Prediction of Customer Behavior using evolutionary Associative Clustering

J. Arunadevi*
Dr. V. Rajamani**

(With Special Reference to Mobile Phone Consumers with Location Reference)

Abstract

Spatial Association Rule Mining (SAR) is an interesting area of the spatial data mining which involves several steps and complexity. We have introduced a two step algorithm in which the first step concentrates on the optimization of SAR using the Hybrid evolutionary algorithm which uses Genetic algorithm and Ant Colony Optimization (ACO). Since Association rule with multiple objectives can be considered as the NP hard problem we are using the Multi objective genetic algorithm and the ACO. The results are appreciable when compared to the existing ones.

In the second step we try to cluster the generated association rules and that can be used for the target group segmentation. We have studied the Customer behaviour of the mobile phone industry based on their location.

Keywords : SAR, MOGA, ACO, Clustering, Segmentation.

1. Introduction

Market is characterized by being global; products are identical and enormous supply. This leads to the customer centric market rather than product centric market. Because of the size of the customers mass marketing is expensive and the returns are not assuring. It leads to the research on the targeted customers. The customers to be targeted can be identified using the model for predicting the customer behavior.

Customer profiling is describing customers by their attributes. This can be used to prospect new customers or to drop out existing bad customers [1]. Customer profiling forms a base for the marketers to market with the existing loyal customers and offer them better services and retain them. This can be achieved by manipulating the collected information. Depending on the need of the hour one has to decide which profile will be beneficial at that time. We can use the specialized data mining technique such Spatial data mining for achieving the customer behaviors based on the spatial attributes.

Spatial Association Rule mining (SAR) is about generating association rules about spatial data objects. Either the antecedent or the consequent of the rule must contain some spatial predicates (such as near) [2]. Spatial association rules are

*Assistant Professor, Dept of MCA, Thiagarajar School of Management, Madurai, Tamilnadu, India
**Principal, Indra Ganesan College of Engineering, Tirchirapalli, Tamilnadu, India

implications of one set of data by another such as the average monthly family income in Madurai for families living near Annanagar is Rs. 100, 000. Due to the relationships involved the spatial components; one entity can affect the behavior of other entity. Spatial data items are naturally linked to neighboring data elements (e.g., contiguous geographic positions), these data elements are not statistically independent. This makes the spatial data mining different from the normal transactional data mining.

The various activities involved in the SAR is computing the spatial relationships, generating the frequent sets and extracting the association rules. In this paper we are concentrating on the second and third step for the SAR. The existing approaches use quantitative reasoning, which computes distance relationships during the frequent set generation [3][4]. These approaches deal only with points, consider only quantitative relationships and do not consider non spatial attributes of geographic data, which may be fundamental importance of knowledge discovery. Qualitative spatial reasoning [5][6][7] considers distance and topological relationships between a reference geographic object type and a set of relevant feature types represented by any geometric primitive (e.g. points, lines, and polygons). [8] uses qualitative spatial reasoning approach with prior knowledge and removes well known patterns completely by early pruning the input space and the frequent item sets.

We present a novel two step refinement algorithm based on hybrid evolutionary algorithm (HEA) which uses genetic algorithm with ant colony optimization for generating the spatial association rules and clustering the generated rules for the required groups. In the first step HEA algorithm is used to enhance the performance of Multi objective genetic algorithm (MOGA) by incorporating local search with Ant colony optimization (ACO), for Multi objective association rule mining. In the proposed HEA algorithm, MOGA is conducted to provide the diversity of associations thereafter; ant colony optimization is performed to come out of local optima. From the experiment results, it is shown that the proposed HEA algorithm has superior performance when compared to other existing algorithms. In the second step we group the rules generated for finding the various target groups by clustering. Rules are grouped based on consequent information of the rules generated by Step 1. Groups of rules are in the form Xi -> Y for i=1,2,…,n. That is, different rule antecedents Xi's are collected into one group for a same rule consequent Y.

