IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 5, Ver. I (Sep – Oct. 2014), PP 01-07
www.iosrjournals.org

Uncertainty Reduction in Data mining Task using Fuzzy based Integrated ACO and GA

Dr.K.Sankar[1] and Dr. V.Venkatachalam[2]
[1]*Professor, Department of Computer Science and Engineering, Sri Venkateswara college of Engineering and Technology, Chennai,*
[2] *Principal, The KAVERY Engineering College, Mecheri, Salem,*

Abstract: *Ant Colony System (ACS) is competitive with other nature-inspired algorithms on some relatively simple problems. This project proposes an ant colony optimization algorithm for tuning generalization of fuzzy rule. The use of Ant Colony Optimization (ACO) for classification is investigated in depth, with the development of the AntMiner+ algorithm. AntMiner+ builds rule based classifiers, with a focus on the predictive accuracy and comprehensibility of the final models. The key differences between the proposed AntMiner+ and previous AntMiner versions are the usage of the better performing MAX-MIN ant system, a clearly defined and augmented environment for the ants to walk through, with the inclusion of the class variable to handle multi-class problems, and the ability to include interval rules in the rule list.*
Ant system is a general purpose algorithm inspired by the study of behavior of ant colonies. It is based on cooperative search paradigm that is applicable to the solution of combinatorial optimization problem. The institutions concern the routing network studies the application of data mining techniques for network traffic risk analysis. The proposed work aims at spatial feature of the traffic load and demand requirements and their interaction with the geo routing environment. In previous work, the system has implemented some spatial data mining methods such as generalization and characterization. The proposal of this work uses intelligent ant agent to evaluate the search space of the network traffic risk analysis along with usage of genetic algorithm for risk pattern.

I. Introduction

Spatial data mining try to find patterns in geographic data. Most commonly used in retail, it has grown out of the field of data mining, which initially focused on finding patterns in network traffic analysis, security threats over a period of time, textual and numerical electronic information. It is considered to be more complicated challenge than traditional mining because of the difficulties associated with analyzing objects with concrete existences in space and time. Spatial patterns may be discovered using techniques like classification, association, and clustering and outlier detection. New techniques are needed for SDM due to spatial auto-correlation, importance of non-point data types, continuity of space, regional knowledge and separation between spatial and non-spatial subspace. The explosive growth of spatial data and widespread use of spatial databases emphasize the need for the automated discovery of spatial knowledge. Our focus of this work is on the methods of spatial data mining, i.e., discovery of interesting knowledge from spatial data of network traffic patterns. Spatial data are related to traffic data objects that occupy space.

II. Review of Literature

An Ant Colony Optimization algorithm (ACO) is essentially a system based on agents which simulate the natural behavior of ants, including mechanisms of cooperation and adaptation. In [5] the use of this kind of system as a new meta-heuristic was proposed in order to solve combinatorial optimization problems [4]. This new meta-heuristic has been shown to be both robust and versatile in the sense that it has been successfully applied to a range of different combinatorial optimization problems [2], [3], [6]. In [1] describe equivalence between the concepts of fuzzy clustering and soft competitive learning in clustering algorithms was proposed on the basis of the existing literature. In [9], systems for clustering with collectives of autonomous agents follow either the ant approach of picking up and dropping objects or the DataBot approach of identifying the data points with artificial life creatures. In [12], Sorting and clustering methods inspired by the behavior of real ants are among the earliest methods in ant-based meta-heuristics.

Data is clustered without initial knowledge of the number of clusters. Ant based clustering is used to initially create raw clusters and then these clusters are refined using the Fuzzy C Means algorithm [7],[8],[11]. Initially the ants move the individual objects to form heaps. The centroids of these heaps are taken as the initial cluster centers and the Fuzzy C Means algorithm is used to refine these clusters. In the second stage the objects obtained from the Fuzzy C Means algorithm are hardened according to the maximum membership criteria to form new heaps. These new heaps are then sometimes moved and merged by the ants. The final clusters formed

are refined by using the Fuzzy C Means algorithm. Results from three small data sets show that the partitions produced are competitive with those obtained from FCM.

