# Efficient Spatial Data mining using Integrated Genetic Algorithm and ACO

**Mr.K.Sankar[1] and Dr. V.Vankatachalam[2]**

**[1]Assistant Professor(Senior), Department of Master of Computer Applications, KSR College of Engineering, Tiruchengode**
**[2] Principal, The KAVERY Engineering College, Mecheri, Salem**

**Abstract**

Spatial data plays a key role in numerous applications such as network traffic, distributed security applications such as banking, retailing, etc., The spatial data is essential mine, useful for decision making and the knowledge discovery of interesting facts from large amounts of data. Many private institutions, organizations collect the number of congestion on the network while packets of data are sent, the flow of data and the mobility of the same. In addition other databases provide the additional information about the client who has sent the data, the server who has to receive the data, total number of clients on the network, etc. These data contain a mine of useful information for the network traffic risk analysis. 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.

The existing work provided a pragmatic approach to multi-layer geo-data mining. The process behind was to prepare input data by joining each layer table using a given spatial criterion, then applying a standard method to build a decision tree. The existing work did not consider multi-relational data mining domain. The quality of a decision tree depends, on the quality of the initial data which are incomplete, incorrect or non relevant data inevitably leads to erroneous results. The proposed model develops an ant colony algorithm integrated with GA for the discovery of spatial trend patterns found in a network traffic risk analysis database. 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.

Keywords: Spatial data mining, Network Traffic, ACO, GA

## 1. Introduction

Data mining is the process of extracting patterns from large data sets by combining methods from statistics and artificial intelligence with database management. Given an informational system, data mining is seen to be used as an important tool to transform data into business intelligence process. It is currently used in wide range of areas such as marketing, surveillance, fraud

detection, and scientific discovery. Automatic data processing is the result of the increase in size and complexity of the data set. This has been used in other areas of computer science as neural networks, support vector machines, genetic algorithms and decision trees. A primary reason for using data mining is to assist in the analysis of collections of observations of network user behavior.

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.

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.

## 2. Literature Review

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. This includes the spatial analysis of risk such as epidemic risk or network traffic accident risk in the router. This work deals with the method of decision tree for spatial data classification. This method differs from conventional decision trees by taking account implicit spatial relationships in addition to other object attributes. Ref [2, 3] aims at taking account of the spatial feature of the packets transmissions and their interaction with the geographical environment.

How are spatial data handled in usual data mining systems? Although many data-mining applications deal at least implicitly with spatial data they essentially ignore the spatial dimension of the data, treating them as non-spatial. This has ramifications both for the analysis of data and for their visualization. First, one of the basic tasks of exploratory data analysis is to present the salient features of a data set in a

format understandable to humans. It is well known that visualization in geographical space is much easier to understand than visualization in abstract space. Secondly, results of a data mining analysis may be suboptimal or even be distorted if unique features of spatial data, such as spatial autocorrelation ([7]), 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. Related work in this direction has been done by Koperski and Han, Ester et al. [4, 9].

Rather than aggregate data, Gridfit [1] avoids overlap in the 2D display by repositioning pixels locally. In areas with high overlap, however, the repositioning depends on the ordering of the points in the database, which might be arbitrary. Gridfit places the first data item found in the database at its correct position, and moves subsequent overlapping data points to nearby free positions, making their placement quasirandom. Cartograms [5] are another common technique dealing with advanced map distortion. Cartogram techniques let data analysts trade shape against area and preserve the map's topology to improve map visualization by scaling polygonal elements according to an external parameter. Thus, in cartogram techniques, the rescaling of map regions is independent of a local distribution of the data points. A cartogram-based map distortion provides much better results, but solves neither the overlap nor the pixel coherence problems. Even if the cartogram provides a perfect map distortion (in many cases, achieving a perfect distortion is impossible), many data points might be at the same location, and there might be little pixel coherence. Therefore, cartogram-based distortion is primarily a preprocessing step.

