International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-3 Number-2 Issue-10 June-2013

Apriori and Ant Colony Optimization of Association Rules

Anshuman Singh Sadh[1], Nitin Shukla[2]

Abstract

Association Rule mining is one of the important and most popular data mining technique. Association rule mining can be efficiently used in any decision making processor decision based rule generation. In this paper we present an efficient mining based optimization techniques for rule generation. By using apriori algorithm we find the positive and negative association rules. Then we apply ant colony optimization algorithm (ACO) for optimizing the association rules. Our results show the effectiveness of our approach.

Keywords

Association Rule Mining, Positive Association, Negative Association, ACO.

1. Introduction

Data Mining is expected to relieve current mining methods from the sequential bottleneck, and provide the ability to scale to massive data sets and improve the response time [1].Data mining has been a powerful technique in analyzing and utilizing data in today's information-rich society. However, privacy is nowadays a major concern in data mining applications, which has led to a new research area, privacy preserving data mining. A large amount of research work has been devoted to this area, and resulted in such techniques as k-anonymity [2], data perturbation [3], [4], [5], [6], and data mining based on [7], [8].

Association Rule Mining (ARM) is one of the most used research are in data mining. ARM can be used for discovering hidden relationship between items. By given a user-specified threshold, also known as minimum support, the mining of association rules can discover the complete set of frequent patterns. That is, once the minimum support is given, the complete set of frequent patterns is determined [9]. In order to retrieve more correlations among items, users may specify a relatively lower minimum support [9].

Anshuman Singh Sadh, M.Tech Scholar, SRIT, Jabalpur, M.P.
Nitin Shukla, Assistant Professor, SRIT, Jabalpur, M.P.

Such a lower support often generates a huge amount of frequent patterns; but most of the patterns are already known or not interested to users. It is a tedious task for users to filter out these valueless patterns.

ARM is also studied in terms of market basket analysis, which is the analysis of the itemset which can be analyzed after the customer purchasing in the mall [9]. It is just like the analysis of the customer of purchasing behavior. Association rules also used in various areas such as telecommunication networks, market, risk management and inventory control etc.[9][10].

In [11] author suggests that Data mining [1] is used everywhere and large amounts of information are gathered: in business, to analyses client behavior or optimize production and sales [2].This signifies the research direction in several fields. We can use ARM and data mining application in health care, medical database, classification and combining these techniques with other approach extensively increases the potential behavior and applicability.

In [12] author suggests that many of the researchers are generally focused on finding the positive rules only but they not find the negative association rules. But it is also important in analysis of intelligent data. It works in the opposite manner of positive rule finding. 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 [12]. So better optimization technique can find a better solution in the above direction.

We provide here the overview of ARM. Other sections are arranged in the following manner: Section 2 introduces literature review; Section 3 discusses about proposed work; Section 4 describes the result analysis; section 5 shows the conclusion.

2. Literature Review

In 2010 Ashutosh Dubey et al. [13] proposed a novel data mining algorithm named J2ME-based Mobile Progressive Pattern Mine (J2MPP-Mine) for effective mobile computing. In J2MPP-Mine, they first

propose a subset finder strategy named Subset-Finder (S-Finder) to find the possible subsets for prune. Then, they propose a Subset pruner algorithm (SBPruner) for determining the frequent pattern. Furthermore, they proposed the novel prediction strategy to determine the superset and remove the subset which generates a less number of sets due to different filtering pruning strategy. Finally, through the simulation their proposed methods were shown to Deliver excellent performance in terms of efficiency, accuracy and applicability under various system conditions. Means if optimization is achieved the data mining is also easily supported on mobile devices.

In 2012, Nikhil Jain et al. [12] discuss about Association rule mining. They suggest that association rule 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. They 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.

In 2012, Ashutosh Dubey et al. [14] Proposes an efficient method for knowledge discovery which is based on subset and superset approach. In this approach they also use dynamic minimum support so that we reduce the execution time. A frequent superset means it contains more transactions then the minimum support. It utilize the concept that if the item set is not frequent but the superset may be frequent which is consider for the further data mining task. By this approach we can also find improved association, which shows that which item set is most acceptable association with others. A frequent subset means it contains less transactions then the minimum support. It utilizes the behavior that the less count may be frequent if we attached the less count with the higher order set. Here we also provide the flexibility to find multiple minimum supports which is useful for comparison with associated items and dynamic support range. Their algorithm provides the flexibility for improved association and dynamic support. Comparative result shows the effectiveness of their algorithm.