Spatial data mining fulfills real needs of many geomantic applications. It allows taking advantage of the growing availability of geographically referenced data and their potential richness. Ref [12] aims at taking account of the spatial feature of the packets transmissions and their interaction with the geographical environment. Results of a data mining analysis may be suboptimal or even be distorted if unique features of spatial data, such as spatial autocorrelation ([10]), are ignored. In sum, convergence of GIS and data mining in an Internet enabled spatial data mining system is a logical progression for spatial data analysis technology.

Many methods have been proposed in the literature, but few of them have taken into account constraints that may be present in the data or constraints on the clustering. These constraints have significant influence on the results of the clustering process of large spatial data. Based on the C-Fuzzy sequential clustering of ACO Problem, we derived a parallel fuzzy ant clustering model to improve the attribute accuracy rate and faster execution on the proposed problem domain. In this project, the system discuss the problem of spatial clustering with obstacles constraints and propose a novel spatial clustering method based on Genetic Algorithms (GAs) and KMedoids, called GKSCOC, which aims to cluster spatial data with obstacles constraints.[9]

2.1 Problem Domain

The objective of the proposed model is to extract classification rules from continuous and nominal data. It use the Fuzzy C means algorithm as the deterministic algorithm for ant optimization. The proposed model is used after reformulation and the partitions obtained from the ant based algorithm were better optimized than those from randomly initialized hard C Means. The proposed technique executes the ant fuzzy in parallel for multiple clusters. This would enhance the speed and accuracy of cluster formation for the required system problem. The proposed ant colony based spatial data mining algorithm applies the emergent intelligent behavior of ant colonies. The experimental results on a network traffic (trend layer) spatial database show that our method has higher efficiency in performance of the discovery process compared to other existing approaches using non-intelligent decision tree heuristics

Ant-based techniques, in the computer sciences, are designed those who take biological inspirations on the behavior of these social insects. Data clustering techniques are classification algorithms that have a wide range of applications, from Biology to Image processing and Data presentation. Since real life ants do perform clustering and sorting of objects among their many activities, we expect that a study of ant colonies can provide new insights for clustering techniques. Data may be clustered using an iterative version of the Fuzzy C means (FCM) algorithm, but the drawback of FCM algorithm is that it is very sensitive to cluster center initialization because the search is based on the hill climbing heuristic. The modified ACO algorithm proposed in this work handles the nominal attribute straight away, unlike the traditional ACO algorithm. In addition the modified ACO handles both continuous and nominal attribute-values as well.

The spatial data is essential mine, useful for decision making and the knowledge discovery of interesting facts from large amounts of data. Initially study was conducted to identify and predict the number of nodes in the system, the nodes can either be a client or a server. It used a decision tree that studies from the traffic risk in a network. However, this method is only based on tabular data and does not exploit geo routing location. Using the data, combined to trend data relating to the network, the traffic flow, demand, load, etc., this work aims at deducing relevant risk models to help in network traffic safety task.

III. Ant Based Rule Mining with Parallel Fuzzy Cluster

Clustering approaches are typically quite sensitive to initialization. In this thesis, the system examine a swarm inspired approach to building clusters which allows for a more global search for the best partition than iterative optimization approaches. The approach is described with cooperating ants as its basis. The ants participate in placing cluster centroids in feature space. They produce a partition which can be utilized as is or further optimized. The further optimization can be done via a focused iterative optimization algorithm. The algorithms are from the C-means family. These algorithms were integrated with swarm intelligence concepts to result in clustering approaches that were less sensitive to initialization.

The clustering approach introduced here provides a framework for optimization of any objective function that can be expressed in terms of cluster centroids. It is highly parallelizable which could enable the time cost to be the same or lower than classical clustering. The algorithm provides a high likelihood of skipping most poor local solutions resulting in a quality partition of data. The new algorithms are loosely based on cemetery organization and brood sorting as done by ants. The algorithm introduced here requires the number of clusters be known, but has minimal sensitivity to parameter choices and results in clusters which are often better optimized than those from current algorithms. Further, the system shows that it can be integrated with cluster validity metric to potentially discover the number of classes in the data.

