A Two Step optimized Spatial Association Rule Mining Algorithm by Hybrid Evolutionary Algorithm and Cluster Segmentation

Dr. J.Arunadevi[1]

[1] Thiagarajar School of Management

Abstract

A novel two step approach by adopting hybrid evolutionary algorithm with cluster segmentation for Spatial Association Rule mining (SAR) is presented in this paper.Here first step concentrates on the optimization of SAR using the hybrid evolutionary algorithm which uses genetic algorithm and ant colony optimization (ACO). Multi objective genetic algorithm is used to provide the diversity of associations. ACO is performed to come out of local optima. In the second step, cluster the generated association rules used for the target group segmentation. Preferential based segmentation of the women of various groups belongs to the Madurai city, Tamilnadu, India. Here, number of rules generated by the first step of our SAR is minimized, also time generation for the rules are also minimized. Lift ratio increased for the generated rules.

Index terms— SAR, MOGA, ACO, clustering, segmentation.

1 Introduction

n association rule is a general form of dependence where an element has some dependence with another element. Margaret H.Dunham and S.Sridhar [1] said 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). Spatial association rules are implications of one set of data by another. 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.

Various activities involved in the SAR is computing the spatial relationships, generating the frequent sets and extracting the association rules. 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, Author ? : Assistant Professor, Dept of MCA, Thiagarajar School of Management, Madurai,Tamilnadu, India. Author ? : Principal, Indra Ganesan College of Engineering, Tirchirapalli,Tamilnadu, India. E-mail-arunasethupathy@gmail.com which may be fundamental importance of knowledge discovery. Qualitative spatial reasoning [5]- [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). Bogorny.V [8] used qualitative spatial reasoning approach with prior knowledge and removes well known patterns completely by early pruning the input space and the frequent item sets. Salvatore Orlando [9] discussed about the various kinds of spatial predicates which can be involved in spatial association rules. In this paper, we are concentrating on the second and third step for the SAR.

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 is developed and implemented. 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 simulation 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 based on genetic algorithm. The proposed methodology is to use genetic algorithm over the existing rule cover algorithm for keeping nearest neighbors in common cluster and to be insensitive to data input order. Rules are grouped based on consequent information of the rules generated by step 1. Groups of rules are in the form X i -> Y for i=1,2,?,n. That is, different rule antecedents X i 's are collected into one group for a same rule consequent Y.

Moses santhakumar said [29] it is the second largest city in Tamilnadu state, having a very old history of about two thousand six hundred years and is often referred to as the Athens of East. Having different locations for residential, commercial, industrial, educational, public and semi-public habitats, we have chosen Madurai as the place of study and collected data from various parts of it.

Aditi pai and Deepika Khatri [30] said women are now a critical consumer segment for marketers to tap, not only for household and conventional women's products but also for services. Yet all women do not behave the same, their habits differ according to their status such as marital and working status. In our paper we have taken women in city Madurai for the target segmentation.

The paper is organized as follows: Section 2 deals with the background concepts of SAR, MOGA and the ACO applied for the optimization of the rule generation, method of clustering the rules. Section 3 deals first step of the algorithm developed, Section 4 explains the second step of the algorithm. Section 5 deals with approach followed in this paper and discuss the results obtained and Section 6 gives the conclusion of the paper.

2 II.

3 Background Study

This section is divided into three parts. Section a.) discusses the need for considering SAR as a multi objective problem, section b.) discusses the use of genetic algorithm and the ACO for the optimizing SAR and section c.) says about the Clustering of association rules.

4 a) Multi objective view on SAR

Existing algorithms for the SAR try to measure the quality of generated rule by considering only one evaluation criterion. Due to 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.

5 b) Evolutionary algorithms for optimization

Over the past decade, population-based evolutionary algorithms (EAs) 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. Alex A. Freitas [10] explained the motivation for using genetic algorithms in the discovery of high-level prediction rules. GA perform a global search and cope better with attribute interaction than the greedy rule induction algorithms often used in data mining. According to Dehuri. S et al [12] genetic algorithms for rule discovery can be divided into two broad approaches, the Michigan approach and the Pittsburgh approach. Peter P. Wakabi-Waiswa and Venansius Baryamureeba [13] said the biggest distinguishing feature between the two is that in the mapping of chromosome to the rule.

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 discussed by Dorigo [15] in his Ph.D. thesis. 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]- [18].

In order to reduce the total number of rules generated we used evolutionary algorithm. We tend to use the evolutionary algorithms since the solution space and the number of rules generated will be exponentially growing. We use the HEA, which inturn is using MOGA and ACO.