The paper is organized as follows; Section 2 deals with the concepts of SAR and their interesting measures, Section 3 deals with MOGA and the ACO applied for the optimization of the rule generation, method of clustering the rules is discussed in the Section 4. Section 5 deals with approach followed in this paper. Section 6 discusses the results obtained and the Section 7 gives the conclusion of the paper.

2. Spatial Association Rule Mining

A spatial association rule is of the form X ->Y , between two disjoint item sets, where X is called antecedent and Y is the consequent of the rule. The antecedent contains a set of predicates from the exploring database, the consequent only represents one predicate, which is not yet included in the antecedent. The rule itself then reflects an existing relationship between predicates in antecedent and consequent. The association rule generated is generally measured by the two metrics called support and the confidence. Support is defined as the ratio between number of transactions that contains both X and Y to the total number of transactions. Confidence is the ratio of the number of transactions with all the items to the number of transactions with just the "if" items. Another metric used is the Lift (improvement) tells us how much better a rule is at predicting the result than just assuming the result in the first place. It is defined as the ratio of the records that support the entire rule to the number that would be expected, assuming there was no relationship between the items. Spatial association rules represent object/predicate relationships containing spatial predicates. For example, the following rules are spatial association rules.

- Nonspatial consequent with spatial antecedent(s)
  is_a(x,house) ^ close_to(x,beach) Is_expensive(x)
- Spatail consequent with non-spatial /spatial antecedent(s).
  is_a(x,gas_station) close_to(x,highway).

Various kinds of spatial predicates can be involved in spatial association rules [9].

3. Optimizing The Rule Generation using Evolutionary Computation Techniques

Existing algorithms for the SAR try to measure the quality of generated rule by considering only one evaluation criterion, but because of the growing need of the knowledge from the spatial data we can consider the problem as a Multi objective one rather than the single objective. Multi-objective optimization deals with solving optimization problems which involve multiple objectives. Most real-world search and optimization problems involve multiple objectives (such as minimizing fabrication cost and maximize product reliability and others) and should be ideally formulated and solved as a multi-objective optimization problem.

Over the past decade, population-based evolutionary algorithms (EAs) (genetic algorithms (GAs) and evolution strategies (ESs) have been found to be quite useful in solving multi-objective optimization problems, simply because of their ability to find multiple optimal solutions in a single simulation run. In general the main motivation for using Genetic Algorithms in the discovery of high-level prediction rules is that they perform a global search and cope better with attribute interaction than the greedy rule induction algorithms often used in data mining.[10].

Genetic algorithms for rule discovery can be divided into two broad approaches, the Michigan approach and the Pittsburgh approach [11]. The biggest distinguishing feature between the two is that in the Michigan approach (also referred to as Learning Classifier Systems) an individual is a single rule, whereas in the Pittsburgh approach each individual represents an entire set of rules [12]. In this paper we follow the fist approach ie the Michigan approach for the SAR.

The MOGA is used to achieve the multi objective by with a Pareto based multiple-objective genetic algorithm. The possible rules are represented as chromosomes and a suitable encoding/decoding scheme has been defined, it also provides

the diversity of associations among the rules generated by elitism. To increase the efficiency of the MOGA we are using the ACO, which limits the algorithm from falling to the local optimal solution.

ACO is a paradigm for designing meta heuristic algorithms for combinatorial optimization problems. The ACO algorithm was first introduced by Colorni, Dorigo and Maniezzo [13] [14] and the first Ant System (AS) was proposed by Dorigo in his Ph.D. thesis [15]. The ACO is a meta-heuristic algorithm, which utilizes the inspiration from real ant colonies behaviours to find a shortest path from a food source to the nest without using visual cues by exploiting pheromone information [16] [17] [18]. When ant colonies are seeking for food, they leave a kind of chemical compositions, which is called pheromone. The more ants walk through the path, the more pheromone left on the ground. Then, the next ant will choose one path with a probability proportional to the amount of pheromone. Finally this positive feedback process will construct a shortest path from their nest to the food source.

The characteristic of ACO algorithms is their explicit use of elements of previous solutions.