IV. Max-Min Ant Optimizer for Problem of Uncertainty

The MAX–MIN Ant System (MMAS) algorithm achieves a strong exploitation of the search history by allowing only the best solutions to add pheromone during the pheromone trail update. Also, the use of a rather simple mechanism for limiting the strengths of the pheromone trails effectively avoids premature convergence of the search. Finally, MMAS can easily be extended by adding local search algorithms. In fact, the best performing ACO algorithms for many different combinatorial optimization problems improve the solutions generated by the ants with local search algorithms. As our empirical results show, MMAS is currently one of the best performing ACO algorithms for the TSP. One of the main ideas introduced by max–min Ant System, the utilization of pheromone trail limits to prevent premature convergence, can also be applied in a different way, which can be interpreted as a hybrid between MMAS and Ant Colony System (ACS).

V. Efficient Spatial Data Mining Using Integrated Genetic Algorithm and ACO

The proposed spatial data mining model uses ACO integrated with GA for network risk pattern storage. The proposed ant colony based spatial data mining algorithm applies the emergent intelligent behavior of ant colonies. The proposed system handle the huge search space encountered in the discovery of spatial data knowledge. It applies an effective greedy heuristic combined with the trail intensity being laid by ants using a spatial path. GA uses searching population to produce a new generation population. The proposed system develops an ant colony algorithm for the discovery of spatial trends in a GIS network traffic risk analysis database. Intelligent ant agents are used to evaluate valuable and comprehensive spatial patterns.

5.1 Geo-Spatial Data Mining

Data volume was a primary factor in the transition at many federal agencies from delivering public domain data via physical mechanisms Algorithmic requirements differ substantially for relational (attribute) data management and for topological (feature) data management Geographic data repositories increasingly include ill structured data such as imagery and geo referenced multimedia. The strength of network GIS is in providing a rich data infrastructure for combining disparate data in meaningful ways by using spatial proximity.

The next logical step to take Network GIS analysis beyond demographic reporting to true market intelligence is to incorporate the ability to analyze and condense a large number of variables into a single forecast or score. Depending upon the specific application, Network GIS can combine historical customer or retail store sales data with syndicated demographic, business, network traffic, and market research data. Moreover, existing GIS datasets are often splintered into feature and attribute components that are conventionally archived in hybrid data management systems. Algorithmic requirement differ substantially for relational (attribute) data management and for topological (feature) data management.

5.2 Genetic Algorithm

The proposed algorithm of spatial clustering based on GAs is described in the following procedure. Divide an individual risk pattern of the network traffic generating objects (chromosome) into n part and each part is corresponding to the classification of a datum element. The optimization criterion is defined by a Euclidean distance among the data frequently, and the initial number of packets that has to be sent is produced at random. Its genetic operators are similar to standard GA's. This method can find the global optimum solution and not influenced by an outlier, but it only fits for the situation of small network traffic risk pattern data sets and classification number.

VI. Result and Discussion on Uncertainty Reduction in Data Mining Task Using Fuzzy Based Integrated ACO and GA

6.1 ANT Based Parallel Cluster Evaluation

The system implementation of fuzzy ant based parallel clustering algorithm for rule mining used three real data sets obtained from UCI repository. The simulation conducted in matlab normalizes the feature values between 0 and 1. The normalization is linear. The minimum value of a dataset specific feature is mapped to 0 and the maximum value of the feature is mapped to 1. Three data sets Glass Data Set, Wine Data Set, Iris Data Set were evaluated.

The ant initialized parallel ant fuzzy algorithm always finds better extrema for the Iris data set and for the Wine data set the ant initialized algorithm finds the better extrema 49 out of 50 times. The ant initialized HCM algorithm always finds better extrema for the Iris data set and for the Glass (2 class) data set a majority of the time. The ACO approach was used to optimize the clustering criteria, the ant approach for parallel C Means, found better extrema 64% of the time for the Iris data set. The number of ants is an important parameter of the algorithm. This number only increases when more partitions are searched for at the same time; as ants are (currently) added in increments. The quality of the final partition improves with an increase of ants, but the improvement comes at the expense of increased execution time.