6 c) Clustering of the association 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. Waler A.Kosters et al, [19] 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. Toivonen.H et al [20] formed cluster of rules of the form X i -> 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. Pi Dechang and Qin Xiaolin [21] formed cluster of rules based on structure distance of antecedent. Alipio Jorge [22] formed hierarchical clustering of rules based on different distance methods used for rules. G. Li, and H.J.Hamilton [23] discussed different ways of pruning redundant rules including rule cover method. All Associative Classifier (AC) CBA, CMAR proposed by W. Li et. al [24]. RMR proposed by A. Thabtah, and P. I. Cowling [25] and MCAR proposed by Adriano Veloso et. al [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 association rule mining, confidence measure is used to rank association rules. Also, other measures such as chi-square, laplace-accuracy is used to select highly ranked rules.

7 III.

Application of HEA for Spatial Association rule mining SAR uses apriori algorithm for the generation of the rules. Here the rules generated by apriori using the hybrid 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.We follow the Michigan approach for the optimization. To increase the efficiency of the MOGA, we are using the ACO, which limits the algorithm from falling to the local optimal solution.

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.

8 a) String Representation

Chromosomes are encoded as real numbers the number of genes in each chromosome is equal to the number of item sets considered. Each gene will have 4 digits for vector index. A sample chromosome may look like as follows: 0001 0102 0204 0302 0401 0500 0601 0702 0802 0901 1002 1101 1201

Here, the first two numbers in each gene represents the attribute and the next two denotes the value , fourth gene has the value 0302 where 03 refers to the age group and 02 refers to the third age group ranges from 23 to 25. Like wise all the gene has been encoded, once the initial population is generated now we are ready to apply genetic operators.

9 b) Fitness

The fitness function is calculated as the arithmetic weighted average confidence, comprehensibility and J-Measure. The fitness function is given by f(x) = (w1 * Comprehensibility) + (w2 * J -Measure) + (w3 * Confidence)
_____w1+w2+w3

where w1,w2,w3 are used defined weights.

10 d)

Reproduction (Selection)

The selection process selects chromosomes from the mating pool directed by the survival of the fittest concept of natural genetic systems. In the proportional selection strategy adopted in this paper, a chromosome is assigned a number of copies, which is proportional to its fitness in the population, go into the mating pool for further genetic operations. Roulette wheel selection is used for the proportional selection strategy.

11 e) Crossover

Crossover is a probabilistic process that exchanges information between two parent chromosomes for generating two child chromosomes. In this paper, single point crossover with a fixed crossover probability of C is used. For chromosomes of length l, a random integer, called the crossover point, is generated in the range [1, l -1]. The portions of the chromosomes lying to the right of the crossover point are exchanged to produce two offspring.

12 Mutation

Each chromosome undergoes mutation with a fixed probability M. For binary representation of chromosomes, a bit position (or gene) is mutated by simply flipping its value. Since we are considering real numbers in this paper, a random position is chosen in the chromosome and replace by a random number between 0-9.

Step 1: Pseudo code for optimization of rule generation 1. while (t <= no_of_gen) 2. M_Selection(Population(t))

13 Clustering the rules

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. Further group based on the purpose like segmenting the population by consumer behavior. S.Kannan and R.Bhaskaran [27] proposed an algorithm for clustering the rules.

The key factors to be considered in the spatial clustering algorithm is keeping nearest neighbors in common cluster and the cluster to be insensitive to the data input order. In our novel approach we achieve this objective by with a Pareto based multiple-objective genetic evolutionary algorithm.The MOGA is used to achieve the multi Let R y ={ X i -> Y | i=1,2,?,n } be a set of n rules for some item-set Y and m(X i Y) be rule cover, which is the set of tuples/records covered by the rule X i -> Y in the dataset D.

Let C y be the cluster rule cover for a group or cluster of rules R y . i.e., C y = m(R y) = U i=1,2,?n m(X i Y)

From cluster rule set Ry, find a small set of k rules r y called representative rule set such that m(r y) is almost equal to m(R y). i.e., m(r y) ? m(R y), or U i=1,2,?k m(X i Y) ?U i=1,2,?n m(X i Y), where k« n

Step 2 : Pseudo code for clustering the rules generated Input : set of rules generated by the HEA R y ={ X i -> Y | i=1,2,?,n } and the rule cover. Apply GA for rearranging the rules in various orders based on the fitness preferred by the user.

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 R y 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. r y = r y U r 5. Delete the highest rule cover from the cluster cover 6. End While Output : the representative rule set.

Apply GA for retaining nearest neighbours in common cluster.The optimized representative rule set is used for the segmentation of the consequent.

GA is applied at the first stage for the arrangement of the rules based on the fitness; this is to help the clustering for not suffering from the order of the input. In the second.

14 Results and discussions

We have used the synthesized dataset for our research. The general procedure of data mining is:

? Question raise ? Data preparation (including data selection, data pre treatment and data transformation) ? Data arrangement ? Model building/data mining ? Result evaluation and explanation.

The specific procedure is as following.