Edge Selection:
An ant will move from node i to node j with probability

$$p_{i,j} = \frac{(\tau_{i,j}^{\alpha})(\eta_{i,j}^{\beta})}{\sum(\tau_{i,j}^{\alpha})(\eta_{i,j}^{\beta})}$$

where
τi,j is the amount of pheromone on edge i,j
α is a parameter to control the influence of τi,j
ηi,j is the desirability of edge i,j (a priori knowledge, typically 1 / di,j)
β is a parameter to control the influence of ηi,j

Pheromone Update
τi,j = (1 - ϱ)τi,j + Δτi,j
where
τi,j is the amount of pheromone on a given edge i,j
ϱ is the rate of pheromone evaporation
and Δτi,j is the amount of pheromone deposited, typically given by

$$\Delta\tau_{i,j}^{k} = \begin{cases} 1/L_k & \text{if ant } k \text{ travels on edge } i,j \\ 0 & \text{otherwise} \end{cases}$$

where Lk is the cost of the kth ant's tour (typically length).

4. Clustering The Rules

Clustering association rules is one of the meaningful ways of grouping association rules into different clusters. When the Spatial Association rules are generated in order to identify the group of targets we are using the clustering approach. In [19], the authors selected highly ranked (based on confidence) association rules one by one and formed cluster of objects covered by each rule until all the objects in the database are covered. The authors of [20] formed cluster of rules of the form Xi -> Y, that is, rules with different antecedent but with same consequent Y and they extracted representative rules for each cluster as knowledge for the cluster. In [21], the authors formed cluster of rules based on structure distance of antecedent. The authors of [22] formed hierarchical clustering of rules based on different distance methods used for rules. In [23], the authors discussed different ways of pruning redundant rules including rule cover method. All Associative Classifier (AC) CBA, CMAR[24], RMR[25], and MCAR[26] generate cluster of rules called class-association rule (CAR) with class label as same consequent and they use database (rule) cover to select potential rules to build (AC) classifier model. In most of the ARM work, confidence measure is used to rank association rules. Also, other measures such as chi-square, laplace-accuracy is used to select highly ranked rules.

In this paper we are using the classifier model which uses the consequent information for grouping. The clusters will be formed who are having their consequent as similar pattern. We have first grouped based on the attributes; it may be homogeneous like urban core, suburbs, rural or Hierarchical groups like Metropolitan area, major cities, and neighborhoods. Then this is further grouped based on the purpose like segmenting the population by consumer behavior. We have used the algorithm proposed in [27].

The clustering algorithm groups of rules are in the form Xi -> Y for i=1,2,…,n. That is, different rule antecedents Xi's are collected into one group for a same rule consequent Y. next step is to select small set of representative rules from each group. Representative rules are selected based on rule instance cover as follows.

Let Ry={ Xi -> Y | i=1,2,…,n } be a set of n rules for some item-set Y and m(Xi Y) be rule cover, which is the set of tuples/records covered by the rule Xi -> Y in the dataset D.

Let Cy be the cluster rule cover for a group or cluster of rules Ry. i.e.,
Cy = m(Ry) = U i=1,2,…n m(X i Y)
from cluster rule set Ry, find a small set of k rules ry called representative rule set such that m(ry) is almost equal to m(Ry). i.e.,
m(ry) ≈ m(Ry), or
U i=1,2,…k m(X i Y) ≈ U i=1,2,…n m(X i Y), where k<< n

To find representative rule set ry from Ry, we use the rule cover algorithm proposed in [20].

5. Application of Hea for Spatial Association Rule Mining

The procedures of HEA are as follows. First, MOGA searches the solution space and generates association lists to provide the initial population for ACO. Next, ACO is executed, when ACO terminates, the crossover and mutation operations of MOGA generate new population. ACO and GA search alternately and cooperatively in the solution space. Then the rules are clustered using the rule cover based on the consequent information.