No. Of Iterations	Time	
	Existing Ant Fuzzy Sequential	Proposed Ant Fuzzy Parallel
1	16.2	15.1
2	16.6	15.6
3	17.1	16.2
4	17.5	16.7
5	17.5	17.3
6	18.1	17.8
7	19.2	18.1
8	20.3	18.5
9	20.1	19.08
10	21.3	19.9

Table 1: Execution time of parallel fuzzy based ant clustering for Number of iterations

Table 1 describes Execution time of parallel fuzzy based ant clustering for Number of iterations. As iteration increases, execution time also gets increased. The proposed parallel fuzzy based scheme having better efficiency, Compared with existing model.

Time	Path length	
	Existing Ant Fuzzy Sequential	Proposed Ant Fuzzy Parallel
1	19.7	15.32
2	21.4	17.46
3	23.21	19.34
4	24.9	21.76
5	26.8	23.28
6	28.62	24.35
7	30.45	25.2
8	32.23	27.78
9	33.89	28.93
10	35.03	30.13

Table 2: Execution time of parallel fuzzy based ant clustering for path length

Table 2 values are taken from experimentation results. Execution time of parallel fuzzy based ant clustering is calculated based on path length. When execution time increases both parallel and sequential fuzzy model, path length also increased.

Dataset	Cluster Frequency	
	Existing ACO Scheme	Proposed Fuzzy ACO Scheme
Glass	32.0925	28.805
Iris	7.563	5.539
Wine	7.903	5.619

Table 3: Frequency of different extrema from parallel fuzzy based ant clustering, for Glass Iris and Wine data set

Table 3 describes Frequency of different extrema from parallel fuzzy based ant clustering. From three datasets, Glass dataset having higher cluster frequency. Iris and wine have low cluster frequency. Compared with Existing ACO Scheme, our proposed Fuzzy ACO Scheme is slightly having less cluster frequency. For Iris dataset, algorithm finds a better performance most of the time.

6.2 TSP Max_Min Ant Fuzzy Performance Evaluation

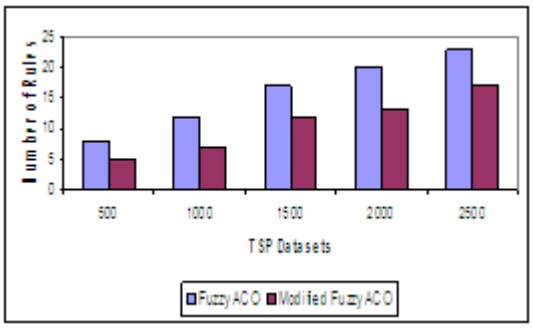
The proposed MMAS algorithm was applied to the identification of the analytical function using equally distributed fuzzy partitions with triangular membership functions for all input and output fuzzy domains and with five sets each of the TSP. The experiment consisted on the identification of the system with randomly generated training sets with three different sizes (10, 20 and 50 examples), and the subsequent run of the proposed MMAS algorithm over the identified fuzzy model. For some of the testing samples, it provides significant improvement on the schedule latency. The biggest saving achieved is 23%. This is obtained when LWID is used as the local heuristic for our algorithm and also as the heuristic for constructing the priority list for the traditional list scheduler.

On average, comparing with the force-directed approach, our algorithm provides a 6.2% performance enhancement for the testing cases, while performance improvement for individual test sample can be as much as 14.7%. Finally, compared with the optimal scheduling results computed by using the integer linear programming model, the results generated by the proposed algorithm are much closer to the optimal than those provided by the list scheduling heuristics and the force directed approach. The MMAS algorithm improves the average schedule latency by 44% comparing with the list scheduling heuristics.

TSP Dataset	Number of Rules	
	Existing fuzzy ACO Scheme	Proposed Modified Fuzzy ACO Scheme
500	8.5	5.3
1000	11.2	6.4
1500	15.3	10
2000	19.8	10.2
2500	23.4	14.4

Table 4: Number of Rules for TSP Dataset from Proposed Modified Fuzzy ACO Scheme

Table 4 gives the values of Number of rules for TSP Dataset. For finding number of rules experimentation, tsp dataset is taken. TSP Datasets are 500, 1000, 1500, 2000, and 2500. As datasets increased, number of rules also increased. Number of rules is reduced from our proposed max min Fuzzy ACO Scheme. Our model having very low number of rules and better performance compare with existing Fuzzy ACO.