In 2012, Preeti Khare et al. [15] discusses that importance of data mining is increasing exponentially since last decade and in recent time where there is very tough competition in the market where the quality of information and information on time play a very crucial role in decision making of policy has attracted a great deal of attention in the information industry and in society as a whole. They use density minimum support so that they reduce the execution time. A frequent superset means it contains more transactions then the minimum support. It utilize the concept that if the item set is not frequent but the superset may be frequent which is consider for the further data mining task. By this approach they can store the transaction on the daily basis, then they provide three different density zone based on the transaction and minimum support which is low(L), Medium(M),High(H). Based on this approach they categorize the item set for pruning. Their approach is based on apriori algorithm but provides better reduction in time because of the prior separation in the data, which is useful for selecting according to the density wise distribution in India.

In 2012, Leena A Deshpande et al. [16] discusses about Semi-structured data which are a huge amount of complex and heterogeneous data sets. Such models capture data that are not intentionally structured, but are structured heterogeneously. These databases evolve so quickly like run time report generated by ERPs, World-Wide Web with its HTML pages, text files, bibliographies, various logs generated etc. These huge and varied become difficult to retrieve relevant information User is often interested in integrating various formats (like in biomedical data text, image or structured) that are generally realized as files, and also wants to access them in an integrated fashion. Users not only query the data to find a particular piece of information, but he is also keen in knowing better understanding of the query. Because of this variety, semi-structured DBs do not come with a conceptual schema.

In 2012, Smruti Rekha Das et al. [17] discusses about Support vector machine (SVM) which has become an increasingly popular tool for machine learning tasks involving classification, regression or novelty detection. SVM is able to calculate the maximum margin (separating hyper-plane) between data with and without the outcome of interest if they are linearly separable. To improve the generalization performance of SVM classifier optimization technique is used. According to the authors Optimization refers to the selection of a best element

from some set of available alternatives. Particle swarm optimization (PSO) is a population based stochastic optimization technique where the potential solutions, called particles, fly through the problem space by following the current optimum particles. They used Principal Component Analysis (PCA) for reducing features of breast cancer, lung cancer and heart disease data sets and an empirical comparison of kernel selection using PSO for SVM is used to achieve better performance. This paper focused on SVM trained using linear, polynomial and radial basis function (RBF) kernels and applying PSO to each kernels for each data set to get better accuracy.

In 2012, Sanat Jain et al. [18] present an Apriori-based algorithm that is able to find all valid positive and negative association rules in a support confidence framework. The algorithm can find all valid association rules quickly and overcome some limitations of the previous mining methods. The complexity and large size of rules generated after mining have motivated researchers and practitioners to optimize the rule, for analysis purpose. Their optimization done using Genetic Algorithm.

In 2011,Huang Qiu-yong et al. [19] present an Apriori's optimization algorithm. The algorithm first uses the order character of itemsets to reduce the times of comparison and connection when it connects and generates the candidate item sets, then compresses the candidate item sets according to the following situation: whether the number of element "a" in the frequent K-item sets is less than K. It improves the efficiency of mining association rules.

In 2013, Anshuman Singh Sadh et al. [20] survey several aspects of optimization techniques by which they can optimize the association rules. So the main motivation of their survey is to find the ways to minimize the rule generation.

3. Proposed Work

In our approach we use different data set for analysis. For finding positive and negative association rules we use apriori algorithm [21].

Apriori Algorithm [21]
Assumptions

- **Itemset:** a set of items
- **k-itemset:** an itemset which consists of k items
- **Frequent itemset (i.e. large itemset):** an itemset with sufficient support
- **L_k or F_k:** a set of large (frequent) k-itemsets
- **c_k:** a set of candidate k-itemsets
- **Appriori property:** if an item X is joined with item Y,

Support(X U Y) = min(Support(X), Support(Y))

//Each iteration i consists of two phases:
1. candidate generation phase:
//Construct a candidate set of large itemsets, i.e. find all the items that could qualify for further consideration by examining only candidates in set L_{i-1}*L_{i-1}
2. candidate counting and selection
//Count the number of occurrences of each candidate itemset
//Determine large itemsets based on predetermined support, i.e. select only candidates with sufficient support.

Set L_k is defined as the set containing the frequent k itemsets which satisfy
Support > threshold.