Step 1: Pseudo code for optimization of rule generation

1. while (t <= no_of_gen)
2. M_Selection(Population(t))
3. ACO_MetaHeuristic
   while(not_termination)

generateSolutions()
pheromoneUpdate()
daemonActions()
end while
end ACO_MetaHeuristic

4. M_Recombination_and_Mutation(Population(t))
5. Evaluate Population(t) in each objective.
6. t = t+1
7. end while
8. Decode the individuals obtained from the population with high fitness function.

Step 2: Pseudo code for clustering the rules generated

Input : set of rules generated by the HEA Ry={ Xi -> Y | i=1,2,…,n } and the rule cover.

1. Generate the cluster rule cover
2. count = number of records in the cluster cover
3. while(no of records in the cluster cover > 2% of count)
   Sort all the rules in the Ry in the descending order of the rule cover.
   Take the first rule r with highest rule cover
   If the no of records in the rule cover is <= 2% of count
   Exit while loop

End if.

4. ry = ry U r
5. Delete the highest rule cover from the cluster cover
6. End While

Output : the representative rule set.

The representative rule set is used for the segmentation of the consequent.

6. Results and Discussions

We have used the synthesized dataset for our research. The area of study is Madurai City.

Fig 1 : Madurai City

Data has been collected in and around the city of Madurai. The main aim of the data collection is about the Mobile phone users based on their service providers, Mode of usage and the amount of recharge done by the customers on the location basis. The general procedure of data mining is: question raise $\rightarrow$ data preparation (including data selection, data pretreatment and data transformation) $\rightarrow$ data arrangement $\rightarrow$ model building / data mining $\rightarrow$ result evaluation and explanation. Data preparation is the key which determines the success of data mining. The process of spatial data is much more complex[28]. After preprocessing we have transformed the spatial data in term of .xls file. We have implemented the basis of the apriori algorithm of association rule, we programmed to complete the calculation in virtue of M-language in Matlab. The specific procedure is as following.

(1) Take advantage of "import wizard" in Matlab to accomplish the import of data file. Until now, the data fields and character fields are saved separately. For example, the default uses a matrix named "data_num" to keep numerical fields and a matrix named "textdata" to keep character fields.
(2) Run algorithm step 1to generate the rules.
(3) Run algorithm step 2 to generate the target group using Java.

Keeping the confidence as 50% we have computed the results. In fig 1 the comparison has been done for the number of rules generated to the support count given with the Apriori algorithm, Apriori algorithm optimized with the MOGA and the Apriori algorithm optimized with HEA proposed in Step 1.

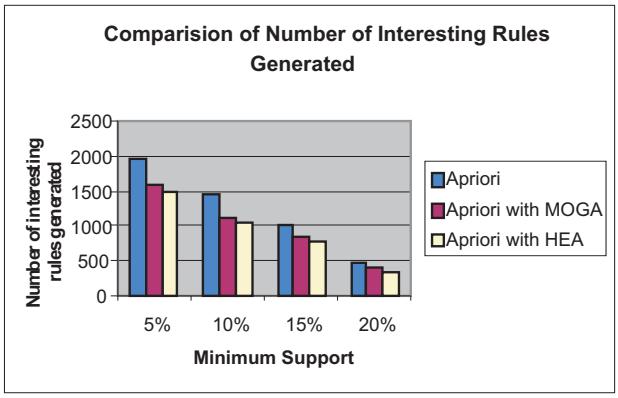

Fig 2 : Comparison of The Three Algorithms based on The Number of Rules Generated

From Fig 2 we can have the following observations

1. When the Support is increased the numbers of rules generated are decreasing and the use of HEA also performs a significant change in the number of rules generated.
2. HEA performance is close with the MOGA, but the application of the ACO reduces the number of needed rules generated .

In fig 3 the comparison has been done for the lift ratio for the top 500 rules generated to the support count given with the Apriori algorithm, Apriori algorithm optimized with the MOGA and the Apriori algorithm optimized with HEA proposed in Step 1.