Graph1: Simulation of Modified ACO

MMAS algorithm provides fuzzy models described with a lower number of rules, when compared with the initial fuzzy models and these rules had also have low complexity.

6.3 Genetic and ACO Based Spatial Data Mining Model Evaluation

Spatial classification provided by the proposed scheme is simple and efficient. It allows adapting to different decision tree algorithm for the spatial modeling of network traffic risk patterns. It uses the structure of

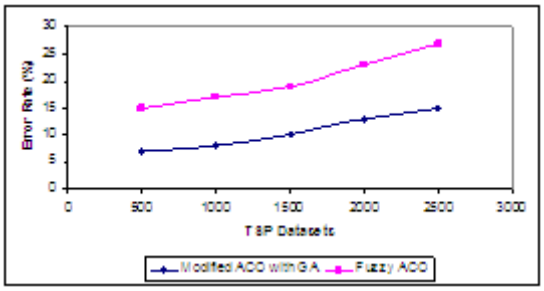
geo-data in multiple trend layers which is characteristic of geographical databases. Finally, the quality of this analysis is improved by enriching the spatial database by multiple geographical trends, and by a close collaboration with a domain specialist in traffic risk analysis. The advantage of proposed technique allows the end-user to evaluate the results without any assistance by an analyst or statistician. Gas automatically achieve and accumulate the knowledge about the search space of the ACO. GA adaptively controls the traffic risk pattern search process to approach a global optimal solution. Perform well in highly constrained traffic risk pattern, where the number of "good" solutions is very small relative to the size of the search space.

The current application results show a use case of spatial decision trees. The contribution of this approach to spatial classification lies in its simplicity and its efficiency. It makes it possible to classify objects according to spatial information (using the distance). It allows adapting any decision tree algorithm or tool for a spatial modeling problem. Furthermore, this method considers the structure of geo-data in multiple trends (patterns) which is characteristic of geographical databases.

TSP Dataset	Error Rate (%)	
	Existing fuzzy ACO Scheme	Proposed Modified Fuzzy ACO with GA Scheme
500	15.2	6.1
1000	17.1	6.8
1500	18.4	8.2
2000	22.7	9.7
2500	26.5	10.5

Table 5: Error Rate for TSP Dataset from Proposed Modified Fuzzy ACO with GA Scheme

Table 5 shows the result of Error Rate for TSP Dataset from Proposed Modified Fuzzy ACO with GA Scheme. For this experiment, TSP Dataset is taken as input dataset. Experiment is conducted on dataset 500, 1000, 1500, 2000, and 2500. Error rate is noted as percentage value. Error rate of Proposed Modified Fuzzy ACO with GA Scheme is calculated and compared with Existing Fuzzy ACO scheme. Our model gives very low error rate value. Averagely it gives 8% error rate. It's very low compared to existing fuzzy ACO scheme.



Graph2: Simulation of Modified ACO with GA

Simulation is conducted on Modified ACO with GA scheme. TSP dataset is taken for experimentation. Error rate is calculated by using the simulation. Error rate is low in our Modified ACO with GA model compared with existing fuzzy ACO model.

VII. Conclusion and Future Enhancements

7.1 Conclusion

A modified ACO based rule induction mining is carried out for 3 synthetic data sets to evaluate the performance in identifying classification rule. The proposed ACO search procedure is used in the structure of an evolutionary fuzzy system. Computer simulations on these three datasets demonstrate high performance of proposed system for continuous and nominal attributes. In this work, MMAS algorithm has been proposed to increase the generality of the fuzzy rules by searching for its structure to be maximal.