In [8] the author proposes an Improved Ant Colony Optimization (IACO) and Hybrid Particle Swarm Optimization (HPSO) method for SCOC. In the process of doing so, the system first use IACO to obtain the shortest obstructed distance, which is an effective method for arbitrary shape obstacles, and then the system develop a novel HPKSCOC based on HPSO and K-Medoids to cluster spatial data with obstacles, which can not only give attention to higher local constringency speed and stronger global optimum search, but also get down to the obstacles constraints. Spatial clustering is an important research topic in Spatial Data Mining (SDM). 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. 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]

## 3. Genetic and ACO Based Spatial Data Mining Model

Before data mining algorithms can be used, a target data set must be collected. As data mining only uncover patterns already present in the data, the target dataset must be large enough to contain these patterns. A common source

for data is a data mart or data warehouse. Pre-process is essential to analyze the multivariate datasets before clustering or data mining. The target set is then cleaned. Cleaning removes the observations with noise and missing data. The clean data are reduced into feature vectors, one vector per observation. A feature vector is a summarized version of the raw data observation. This might be turned into a feature vector by locating the eyes and mouth in the image. The feature vectors are divided into two sets, the "training set" and the "test set". The training set is used to "train" the data mining algorithm(s), while the test set is used to verify the accuracy of any patterns found

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.

**3.1. 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. This is the strength of predictive data mining technology and the reason why there is such a true relationship between Network GIS & data mining. 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. This dataset is then ideal for building predictive models to score new locations or customers for sales potential, cross-selling, targeted marketing, customer churn, and other similar applications. Geospatial data repositories tend to be very large. 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.

**3.2 Ant Colony Optimization**

Ant colony Optimization algorithm (ACO), a probabilistic

technique is deployed for evaluating spatial data inference from network traffic patterns which find load and demand at various instances. In the natural world, ants (initially) wander randomly, and upon finding food return to their colony while laying down pheromone trails. If other ants find such a path, they are likely not to keep traveling at random, but to instead follow the trail, returning and reinforcing it if they eventually find food.

Ant Colony Optimization (ACO) is a paradigm for designing meta-heuristic algorithms for combinatorial optimization problems. Meta-heuristic algorithms are algorithms which, in order to escape from local optima, drive some basic heuristic, either a constructive heuristic starting from a null solution and adding elements to build a good complete one, or a local search heuristic starting from a complete solution and iteratively modifying some of its elements in order to achieve a better one. The metaheuristic part permits the low level heuristic to obtain solutions better than those it could have achieved alone, even if iterated. The characteristic of ACO algorithms is their explicit use of elements of previous solutions

Over time, however, the pheromone trail starts to evaporate, thus reducing its attractive strength. The more time it takes for an ant to travel down the path and back again, the more time the pheromones have to evaporate. A short path, by comparison, gets marched over faster, and thus the pheromone density remains high as it is laid on the path as fast as it can evaporate. Pheromone evaporation has also the advantage of avoiding the convergence to a locally optimal solution. If there were no evaporation at all, the paths chosen by the first ants would tend to be excessively attractive to the following ones. In that case, the exploration of the solution space would be constrained.

Thus, when one ant finds a good (i.e., short) path from the colony to a food source, other ants are more likely to follow that path, and positive feedback eventually leads all the ants following a single path. The idea of the ant colony algorithm is to mimic this behavior with "simulated ants" walking around the graph representing the problem to solve.

### 3.3 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.

## 4. Experimental Evaluation

ACO with GA integration SPDM model is proposed to be tested in the framework of network traffic risk analysis. The analysis is done on a spatial database provided in the

framework of an industrial collaboration. It contains data on the number of packets to be sent and others on the number of nodes that is ready to be served in the network. The objective is to construct a predictive model. The system model looks correspondences between the packet and the other trend layers as the number of nodes, time taken for the packet to reach at the other end etc. It applies classification by decision tree while integrating the number of packets to be transmitted via spatial character and their interaction with the geographical environment. The experimental evaluation is made on a geographical network traffic (trend layer) spatial database to depict higher efficiency in performance of the discovery process. It proves that better quality of trend patterns discovered compared to other existing approaches using non-intelligent decision tree heuristics. Reliable data constitute the key to success of a decision tree. An efficient parallel and near global optimum search for network traffic risk patterns are evaluated using genetic algorithm. It combines the concept of survival of the fittest with a structured interchange. GAs imitates natural selection of the biological evolution. Improvements in the identification of high or low risk areas can assist the emergency preparedness planning and resource evaluation.

### 4.1 Spatial Data Mining on Network Traffic Risk Patterns

Spatial data mining on network traffic risk pattern focuses on the human vulnerability in built environments. It considers issues like differences between common and rare collision, commuting of people, and relations between accidents and networks. Visualization and interaction helps to understand the dependencies within and between data sets. Visualization supports formulating hypotheses and answering questions about correlations between certain variables and collision. Explorative visualization may reveal new variables relevant to the model and relevance of already used variables. It is highly required to analyze the correlations combine spatial data analysis methods with visualization. Risk model development is an interactive and explorative process.

