G.Kiran Kumar, P.Premchand / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 3, May-Jun 2012, pp.1746-1752

A Novel Hybrid Spatial Clustering Algorithm

G.Kiran Kumar
Associate Professor
Department of CSE
MLR IT, Hyderabad
A.P.

P.Premchand
Professor
Department of CSE
University College of Engineering
Hyderabad, A.P.

ABSTRACT. Clustering spatial data is a well-known problem that has been extensively studied to find hidden patterns or meaningful sub-groups, it is the problem of grouping data based on similarity and has many applications such as satellite imagery, geographic information systems, medical image analysis, pattern recognition, data clustering and signal processing. While this problem has attracted the attention of many researchers for many years. In this paper we discuss some very recent clustering approaches and present our research on spatial clustering approaches. Many algorithms have been designed to detect clusters in spatial databases. In this algorithm, clusters are detected and adjusted according to the intra-relationship within clusters and the inter-relationship between clusters and outliers, and vice versa. The experimental results prove that CLARA clustering algorithm when combined with medoid improves the accuracy of detection and increases the time efficiency.

Keywords:
Spatial data Mining, Spatial Clustering, PAM, CLARA, Partitioning Clustering, Hierarchical Clustering..

1.INTRODUCTION

The computerization of many business and government transactions and the advances in scientific data collection tools provide us with a huge and continuously increasing amount of data. This explosive growth of databases has far outpaced the human ability to interpret this data, creating an urgent need for new techniques and tools that support the human in transforming the data into useful information and knowledge. *Knowledge discovery in databases* (*KDD*) has been defined as an important process of discovering valid, novel, and potentially useful, and ultimately understandable patterns from data [1]. The process of KDD is interactive and iterative, involving several steps such as the following ones:

Selection: selecting a subset of all attributes and a subset of all data from which the knowledge should be discovered

Data reduction: using dimensionality reduction or transformation techniques to reduce the effective number of attributes to be considered.

Data mining: the application of appropriate algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data.

Evaluation: interpreting and evaluating the discovered patterns
with respect to their usefulness in the given application.

An important goal of data mining is to extract hidden relationships between objects, in particular relationships between some variables, possibly conditional on the values of other variables. However, looking unspecifically for possibly interesting properties of a data set.

The applications covered by spatial data mining are decisional ones, such as geo-marketing, environmental studies, risk analysis, and so on. For example, in geo-marketing, a store can establish its trade area, i.e. the spatial extent of its customers, and then analyze the profile of those customers on the basis of both their properties and the properties related to the area where they live. Now-a-days, data analysis in geography is essentially based on traditional statistics and multidimensional data analysis and does not take account of spatial data [2]. Yet the main specificity of geographic data is that observations located near to one another in space tend to share similar (or correlated) attribute values. This constitutes the fundamental of a distinct scientific area called "spatial statistics" which, unlike traditional statistics, supposes inter-dependence of nearby observations. An abundant bibliography exists in this area, including well-known geo-statistics, recent developments in Exploratory Spatial Data Analysis (ESDA) by Anselin and Geographical Analysis Machine (GAM) by Openshaw. Multi-dimensional analytical methods have been extended to support contiguity [3]. We maintain that spatial statistics is a part of spatial data mining, since it provides data-driven analyses. Some of those methods are now implemented in operational GIS or analysis tools.

The paper is organized as follows section 2 deals with overview of spatial data mining, section 3 deals with overview of spatial clustering algorithm, section 4 explains about CLARA algorithm, section 5 gives the proposed method. Results are given in section 6, section 7 presents conclusion.