L_k*L_k is defined as:
L_k*L_k = {X U Y, where X, Y belong to L_k and| X $\cap$Y| = k-1}.
find all frequent itemsets

```
Appriori(database D of transactions, min_support) {
    F1 = {frequent 1-itemsets}
    k = 2
    while Fk-1 ≠ EmptySet
        Ck= AprioriGeneration(Fk-1)
        for each transaction t in the database D{
            Ct= subset(Ck, t)
            for each candidate c in Ct {
                count c ++
                }
        Fk = {c in Ck such that countc ≥ min_support}
        k++
    }
    F = U k ≥ 1 Fk
    }
```

After finding the positive and negative rules we then apply the ant colony optimization algorithm. The Ant Colony Optimization algorithm is mainly inspired by the experiments run by Goss et al. [22] which using a grouping of real ants in the real environment. They study and observe the behavior of those real ants and suggest that the real ants were able to select the shortest path between their nest and food resource, in the existence of alternate paths between the two.This ant behavior was first formulated and arranged as Ant

System (AS) by Dorigo et al. [23][24]. Based on the AS algorithm, the Ant Colony Optimization (ACO) algorithm was proposed [25]. In ACO algorithm, the optimization problem can be expressed as a formulated graph G = (C; L), where C is the set of components of the problem, and L is the set of possible connections or transitions among the elements of C.

ACO Algorithm [23][24]

1. Each ant searches for a minimum cost feasible partial solution.
2. An ant k has a memory Mk that it can use to store information on the path it followed so far. The stored information can be used to build feasible solutions, evaluate solutions and retrace the path backward.
3. An ant k can be assigned a start state sks and more than one termination conditions ek.
4. Ants start from a start state and move to feasible neighbor states, building the solution in an incremental way. The procedure stops when at least one termination condition ek for ant k is satisfied.
5. An ant k located in node i can move to node j chosen in a feasible neighborhood Nki through probabilistic decision rules. This can be formulated as follows:

An ant k in state sr =< sr-1; i > can move to any node j in its feasible neighborhood Nki , defined as Nki = {j | (j Є Ni) Λ (< sr, j >Є S)} sr Є S, with S is a set of all states.

6. A probabilistic rule is a function of the following.

a) The values stored in a node local data structure Ai = [aij] called ant routing table obtained from pheromone trails and heuristic values,

b) The ant's own memory from previous iteration, and

c) The problem constraints.

7. When moving from node i to neighbour node j, the ant can update the pheromone trails τij on the edge (i, j).
8. Once it has built a solution, an ant can retrace the same path backward, update the pheromone trails and die.

4. Result Analysis

For showing the effectiveness of the above algorithm, we taken the below example of table 1.

Table 1: Sample Database

S1							
Transaction	**A**	**B**	**C**	**D**	**E**	**F**	**G**
D1	0	1	0	0	0	0	0
D2	0	1	1	0	0	0	0
D3	0	1	1	1	0	0	0
D4	1	1	1	1	0	0	0
D5	1	0	0	1	0	0	0
D6	1	0	0	0	0	0	0
D7	1	0	0	0	1	1	1

Then we apply Apriori algorithm. In each step candidate itemsets are generated and counted on-the-fly as the database is scanned and we find the support on each step.

Support shows the frequency of the patterns in the rule; it is the percentage of transactions that contain both of the transactions.

Support = Probability(A and B)
Support = (# of transactions involving A and B) / (total number of transactions)

The frequent item sets in L1 are shown below. The first column lists the item, the second lists the support of the item. Item A in the first row, for example, appears in 4 transactions in Table 1 (transactions D4, D5, D6 and D7). This represents 57.14% of the 7 total transactions.

Table 2: Frequent Item set L1

S1_L1	
Itemset	**Percentage**
A	57.14
B	57.14
C	42.86
D	42.86
E	14.29
F	14.29
G	14.29

The frequent item sets in L2 are shown below. The first column lists the item; the second lists the support of the item. Item AB in the first row, for example, appears in 1 transaction in Table 1 (transactions D4). This represents 14.29% of the 7 total transactions.

Table 3: Frequent Item set L2

S1_L2	
Itemset	**Percentage**
AB	14.29
AC	14.29
AD	28.57
AE	14.29

AF	14.29
AG	14.29
BC	42.86
BD	28.57
BE	0
BF	0
BG	0
CD	28.57
CE	0
CF	0
CG	0
DE	0
DF	0
DG	0
EF	14.29
EG	14.29
FG	14.29

The frequent item sets in L3 are shown below. The first column lists the item; the second lists the support of the item. Item ABC in the first row, for example, appears in 1 transaction in Table 1 (transactions D4). This represents 14.29% of the 7 total transactions.