The MMAS model presents an exponential pheromone deposition approach to improve the performance of classical ant system algorithm which employs uniform deposition rule. The Spatial data mining

system of ACO with GA have shown that network traffic risk patterns are discovered efficiently and recorded in the genetic property for avoiding the collision risk in highly dense spatial regions. The proposal of our system analyzes existing methods for spatial data mining and mentioned their strengths and weaknesses. This work gives an efficient approach to multi-layer geo-data mining.

7.2 Future Enhancement

Our future work will focus on adapting recent work in multi-relational data mining domain, in particular on the extension of the spatial decision trees based on neural network. Another extension will concern automatic filtering of spatial relationships. The system will study of its functional behavior and its performances for concrete cases, which has never been done before. Finally, the quality of this analysis could be improved by enriching the spatial database by other geographical trends, and by a close collaboration with a domain specialist in traffic risk analysis. Indeed, the quality of a decision tree depends, on the whole, of the quality of the initial data.

References

[1]. Holden, N., Freitas, A. A Hybrid PSO/ACO Algorithm for is covering Classification Rules in Data Mining, In Journal of Artificial Evolution and Applications (JAEA), 2008.
[2]. D. Martens, M. De Backer, R. Haesen, M. Snoeck, J. Vanthienen, and B. Baesens. Classification with ant colony optimization IEEE Transaction on Evolutionary Computation, 11(5):651–665, 2007.
[3]. Ruckert, U., Richter, L., and Kramer, S. Quantitative association rules based on half-spaces: An optimization approach. In Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04), pages 507–510, 2004.
[4]. Parpinelli, R.S., Lopes, H.S., and Freitas, A.A. Data Mining with an Ant Colony Optimization Algorithm, IEEE Trans. on Evolutionary Computation, special issue on Ant Colony algorithms, 6(4), p. 321-332, 2002.
[5]. Parsopoulos, K. E., and Vrahatis, M. N. On the computation of all global minimizers through particle swarm optimization. IEEE Transactions on Evolutionary Computation, 8(3):211-224. 2004.
[6]. D. Martens, B. Baesens, T. Van Gestel, and J. Vanthienen. Comprehensible credit scoring models using rule extraction from support vector machines. European Journal of Operational Research, 183(3):1466–1476, 2007
[7]. D. Martens, T. Van Gestel, and B. Baesens. "Decompositional rule extraction from support vector machines by active learning" IEEE Transactions on Knowledge and Data Engineering, February 2009
[8]. P. M. Kanade and L. O. Hall, "Fuzzy ants as a clustering concept", North American Fuzzy Information Processing Society, NAFIPS 2003, 22nd International Conference of the, pp. 227 -232, 2003.
[9]. P. Kanade, "Fuzzy ants as a clustering concept", Master's thesis, University of South Florida, Tampa, FL, 2004.
[10]. A. Ultsch, "Strategies for an artificial life system to cluster high dimensional data," in Abstracting and Synthesizing the Principles of Living Systems, GWAL-6, U. Brggemann, H. Schaub, and F. Detje, Eds., 2004, pp. 128-137.
[11]. W. Lai, K. Hoe, T. Tai, and M. Seah, "Classifying english web pages with "smart" ant-like agents soft computing", in Multimedia Biomedicine, Image Processing and Financial Engineering Vol.13, 2002, pp. 411-416.
[12]. Y. Yang and M. Kamel, "Clustering ensemble using swarm intelligence," in Proceedings of the 2003 IEEE Swarm Intelligence Symposium. SIS'03, 2003, pp. 65-71.

Dr.K.Sankar is a Professor at Sri Venkateswara College of Engineering and Technology, Chennai. He obtained his Ph.D degree in Computer Science and Engineering from Anna University Chennai. His Research interests are in the field of Data Mining and Optimization Techniques.

Dr.V.Venkatachalam is a principal of The Kavery Engineering College. He received his B.E in Electronics and Communication at Coimbatore Institute of Technology Coimbatore. He obtained his M.S degree in Software systems from Birla Institute of Technology Pilani. He did his M.Tech in Computer Science at Regional Engineering College (REC) Trichy. He obtained his Ph.D degree in Computer Science and Engineering from Anna University Chennai. He has published 3 papers in International Journal and 20 papers in International & National conferences.