2. SPATIAL DATA MINING

Climate change, natural risk prevention, human demography, deforestation, and natural resources atlas are examples from a large variety of issues arisen as result of the interaction among people and their natural environment, the earth planet. Data generated from those issues are known as spatial data. Spatial data mining methods focuses in the discovery of implicit and previously unknown knowledge in spatial databases [4]. Spatial data have many features that distinguish them from relational data. For example, the spatial objects may have topological, distance, and direction information, the complexity and the query language used to access them. Different approaches have been developed for knowledge discovery from spatial data such as Generalization-based method, Clustering, Spatial associations , Approximation and aggregation [5], Mining in image and raster databases, Classification learning , and Spatial trend detection [6]. In the following subsections we present a description of these spatial data mining approaches.

Spatial data mining is a process to discover interesting, potentially useful and high utility patterns embedded in spatial databases. Efficient tools for extracting information from spatial data sets can be of great importance to the organizations which own, generate and manage large databases. The objective of co-location pattern mining is to find the subset of features frequently located together in the same region. The spatial co-location rule problem is different from the association rule problem. There is no notion of transactions in spatial data mining. To specify group of items neighborhoods have been proposed in place of transactions.

The early research of data mining concentrated on non-spatial data, after great achievement in business, insurance, finance, etc., research emphasis has been inevitably shifted from non-spatial data (market basket data) to spatial data. We all know that there is marked and important difference between non-spatial data and spatial data: all spatial data are governed by scales in nature. To characterize a spatial object precisely and accurately, it is essential to state exactly the concomitant scale of the spatial objects.

In general, spatial data mining, or knowledge discovery in spatial databases, is the extraction of implicit knowledge, spatial relations and discovery of interesting characteristics and patterns that are not explicitly represented in the databases. These techniques can play an important role in understanding spatial data and in capturing intrinsic relationships between spatial and non-spatial data. Moreover, such discovered relationships can be used to present data in a concise manner and to reorganize spatial databases to accommodate data semantics and achieve high performance. Spatial data mining has wide applications in many fields, including GIS systems, image database exploration, medical imaging, etc. [7]. The amount of spatial data obtained from satellite, medical imagery and other sources has been growing tremendously in recent years. A crucial challenge in spatial data mining is the efficiency of spatial data mining algorithms due to the often huge amount of spatial data and the complexity of spatial data types and spatial accessing methods.

Nowadays, large amount of spatial data have been collected from many applications and data collection tools. "The spatial data explode but knowledge is poor" [8], therefore, "We are drowning in data, but starving for knowledge!" The implicit knowledge hidden in those spatial data cannot be extracted using traditional database management systems.

Spatial data mining is a budding research field that is still at its infancy. In the last decade, due to the extensive applications of GPS technology, web-based spatial data sharing and mapping, high-resolution remote sensing, and location-based services, more and more research domains have created or gained access to high-quality geographic data to incorporate spatial information and analysis in various studies, such as social analysis and business applications [9]. Besides the research domain, private industries and the general public also have enormous interest in both contributing geographic data and using the vast data resources for various application needs. In the coming years more and more new uses of spatial data and novel spatial data mining approaches are going to be developed. Although we attempt to present an overview of common spatial data mining methods in this section, readers should be aware that spatial data mining is a new and exciting field that its bounds and potentials are yet to be defined. Spatial data mining encompasses various tasks and, for each task, a number of different methods are often available, whether computational, statistical, visual, or some combination of them. Here we only briefly introduce a selected set of tasks and related methods, including classification (supervised classification), association rule mining, clustering (unsupervised classification), and multivariate geo-visualization.

2.1. Spatial classification and prediction

Grouping of data items according to their attribute values into categories is known as Classification. It is also called as supervised classification, unsupervised classification is called as clustering. ‘‘Supervised” classification needs a training dataset to train or configure the classification model, a validation dataset to validate (or optimize) the configuration, and a test dataset to test or evaluate the performance of the trained model.