(1) As done by Xinqi Zheng and Lu Zhao [28] we 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.

(2) Run algorithm step 1 to generate the rules.

(3) Run algorithm step 2 to generate the target group.

The environmental setup are Population size = 100, Mutation rate (M) = 0.5, Crossover rate (C)=0.8, Elitism = 0.5. The stopping criterion used is the non evolution of the archive during 100 generations, once the minimal number of generations has been over passed.

Keeping the fitness 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. the nearest neighbors based on the metrics given by the user. Application of GA is to increase the quality of 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 fitness, this rectify the sensitiveness to the data input order. After applying the rule cover, GA is used to retain better compared to the other two algorithms. HEA outperforms MOGA by an average of 8% for the four support levels (5%,10%,15% and 20%). It shows the good improvement in the 20% support level.

In 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 by improving the information about the increase in probability of the consequent to the given antecedent part.

In By the refinement of the rules generated HEA algorithm in step 2 by the cluster concept is useful in narrowing the segmentation. From the synthetic dataset we have used the representative rules to find the preference based segmentation for the women based on 7 preferences and the distance they can maximum travel to get it. We have segmented the women into five segments and the representative rules have minimum support 20% fitness as 50% and the lift ratio as 38 %.

By the application of genetic algorithm for the keeping nearest neighbors in common cluster and to be insensitive to data input order, the time of generation of the clusters has been significantly reduced. This is depicted in Fig 4 and 1. On an average 11 % decrease in the time of forming the clusters is performed by the proposed approach. This is due to the reason of applying GA for the rearrangement of rules before clustering. 2.

The proposed approach works well when the number of rules applied for clustering is 1000 and the time reduction is about 16.6% 3. The average increase in the Dunn's Validity Index is 6.91% The students and the just married women are willing to relocate more for the preference. 2. Young mothers are limited to relocate 3. Most of the target group is willing to relocate less for the bus stop 4. The preference of the students and the just married women were alike.

15 Comparision based on the Lift ratio generated for various support levels

We can generate number of observations and this can be used to the planning of the city based on the preferences, marketing and the logistics can be planned.

16 VI.

17 Conclusion

This paper proposed a methodology for the optimization of SAR using the Hybrid Evolutionary Algorithm in the first phase and finding the target groups by the refinement of the rules by clustering. The results for the first phase is promising and the second phase also lay a opening for the identification of target groups which can be further extended to the preferential based mining. In this paper the case taken for the phase two is the women target groups with various categories and their preferences. The work can be extended to the classification of the rule generated and the phase two can be also optimized with the help of the evolutionary algorithms. [1] [2] [3]

Figure 1: Fig 1 :

[4] [5] [6] [7] [8]

Figure 2:

A Two Step Optimized Spatial Association Rule Mining Algorithm By Hybrid
Evolutionary Algorithm
And Cluster Segmentation

Figure 3:

[Thabtah and Cowling (2007)] 'A greedy classification algorithm based on association rule'. A Thabtah , P I Cowling . *Appl. Soft Comput* June 2007. 7 (3) p. .

[Yoo et al. ()] 'A Joinless Approach for Co-location Pattern Mining: A Summary of Results'. J S Yoo , S Shekhar , M Celik . *5th IEEE-ICDM*, (Houston) 2005. 2005. IEEE Computer Society. p. .

[Dechang and Xiaolin ()] 'A new Fuzzy Clustering algorithm on Association rules for knowledge management'. Pi Dechang , Qin Xiaolin . *Information Technology Journal* 2008. 7 (1) p. .

[Yoo and Shekhar ()] 'A partial join approach for mining co-location patterns'. J S Yoo , S Shekhar . *12th ACM-GIS*, (Washington) 2004. ACM Press. p. .

[Agarwal and Srikant ()] R Agarwal , R Srikant . *fast Algorithms for mining association rules in large databases, International conference on very large databases, VLDB, 20*, (San Fransico. Proceedings? California) 1994. 1994. Morgan Kaufmann. p. .

[Zheng and Zhao ()] 'Association Rule Analysis of Spatial Data Mining Based on Matlab'. Xinqi Zheng , Lu Zhao . IEEEDOI10.1109/WKDD.2008.21. *Workshop on Knowledge Discovery and Data Mining*, 2008.

[Kannan and Bhaskaran ()] 'Association Rule Pruning based on Interestingness Measures with Clustering'. S Kannan , R Bhaskaran . *IJCSI International Journal of Computer Science Issues* 2009. 6 (1) p. .

[Li and Hamilton ()] 'Basic association rules'. G Li , H J Hamilton . *Proceedings of the 4th SIAM International Conference on Data Mining*, (the 4th SIAM International Conference on Data MiningOrlando, FL) 2004. p. .