Table 4: Frequent Item set L3

S1_L3	
Itemset	**Percentage**
ABC	14.29
ABD	14.29
ABE	0
ABF	0
ABG	0
ACD	14.29
ACE	0
ACF	0
ACG	0
ADE	0
ADF	0
ADG	0
AEF	14.29
AEG	14.29
AFG	14.29
BCD	28.57
BCE	0
BCF	0
BCG	0
BDE	0
BDF	0
BDG	0
BEF	0
BEG	0
BFG	0
CDE	0
CDF	0
CDG	0
CEF	0
CEG	0
CFG	0
DEF	0
DEG	0
DFG	0
EFG	14.29

The frequent item sets in L4 are shown below. The first column lists the item; the second lists the support of the item. Item ABCD in the first row, for example, appears in 1 transaction in Table 1 (transactions D4). This represents 14.29% of the 7 total transactions.

Table 5: Frequent Item set L4

S1_L4	
Itemset	**Percentage**
ABCD	14.29
ABCE	0
ABCF	0
ABCG	0
ABDE	0
ABDF	0
ABDG	0
ABEF	0
ABEG	0
ABFG	0
ACDE	0
ACDF	0
ACDG	0
ACEF	0
ACEG	0
ACFG	0
ADEF	0
ADEG	0
ADFG	0
AEFG	14.29
BCDE	0
BCDF	0
BCDG	0
BCEF	0
BCEG	0
BCFG	0
BDEF	0
BDEG	0
BDFG	0
BEFG	0

CDEF	0
CDEG	0
CDFG	0
CEFG	0
DEFG	0

Table 6: Frequent Item set L5

S1_L5	
Itemset	**Percentage**
ABCDE	0
ABCDF	0
ABCDG	0
ABCEF	0
ABCEG	0
ABCFG	0
ABDEF	0
ABDEG	0
ABDFG	0
ABEFG	0
ACDEF	0
ACDEG	0
ACDFG	0
ACEFG	0
ADEFG	0
BCDEF	0
BCDEG	0
BCDFG	0
BCEFG	0
BDEFG	0
CDEFG	0

Table 7: Frequent Item set L6

S1_L6	
Itemset	**Percentage**
ABCDEF	0
ABCDEG	0
ABCDFG	0
ABCEFG	0
ABDEFG	0
ACDEFG	0
BCDEFG	0

Table 8: Frequent Item set L7

S1_L7	
Itemset	**Percentage**
ABCDEFF	0

Then we consider only those sets which are frequent. For the above example we consider the minimum support value as 50%. So only those set which are greater than or equal to 50 % are in positive category and all other in the negative category. This phenomenon is shown in figure 1. So we clearly understand that only A and B are in positive set and rest of the items is in negative set.

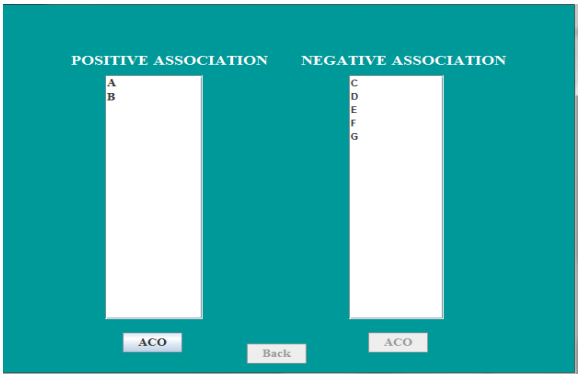

Figure 1: Positive and Negative Association

The percentage of A and B before applying optimization is 57.14, when we apply optimization we achieve the optimization of 94 % as shown in table 9. This phenomenon is also shown in figure 2 and figure 3.

Table 9: ACO on Positive Rules

PACO	
Column	**value**
A	0.94
B	0.94

The percentage of C, D, E, F and G before applying optimization is 42.86, 42.86, 14.29, 14.29 and 14.29, when we apply optimization we achieve the optimization as shown in table 10. This phenomenon is also shown in figure 4 and figure 5.

Table 10: ACO on Negative Rules

NACO	
Column	**value**
C	0.4286
D	0.4286
E	0.5714
F	0.5714
G	0.5714

By the below results we can proof that by applying optimization can provide better optimization association rules.