Classification methods include, for example, decision trees, linear discriminant function (LDF), support vector machines (SVM), artificial neural networks (ANN), maximum likelihood estimation (MLE), nearest neighbor

methods and case-based reasoning (CBR). General-purpose classification methods have been extended in Spatial classification, which consider not only attributes of the object to be classified but also the attributes of neighboring objects and their spatial relations [10]. In [11] spatial classification a visual approach for was initiated, where the decision tree is combined with map visualization to reveal spatial patterns of the classification rules. In [12] Decision tree induction has also been used to analyze and predict spatial choice behaviors. Artificial neural networks (ANN) have been used for a broad variety of problems in spatial analysis. Remote sensing is one of the major areas that commonly use classification methods to classify image pixels into labeled categories .

In regression analysis, Spatial regression or prediction models form a special group, which considers either the independent variable or dependent variable or both of nearby neighbors in predicting the dependent variable at a specific location, such as the spatial autoregressive models (SAR). However, SAR often involve the manipulation of an n X n spatial weight matrix, which is computationally intensive if n is large[13]. To find approximate solutions for SAR more recent research efforts have been sought so that it can process very large data sets.

2.2. Spatial association rule mining

Association rule mining was originally intended to discover regularities between items in large transaction databases. Confidence denotes the strength and support indicates the frequencies of the rule. It is often desirable to pay attention to those rules that have reasonably large support [7]. Similar to the mining of association rules in transactional or relational databases, spatial association rules can be mined in spatial databases by considering spatial properties and predicates [14]. A spatial association rule is expressed in the form A →B [s%, c%], where A and B are sets of spatial or non-spatial predicates, s% is the support of the rule, and c% is the confidence of the rule. Obviously, many possible spatial predicates (e.g., close_to, far_away, intersect, overlap, etc.) can be used in spatial association rules. It is computationally expensive to consider various spatial predicates in deriving association rules from a large spatial datasets. Another potential problem with spatial association rule mining is that a large number of rules may be generated and many of them are obvious or common knowledge[15]. Domain knowledge is needed to filter out trivial rules and focus only on new and interesting findings. Spatial co-location pattern mining is spiritually similar to, but technically very different from, association rule mining [16]. Given a dataset of spatial features and their locations, a co-location pattern represents subsets of features frequently located together, such as a certain species of bird tend to habitat with a certain type of trees. Of course a location is not a transaction and two features rarely exist at exactly the same location.

2.3. Spatial clustering

In Data analysis Cluster analysis is very frequently used , which organizes a set of data items into groups (or clusters) so that items in the same group are similar to each other and different from those in other groups [17]. Many different clustering methods have been developed in various research fields such as statistics, pattern recognition, data mining, machine learning, and spatial analysis. Clustering methods can be broadly classified into two groups: partitioning clustering and hierarchical clustering. Partitioning clustering methods, such as K-means and self-organizing map (SOM) , divide a set of data items into a number of non-overlapping clusters. A data item is assigned to the ‘‘closest” cluster based on a proximity or dissimilarity measure. Hierarchical clustering, on the other hand, organizes data items into a hierarchy with a sequence of nested partitions or groupings. Commonly-used hierarchical clustering methods include the Ward’s method, single-linkage clustering, average- linkage clustering, and complete-linkage clustering [18].

To consider spatial information in clustering, three types of clustering analysis are existing, spatial clustering (i.e., clustering of spatial points), regionalization (i.e., clustering with geographic contiguity constraints), and point pattern analysis (i.e., hot spot detection with spatial scan statistics).

2.4. Geovisualization

Geovisualization concerns the development of theory and method to facilitate knowledge construction through visual exploration and analysis of geospatial data and the implementation of visual tools for subsequent knowledge retrieval, synthesis, communication and use. In this paper our work is focused on spatial clustering.