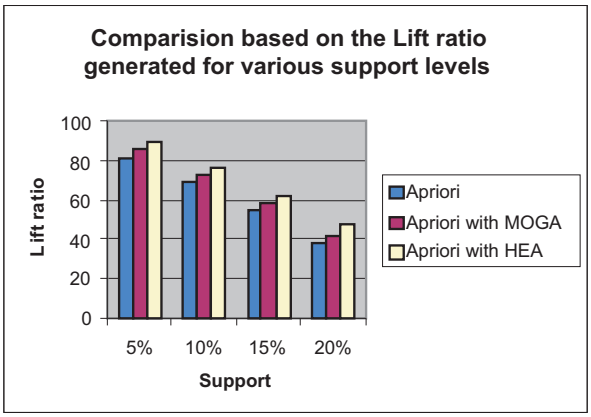

Fig 3 : Comparison of the Three Algorithms based on the Lift Ratio

Lift ratio says us how much better the rule is better as predicting the result than just assuming the result in the first place. It is defined as the ratio of the records that support the entire rule to the number that would be expected, assuming there was no relationship between the items. From Fig 2 we can have the following observation, Lift ratio for the HEA is better than the other two algorithms. This shows the efficiency of the HEA to identify the rules for predicting the result.

In fig 4 the comparison has been done for computational time for the support count given with the Apriori algorithm, Apriori algorithm optimized with the MOGA and the Apriori algorithm optimized with HEA proposed in Step 1.

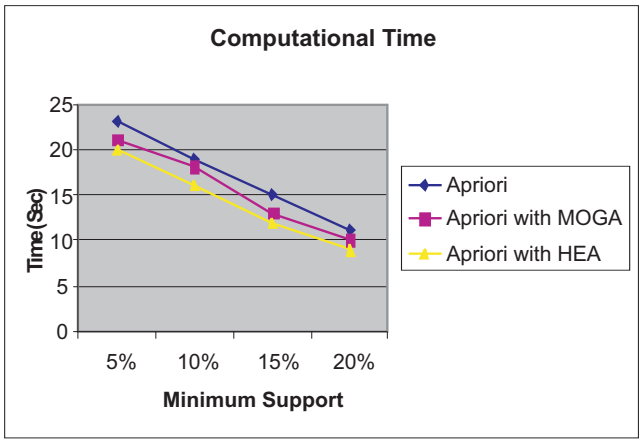

Fig 4 : Comparison of the Three Algorithms based on the Computational Time

By the refinement of the rules generated HEA algorithm in step 2 by the cluster concept is useful in narrowing the segmentation.

The segmentation has been done to find the popular Service provider in the various locations of Madurai, Mode of usage used and the amount of recharge done

