. 6.

Yan et al (2009), Genetic Algorithm-Based Strategy For Identifying Association Rules Without Specifying Actual Minimum Support, *Available online at www.sciencedirect.com*, Expert Systems with Applications 36,3066–3076.

Sotiris Kotsiantis, Dimitris Kanellopoulos (2006), Association Rules Mining: A Recent Overview, *GESTS International Transactions on Computer Science and Engineering*, Vol.32 (1), pp. 71-82

Jiao Yabing (Jan 2013), Research of an Improved Apriori Algorithm in Data Mining Association Rules, *International Journal of Computer and Communication Engineering*, Vol. 2, No. 1.

Markus Hegland (Mar 30, 2005), The Apriori Algorithm a Tutorial, 9:7 WSPC/Lecture Notes Series: 9in x 6in heg05a.

Vittorio Maniezzo, Luca Maria Gambardella, Fabio de Luigi, Ant Colony Optimization.

Theresa Barker (2005), Meggie von Haartman, Ant Colony Optimization, *IEEE 516 Spring.*

M. Dorigo, M. Birattari, and T. Stitzle (Nov. 2006), Ant Colony Optimization: Arificial Ants as a Computational Intelligence Technique, *IEEE computational intelligence magazine*.

M. Dorigo and G. Di Caro (1999), The Ant Colony Optimization meta-heuristic, *in New Ideas in Optimization*, D. Corne et al., Eds., McGraw Hill, London, UK, pp. 11-32.

Mr.Avdhesh Mann, Dr.Rajneesh Talwar, Dr.Bharat Bhushan , Mr.Rakesh Gupta (July-2012), A Review Of Ant Colony Optimization, ISSN: 2249-9482, IJESS Vol. 2, Issue7.

Vittorio Maniezzo, Luca Maria Gambardella, Fabio de Luigi (2004), Ant Colony Optimization, http://www.idsia.ch/~luca/aco2004.pdf.

Rafael S. Parpinelli, Heitor S. Lopes, and Alex A. Freitas, Data Mining with an Ant Colony Optimization Algorithm.

Badri Prasad Patel, Nitesh Gupta, Rajneesh K. Karn and Y.K. (2011), Optimization of Association Rules Mining Apriori Algorithm Based on ACO, *International Journal on Emerging Technologies* 2(1): 87-92, ISSN: 0975-8364.

Huan Wu, Zhigang Lu, Lin Pan, Rongsheng Xu, Wenbao Jiang (2009), An Improved Apriori-based Algorithm for Association Rules Mining, *Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, IEEE Society community.