[Li et al. ()] 'CMAR: accurate and efficient classification based on multiple classassociation rules'. W Li , J Han , J Pei . *Proceedings of IEEE International Conference on Data Mining (ICDM2001)*, (IEEE International Conference on Data Mining (ICDM2001)) 2001. p. .

[Margaret et al. ()] *Data Mining Introductory and Advanced Topics*, H Margaret , S Dunham , Sridhar . 2006. Pearson Education.

[Koperski and Han ()] 'Discovery of spatial association rules in geographic information databases'. K Koperski , J Han . *4th SSD*, (Portland) 1995. Springer. p. .

[Bogorny (2006)] 'Enchancing spatial association rule mining in geographic databases'. V Bogorny . *Instituto de Informatica da UFRGS* October 2006. (PhD Thesis)

[Peter et al.] 'Extraction of Interesting Association Rules Using Genetic Algorithms'. P Peter , Venansius Wakabi-Waiswa , Baryamureeba . `http:www.ijcir.org/volume2-number1/article4.pdf` *International Journal of Computing and ICT Research* 2 (1) p. .

[Jorge ()] 'Hierarchical Clustering for thematic Rules'. Alipio Jorge . *Advances in Intelligent Data Analysis*, Lecture Notes in Computer Science J Ida-99) (d, J N Hand, M R Kok, Berthold (ed.) 1999. Springer. 1642 p. .

[Appice et al. ()] 'Mining and Filtering Multi-level Spatial Association Rules with ARES'. M Appice , M Berardi , M Ceci , D Malerba . *15th ISMIS*, (New York) 2005. Springer. p. .

[Mennis and Liu (2005)] 'Mining Association Rules in Spatio-Temporal Data: An Analysis of Urban Socioeconomic and Land Cover Change'. J Mennis , J W Liu . *Transactions in GIS* 2005. January. 9 (1) p. .

[Waler et al. ()] 'Mining Clusters with Association browsing and summarization of large sets of Association Rules'. A Waler , Elena Kosters , Marchiori , A J Ard , Oerlemans . *Proceedings of the 4th SIAM International Conference on Data Mining*, (the 4th SIAM International Conference on Data MiningOrlando, FL) 2004. p. .

[Veloso et al. ()] *Multi-label Lazy Associative Classification*, Adriano Veloso , Wagner Meira , Marcos Gonçalves , Mohammed Zaki . 2007. 2007. p. . (Knowledge Discovery in Databases: PKDD)

[Dehuri et al. ()] 'Multi-objective Genetic Algorithm for Association Rule Mining Using a Homogeneous Dedicated Cluster of Workstations'. S Dehuri , A K Jagadev , A Ghosh . *American Journal of Applied Sciences* 1546-9239. 2006. 2006. 3 (11) p. . (And Mall R)

[Dorigo ()] *Optimization, Learning and Natural Algorithms*, M Dorigo . 1992. Italy. Politecnico di Milano (Ph.D. Thesis)

[Colorni et al. ()] *Positive feedback as a search strategy*, A Colorni , M Dorigo , V Maniezzo . No. 91-016. 1991. Milano, Italy. (Technical Report) (Politecnico di)

[Freitas] *Postgraduate Program in Computer Science, Pontificia Universidade Catolica do Parana Rua Imaculada Conceicao*, Alex A Freitas . 1155. Curitiba -PR. 80215- 901. Brazil. (A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery)

[Toivonen et al. (1995)] 'Pruning and grouping discovered association rules'. H Toivonen , M Klemettinen , P Ronkainen , K Hatonen , H Mannila . *Proc. ECML-95 Workshop on Statistics, Machine Learning, and Knowledge Discovery in Database*, (ECML-95 Workshop on Statistics, Machine Learning, and Knowledge Discovery in Database) April 1995. p. .

[Goss et al.] 'Self-organized shortcuts in the argentine ant'. S Goss , S Aron , J L Deneubourg , J M Pasteels . *Naturwissenschaften* 76 (12) p. .

[Pai and Khatri (2008)] *She buys to conquer*, Aditi Pai , Deepika Khatri . April 25, 2008. India Today.

[Colorni et al. ()] *The ant system: an autocatatlytic process*, A Colorni , M Dorigo , V Maniezzo . No. 91-016. 1991. Milano, Italy. (Technical Report) (Politecnico di)

[Holldobler and Wilson ()] *The Ants*, B Holldobler , E O Wilson . 1990. Berlin: Springer-Verlag.

[Beckers et al. ()] 'Trails and Uturns in the selection of the shortest path by the ant lasius niger'. R Beckers , J L Deneubourg , S Goss . *Journal of Theoretical Biology* 1992. 159 p. .

[Santhakumar ()] 'Transportation system management for Madurai city using GIS'. Moses Santhakumar . *Map India*, 2003.

[Dunn ()] 'Well separated clusters and optimal fuzzy partitions'. J C Dunn . *J. Cybern* 1974. 4 p. .