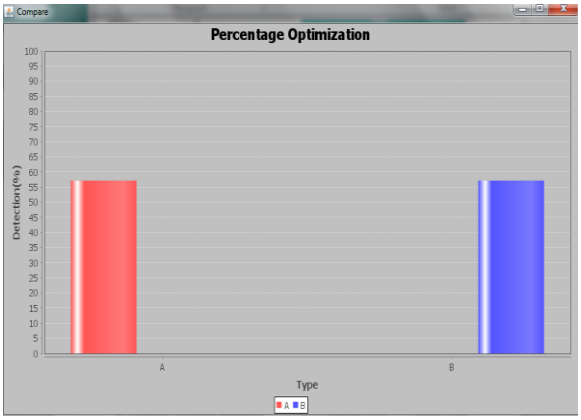

Figure 2: Percentage Optimization before ACO (Positive Rules)

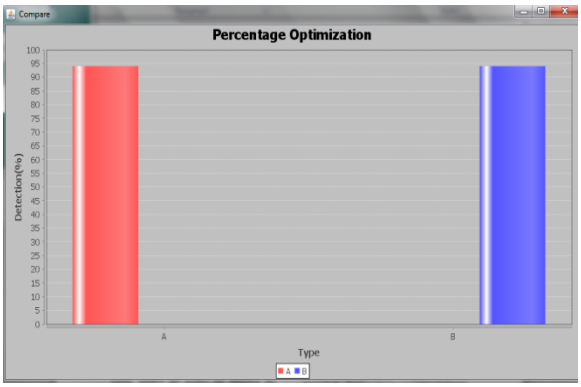

Figure 3: Percentage Optimization after ACO (Positive Rules)

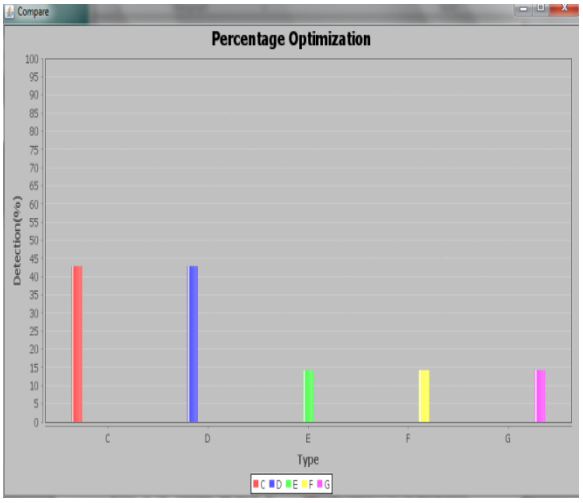

Figure 4: Percentage Optimization before ACO (Negative Rules)

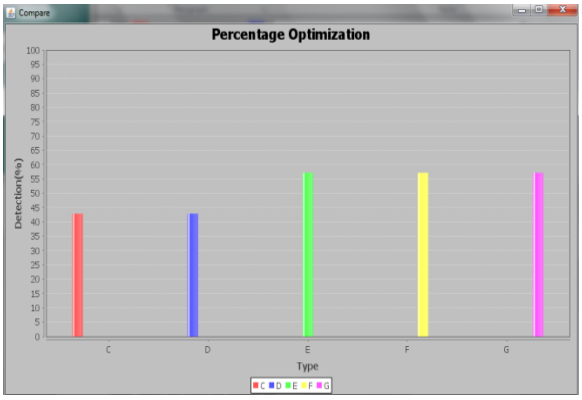

Figure 5: Percentage Optimization after ACO (Negative Rules)

5. Conclusions

In this paper we apply apriori and ant colony optimization technique to achieve the positive and negative association rule optimization. By our approach we achieve better optimization as we discuss in the result section.

References

[1] J.Z.Mohammed.Parallel and distributed data mining: an introduction. M.J.Zaki, C.-T.Ho (Eds.) Large-scale parallel data mining, Lecture Notes in Artificial Intelligence, 1759:1-23, 2000.

[2] Sweeney, L. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):557-570, 2002.

[3] Agrawal, R. and Srikand R. Privacy preserving data mining. In Proc. Of ACM SIGMOD Conference, pp. 439-450, 2000.

[4] Chen, K. and Liu. L. A random rotation perturbation approach to privacy data classification. In Proc of IEEE Intl. Conf. on Data Mining (ICDM), pp. 589-592, 2005.

[5] Xu, S., Zhang, J., Han, D. and Wang J. Singular value decomposition based data distortion strategy for privacy distortion. Knowledge and Information System, 10(3):383-397, 2006.

[6] Mukherjeee, S., Chen, Z. and Gangopadhyay, A. A privacy-preserving technique for Euclidean distance-based mining algorithms using Fourier related transforms. Journal of VLDB, 15(4):293-315, 2006.