3. SPATIAL CLUSTERING

Clustering is one of the most important tasks in data mining and knowledge discovery [19]. The main goal of clustering is to group data objects into clusters such that objects belonging to the same cluster are similar, while those belonging to different ones are dissimilar. By clustering one can identify dense and sparse regions and, therefore, discover overall distribution patterns and interesting correlations among the attributes. Finding clusters in data is challenging when the clusters are of widely differing sizes, shapes and densities and when the data contains noise and outliers. A survey of clustering algorithms can be found in [20]. Although many algorithms exist for finding clusters with different sizes and shapes, there are a few algorithms that can detect clusters with different densities. Basic density based clustering techniques such as DBSCAN [21] and DENCLUE [22] treats clusters as regions of high densities separated by regions of no or low densities. So they are able to suitably handle clusters of different sizes and shapes besides effectively separating noise (outliers). But they fail to identify clusters with differing densities unless the clusters are separated by sparse regions. The following are the typical requirements for a good clustering technique in data mining [19].

Scalability: The cluster method should be applicable to huge databases and performance should decrease linearly with data size increase.

Versatility: Clustering objects could be of different types numerical data, Boolean data or categorical data. Ideally a clustering method should be suitable for all different types of data objects.

Ability to discover clusters with different shapes: This is an important requirement for spatial data clustering. Many clustering algorithms can only discover clusters with spherical shapes.

Minimal input parameter: The method should require a minimum amount of domain knowledge for correct clustering. However, most current clustering algorithms have several key parameters and they are thus not practical for use in real world applications.

Robust with regard to noise: This is important because noise exists everywhere in practical problems. A good clustering Algorithm should be able to perform successfully even in the presence of a great deal of noise.

Insensitive to the data input order: The clustering method should give consistent results irrespective of the order the data is presented.

Scalable to high dimensionality: The ability to handle high dimensionality is very challenging but real data sets are often multidimensional.

A fundamental task in knowledge discovery is the unraveling of clusters intrinsically formed in spatial databases. These clusters can be natural groups of variables, data-points or objects that are similar to each other in terms of a concept of similarity. They render a general and high-level scrutiny of the databases that can serve as an end in itself or a means to further data mining activities. Segmentation of spatial data into homogenous or interconnected groups, identification of regions with varying levels of information granularity, detection of spatial group structures of specific characteristics, and visualization of spatial phenomena under natural groupings are typical purpose of clustering with very little or no prior knowledge about the data. Often, clustering is employed as an initial exploration of the data that might form natural structures or relationships. It usually sets the stage for further data analysis or mining of structures and processes.

Clustering has long been a main concern in statistical investigations and other data-heavy researches. It is essentially an unsupervised learning, a terminology used in the field of pattern recognition and artificial intelligence, which aims at the discovery from data a class structure or classes that are unknown a priori. It has found its applications in fields such as pattern recognition, image processing, micro array data analysis, data storage, data transmission, machine learning, computer vision, remote sensing, geographical information science, and geographical research. Novel algorithms have also been developed arising from these applications. The advancement of data mining applications and the associated data sets have however posed new challenges to clustering, and it in turn intensifies the interest in clustering research. Catering for very large databases, particularly spatial databases, some new methods have also been developed over the years. These are two basic approaches to perform clustering: hierarchical clustering and partitioning clustering.

3.1 Hierarchical clustering

With reference to some criteria for merging or splitting clusters on the basis of a similarity or dissimilarity/distance measure, hierarchical clustering algorithms produce (via an agglomerative or divisive manner) a dendrogram which is a tree showing a sequence of clustering with each being a partition of the data set.

According to the structure adopted, hierarchical clustering can be further categorized into nested hierarchical clustering and non-nested hierarchical clustering. In nested hierarchical clustering, each small cluster fits itself in whole inside a larger cluster at a merging scale (or threshold) and every datum is not permitted to change cluster membership once an assignment has been made. The single-link (nearest-neighbor) algorithms, the complete-link (farthest-neighbor) algorithms, and the average-link (average-neighbor) algorithms are typical nested hierarchical clustering algorithms. The single-link method is more efficient but is sensitive to noise and tends to generate elongated clusters. Complete link and average link methods give more compact clusters but are computationally more expensive.