### 4.2 ACO on SPDM

ACO has been recently used in some data mining tasks, e.g. classification rule discovery. Considering the challenges faced in the problem of spatial trend detection, ACO suggest efficient properties in these aspects. Ant agents search for the trend starting from their own start point in a completely distributed manner. This guides the search process to infer to a better subspace potentially containing more and better trend patterns. Finally some measures of attractiveness can be defined for selecting a feasible spatial object from the neighborhood graph. ACO on SPDM Effectively guide the trend detection process of an ant ACO has been recently used in some data mining tasks, e.g., classification rule discovery. Considering the challenges faced in the problem of spatial trend detection, ACO suggest efficient properties in these aspects. Ant agents search for the trend starting from their own start point in a completely distributed manner. Finally some measures of attractiveness can be

defined for selecting a feasible spatial object from the neighborhood graph.

### 4.3 Spatial Clustering GA

Genetic algorithms are an efficient parallel and near global optimum search method based on nature genetic and selection. GA combines the concept of survival of the fittest with a structured interchange. Gas imitates natural selection of the biological evolution. It uses searching population (set) to produce a new generation population. GAs automatically achieve and accumulate the knowledge about the search space. GA adaptively controls the search process to approach a global optimal solution. GA performs well in highly constrained problems, where the number of "good" solutions is very small relative to the size of the search space. GAs provides better solution in a shorter time, including complex problems to solve by traditional methods.

## 5. Result and Discussions

The proposed results provide spatial decision trees for network traffic risk patterns with optimized route structure with the ant agents. The proposed model classifies objects according to spatial information (using the ant agent and the distance pheromone). 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. The graph below indicates the number of trends found and paths examined using SPDM Decision Tree and SPDM-ACO-GA models for traffic risk pattern analysis.

## 6. Conclusion

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. The variety of yet unexplored topics and problems

makes knowledge discovery in spatial databases an attractive and challenging research field. This work gives an efficient approach to multi-layer geo-data mining. The main idea is to prepare input data by joining each layer table using a given spatial criterion, then applying a standard method to build' a decision tree. The most advantage is to demonstrate the feasibility and the interest of integrating neighborhood properties when analyzing spatial objects. 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] . D.A. Keim and A. Herrmann, "The Gridfit Algorithm: An Efficient and Effective Approach to Visualizing Large Amounts of Spatial Data," *Proc. IEEE Visualization Conf.*, IEEE CS Press, 1998, pp. 181-188.

[2] Anselin, L. 1988. Spatial Econometrics: Methods and Models. Dordrecht, Netherlands: Kluwer. Anselin, L. 1994. Exploratory Spatial Data Analysis and Geographic Information Systems. In Painho, M., ed., New Tools for Spatial Analysis, 45-54.

[3] Anselin, L. 1995. Local Indicators of Spatial Association: LISA. Geographical Analysis 27(2):93-115.

[4] Ester, M., Frommelt, A., Kriegel, H.P, Sander, J., "Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS Support", in *Data Minining and Knowledge Discovery, an International Journal*, 1999

[5] D.A. Keim, S.C. North, and C. Panse, "Cartodraw: A Fast Algorithm for Generating Contiguous Cartograms," *IEEE Trans. Visualization and Computer Graphics* (TVCG), vol. 10, no. 1, 2004, pp. 95-110.

[6] Haining, R. Spatial data analysis in the social and environmental sciences, Cambridge Univ. Press, 1991

[7] Hawkins, D. 1980. Identification of Outliers. Chapman and Hall. [Jain & Dubes1988] Jain, A., and Dubes, R. 1988. Algorithms for Clustering Data. Prentice Hall.

[8] Jhung, Y., and Swain, P. H. 1996. Bayesian Contextual Classification Based on Modified M-Estimates and Markov Random Fields. IEEE Transaction on Pattern Analysis and Machine Intelligence 34(1):67-75.

[9] Koperski, K., Han, J. "GeoMiner: A System Prototype for Spatial Mining", Proceedings ACMSIGMOD, Arizona, 1997.

**K.Sankar** is a Research Scholar at the Anna University Coimbatore. He is now working as a Assistant Professor(Sr) at KSR College of Engineering, Tiruchengode. His Research interests are in the field of Data Mining and Optimization Techniques.

**Dr.V.Vankatachalam** 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.