[7] Vaidya, J. and Clifton, C. Privacy preserving k-means clustering over vertically partitioned data. In Prof. of ACM SIGKDD Conference, pp.206-215, 2003.

[8] Vaidya, J., Yu, H. and Jiang, X. Privacy preserving SVM classification. Knowledge and Information Systems, 14:161-178, 2007.
[9] Ms. Kumudbala Saxena, Dr. C.S. Satsangi, "A Non-Candidate Subset-Superset Dynamic Minimum Support Approach for sequential pattern Mining", International Journal of Advanced Computer Research (IJACR), Volume-2, Number-4, Issue-6, December-2012.
[10] Dr. Manish Shrivastava, Mr. Kapil Sharma, MR. Angad Singh, "Web Log Mining using Improved Version of Proposed Algorithm", International Journal of Advanced Computer Research (IJACR),Volume 1, Number 2, December 2011.
[11] Pragati Shrivastava, Hitesh Gupta," A Review of Density-Based clustering in Spatial Data", International Journal of Advanced Computer Research (IJACR), Volume-2, Number-3, Issue-5, September-2012.
[12] Nikhil Jain,Vishal Sharma,Mahesh Malviya, "Reduction of Negative and Positive Association Rule Mining and Maintain Superiority of Rule Using Modified Genetic Algorithm", International Journal of Advanced Computer Research (IJACR) ,Volume-2, Number-4, Issue-6, December-2012.
[13] Ashutosh K. Dubey and Shishir K. Shandilya," A Novel J2ME Service for Mining Incremental Patterns in Mobile Computing",Communications in Computer and Information Science, 2010,Springer LNCS.
[14] Ashutosh Kumar Dubey, Animesh Kumar Dubey, Vipul Agarwal, Yogeshver Khandagre, "Knowledge Discovery with a Subset-Superset Approach for Mining Heterogeneous Data with Dynamic Support",Conseg-2012.
[15] Preeti Khare, Hitesh Gupta, "Finding Frequent Pattern with Transaction and Occurrences based on Density Minimum Support Distribution", International Journal of Advanced Computer Research (IJACR), Volume-2, Number-3, Issue-5, September-2012.
[16] Leena A Deshpande, R.S. Prasad, "Efficient Frequent Pattern Mining Techniques of Semi Structured data: a Survey", International Journal of Advanced Computer Research (IJACR) Volume-3, Number-1, Issue-8, March-2013.
[17] Smruti Rekha Das, Pradeepta Kumar Panigrahi, Kaberi Das and Debahuti Mishra, "Improving RBF Kernel Function of Support Vector Machine using Particle Swarm Optimization", International Journal of Advanced Computer Research (IJACR) Volume-2, Number-4, Issue-7, December-2012.
[18] Sanat Jain, Swati Kabra, "Mining & Optimization of Association Rules Using Effective Algorithm", International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 4, April 2012.
[19] Huang Qiu-yong, Tang Ai-long and Sun Zi-guang, "Optimization Algorithm of Association Rule Mining Based on Reducing the Time of Generating Candidate Itemset", IEEE 2011.
[20] Anshuman Singh Sadh, Nitin Shukla," Association Rules Optimization: A Survey", International Journal of Advanced Computer Research (IJACR), Volume-3, Number-1, Issue-9, March-2013.
[21] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. Research Report RJ 9839, IBM Almaden Research Center, San Jose, California, June 1994.
[22] S. Goss, S. Aron, J. L. Deneubourg, and J. M. Pasteels. Self-organized Shorcuts in the Argentine Ant. Naturwissenschaften, 76:579–581, 1989.
[23] M. Dorigo, Gianni Di Caro, and Luca M. Gambardella. Ant Algorithms for Discrete Optimization. Technical Report Tech. Rep. IRIDIA/98-10, IRIDIA, Universite Libre de Bruxelles, Brussels, Belgium, 1998.
[24] M. Dorigo and M. Maniezzo and A. Colorni. The Ant Systems: An Autocatalytic Optimizing Process. Revised 91-016, Dept. of Electronica, Milan Polytechnic, 1991.
[25] M. Dorigo and G. Di Caro. New Ideas in Optimisation. McGraw Hill, London, UK, 1999.

I completed my MCA and M.Phil (Computer Science). Currently I am pursuing M.Tech (CTA) from Shri Ram Group of Institutes, Jabalpur. I having experience of 07 Year in Teaching. My interest areas are data mining, optimization, cloud computing and object oriented programming.