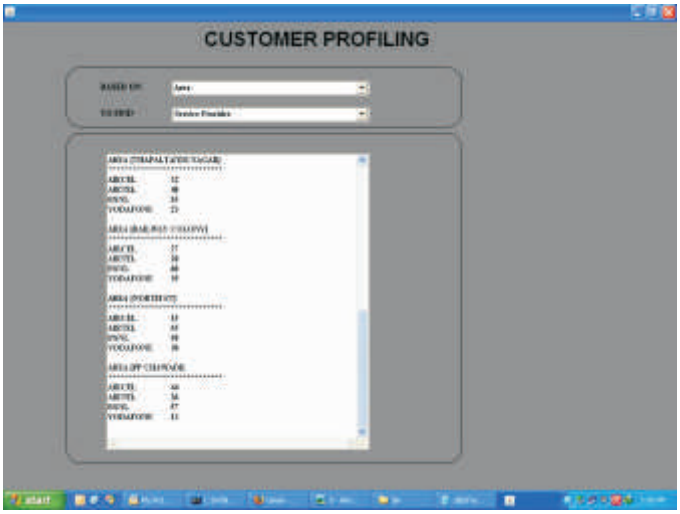

Fig 5 : Customer profiling based on Area and Service Provider

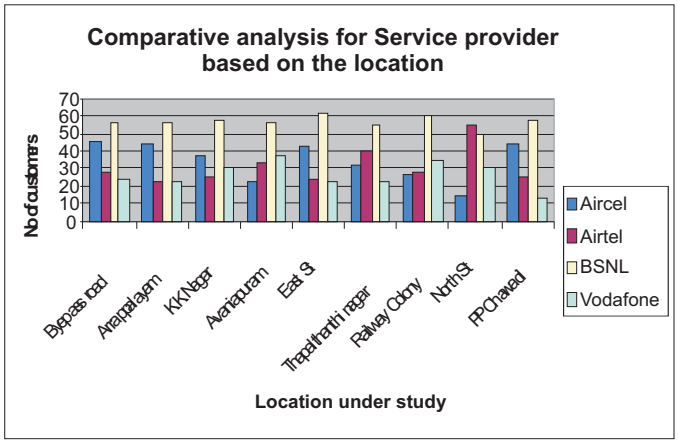

Fig 6 : Comparative analysis for Service Provider based on the Location

Based on the above analysis we can find the following facts.

1. The maximum usage is based on the BSNL service provider
2. In the north street area Airtel provides the maximum usage

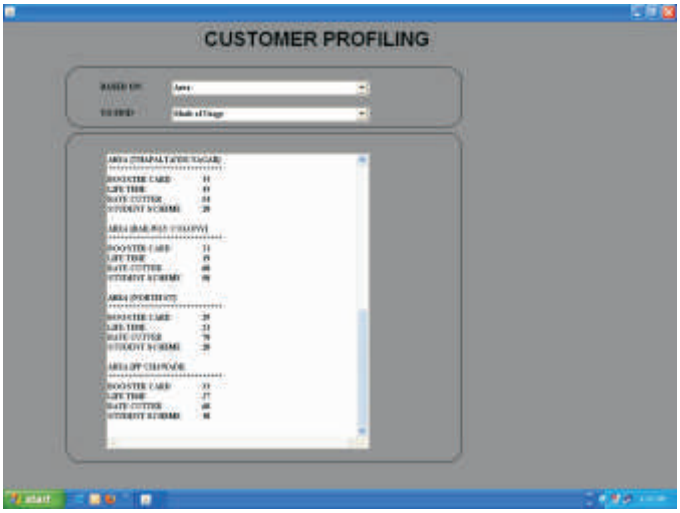

Fig 7 : Customer profiling based on area and mode of usage

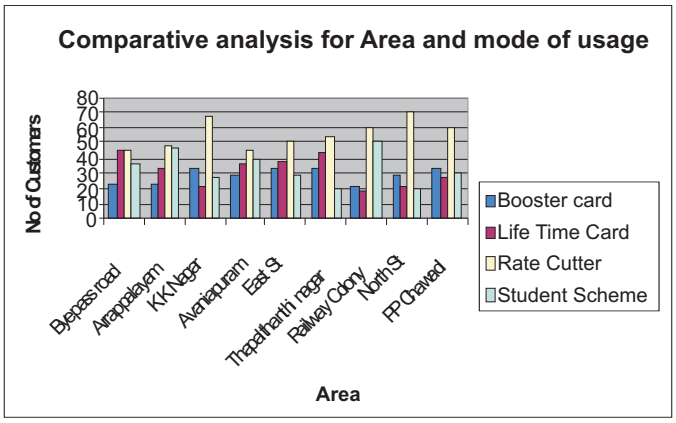

Fig 8 : Comparative analysis for area and mode of usage

Based on the above analysis we can find the following facts

1. Rate cutter is preferred by the users

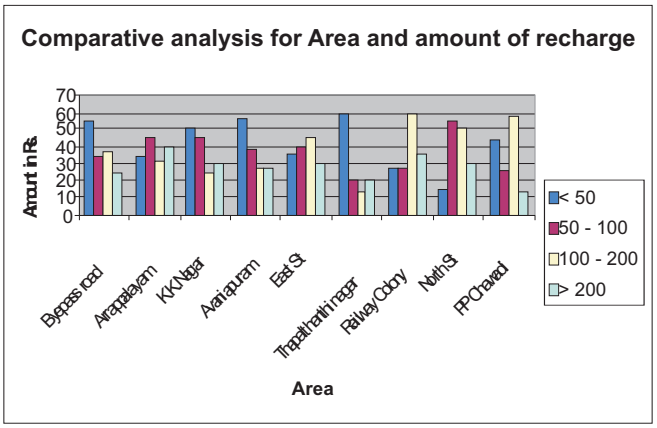

Fig 9 : Customer profiling based on area and amount of recharge

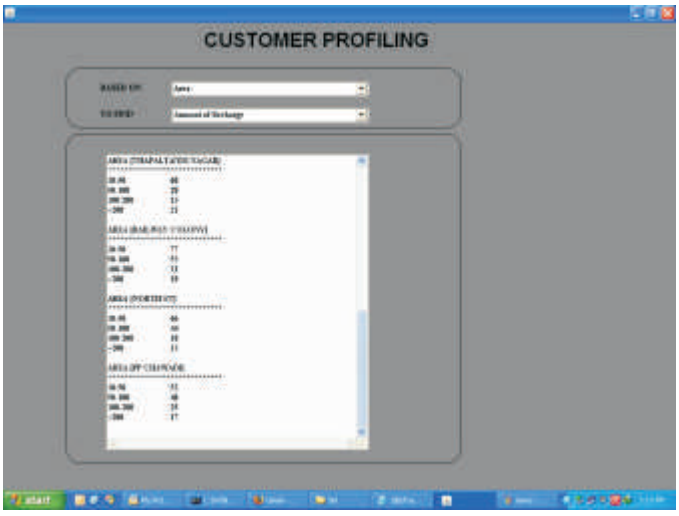

Fig 10 : comparative analysis for area and amount of usage

2. In three areas the amount of recharge is mostly around Rs. 100-200.

7. Conclusion

Prediction of Customer Behavior for the Mobile communication in the area of study gives us knowledge about the behavioral trends of the Customers. The usage of the Associative clustering algorithm optimized by the Hybrid evolutionary algorithm provides efficient usage of the given data to form the segmentations. HEA reduces the time for the prediction and increases the lift ratio of the rules generated. Clustering the Association rules gives the formation of the segmentation of the preferred customer behavior. In future we can use Classification over the Clustered Rules.

8. References

1. *K.Thearling, " Data Mining and Customer relationships", http://www3.shore.net/~kht/index.htm.*
2. *Margaret H.Dunham and S.Sridhar. "Data Mining Introductory and Advanced Topics", Pearson Education, 2006.*
3. *Yoo J.S., Shekhar S. and Celik M. (2005). "A Join-less Approach for Co-location Pattern Mining: A Summary of Results", In 5th IEEE-ICDM, Houston, 2005, p.813-816. IEEE Computer Society.*
4. *Yoo, J.S. and Shekhar S. (2004). "A partial join approach for mining co-location patterns". In 12th ACM-GIS, Washington, p.241-249, ACM Press.*
5. *Appice, M., Berardi, M., Ceci, M. and Malerba, D. (2005) "Mining and Filtering Multi-level Spatial Association Rules with ARES". In 15th ISMIS, New York, p.342-353. Springer.*
6. *Mennis, J. and Liu, J.W. (2005) Mining Association Rules in Spatio-Temporal Data: An Analysis of Urban Socioeconomic and Land Cover Change. Transactions in GIS, v9 (1), (January), p. 5-17.*
7. *Koperski, K. and Han, J. (1995) "Discovery of spatial association rules in geographic information databases," In 4th SSD, Portland, p. 47-66. Springer.*
8. *Bogorny, V. Enchancing spatial association rule mining in geographic databases. PhD Thesis, Instituto de Informatica da UFRGS, Brazil, October 2006.*
9. *www.dsi.unive.it/~dm/ssd95.pdf*
10. *Alex A. Freitas, "A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery" Postgraduate Program in Computer Science, Pontificia Universidade Catolica do Parana Rua Imaculada Conceicao, 1155. Curitiba - PR. 80215-901. Brazil*
11. *Dehuri, S., Jagadev, A. K., Ghosh A. And Mall R. 2006. Multi-objective Genetic Algorithm for Association Rule Mining Using a Homogeneous Dedicated Cluster of Workstations. American Journal of Applied Sciences 3 (11): 2086-2095, 2006 ISSN 1546-9239*
12. *Peter P. Wakabi-Waiswa and Venansius Baryamureeba. Extraction of Interesting Association Rules Using Genetic Algorithms. International Journal of Computing and ICT Research, Vol. 2, No. 1, pp. 26 – 33. http:www.ijcir.org/volume2-number1/article4.pdf.*
13. *A . Colorni, M. Dorigo, and V. Maniezzo. Positive feedback as a search strategy. Technical Report No. 91-016, Politecnico di Milano, Italy,1991.*
14. *A. Colorni, M. Dorigo, and V. Maniezzo . The ant system: an autocatatlytic process. Technical Report No. 91- 016, Politecnico*

di Milano, Italy, 1991.

15. *R. Beckers, J.L. Deneubourg and S. Goss. Trails and Uturns in the selection of the shortest path by the ant lasius niger. Journal of Theoretical Biology, 159, 1992, pp. 397- 415.*
16. *M. Dorigo. Optimization, Learning and Natural Algorithms. Ph.D. Thesis, Politecnico di Milano, Italy,1992*
17. *S. Goss, S. Aron, J.L. Deneubourg and J.M. Pasteels (). Self-organized shortcuts in the argentine ant. Natur wissenschaften, 76, 1989, pp. 579-581.*
18. *B. Holldobler and E.O. Wilson (1990). The Ants. Springer-Verlag: Berlin.*
19. *Waler A.Kosters, Elena Marchiori and Ard A.J.Oerlemans, "Mining Clusters with Association Rules", Advances in Intelligent Data Analysis (IDA-99) (D.J.Hand, J.N.Kok and M.R.Berthold, Eds.), Lecture Notes in Computer Science 1642, Springer, 1999, pp. 39-50.*
20. *H. Toivonen, M. Klemettinen, P. Ronkainen, K. Hatonen, and H. Mannila, "Pruning and grouping discovered association rules", In Proc. ECML-95 Workshop on Statistics, Machine Learning, and Knowledge Discovery in Database, April 1995,pp 47-52.*
21. *Pi Dechang and Qin Xiaolin, "A new Fuzzy Clustering algorithm on Association rules for knowledge management", Information Technology Journal 7(1), 2008, pp. 119-124.*
22. *Alipio Jorge, "Hierarchical Clustering for thematic browsing and summarization of large sets of Association Rules", In Proceedings of the 4th SIAM International Conference on Data Mining. Orlando, FL, 2004, pp. 178-187.*
23. *G. Li, and H.J.Hamilton, "Basic association rules", In Proceedings of the 4th SIAM International Conference on Data Mining, Orlando, FL, 2004, pp. 166–177.*
24. *W. Li, J. Han, and J. Pei, "CMAR: accurate and efficient classification based on multiple class-association rules", In: Proceedings of IEEE International Conference on Data Mining (ICDM2001), 2001. pp. 369–376.*
25. *A. Thabtah, and P. I. Cowling, "A greedy classification algorithm based on association rule", Appl. Soft Comput., Vol. 7, No. 3. June 2007, pp. 1102-1111.*
26. *Adriano Veloso, Wagner Meira, Marcos Gonçalves, and Mohammed Zaki, "Multi-label Lazy Associative Classification", Knowledge Discovery in Databases: PKDD 2007, 2007, pp. 605-612.*
27. *S.Kannan and R.Bhaskaran, "Association Rule Pruning based on Interestingness Measures with Clustering", IJCSI International Journal of Computer Science Issues, Vol. 6, No. 1, 2009, p. 35-43.*
28. *Xinqi Zheng, Lu Zhao, "Association Rule Analysis of Spatial Data Mining Based on Matlab", Workshop on Knowledge Discovery and Data Mining , 2008 IEEE DOI 10.1109/WKDD.2008.21*