In non-nested hierarchical clustering, a cluster obtained at small scale may divide itself into several small parts and fits these parts into different clusters at the merging scale and, therefore, each datum is permitted to change its cluster membership as the scale varies. The algorithms proposed in generate non-nested hierarchical clustering.

Early hierarchical clustering algorithms such as AGENS (Agglomerative nesting) and DIANA (DIvisive ANAlysis) are under the curse of dimensionality and do nor scale well for large data sets because of the difficulties in deciding on the merge or split points. To handle large data sets, BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) obtains clusters by compressing data into smaller sub-clusters. The algorithm appears to be linearly scalable and gives reasonably good-quality clustering. Clusters are spherical in shape but they may not be natural clusters. By combining random sampling and partitioning, CURE (Clustering Using REpresentatives) merges clusters via the concepts of representative objects and shirking factor. It is relatively robust to outliers (objects in non-dense regions) and can identify clusters with non-spherical shapes and large variance. Somewhat similar to CURE, CHAMELEON employs the concepts of interconnectivity and closeness to merge clusters. The algorithm appears to be more effective than CURE in identifying clusters with arbitrary shapes and varying density.

The advantage of hierarchical clustering algorithms is that it is more versatile. They give a series of clusterings along some scales. The time complexity for agglomerative

algorithms is $O(n^2 \log n)$ and the space complexity is $O(n^2)$, where n is the number of objects. The disadvantage of hierarchical clustering is that it is often difficult to determine at which level the clustering gives the optimal clusters essential to an investigation.

3.2 Partitioning Clustering:

Differing from the hierarchical approach, partitioning algorithms give only a *single partition of a data set*. The majority of such algorithms partition a data set into clusters through the minimization of some suitable measures such as a cost function. The K-means method, FORGY, ISODATA, WISH, and Fuzzy ISODATA are essentially based on the minimization of a squared-error function. The K-means methods use the mean value of the objects in a cluster as the cluster center. Its time complexity is O(nkt), where n is the number of objects, k is the number of clusters, and t is the number of iterations. That is, for fixed k and t, the time complexity is O(n). Thus, it is essentially linear in the number of objects and this becomes its advantage. However, the K-means method is sensitive to initial partition, noise, and outliers (objects whose removal improves significantly the tightness of the clusters), and it cannot discover clusters of arbitrary shapes. By using the most centrally located object (medoid) in a cluster as the cluster center, the K-medoid is less sensitive to noise and outliers but in the expense of a higher computational cost. PAM (Partitioning Around Medoids) is an earlier K-medoid method that uses a complex iterative procedure to replace k cluster centers. The computational complexity in a single iteration is $O(k(n-k)^2)$. Thus, the algorithm is very costly for large data sets. To deal with large volume of data, CLARA (Clustering LARge Application) takes multiple samples of the whole data set and applies PAM to each sample to give the best clustering as the output. The computational complexity for each iteration becomes $O(ks^2+k(n-k))$, where s is the sample size. So, the success of CLARA depends on the sample chosen. Good-quality clustering will not be achieved if the samples are biased.

By combining PAM and CLARA, CLARANS (Clustering Large Applications based upon RANdomized Search) is constructed to search only the subset of a data set but not confining itself to any sample at any time. The process is similar to searching a graph as if everyone if its nodes are potential solutions. The algorithm attempts to search for a better solution by replacing the current one with a better neighbor in an iterative manner. Though CLARAN appears to be more effective than PAM and CLARA, its computational complexity is roughly $O(n^2)$. Furthermore, it assumes that all objects to be clustered are stored in the main memory. It should be noted that most of the partitioning methods cluster objects on the basis of the distance between them. It actually constitutes the expensive step of the algorithms. Since the minimization problems involved are generally NP-hard and combinatorial in nature, techniques such as simulated annealing, deterministic annealing, and EM (expectation maximization) algorithms are often utilized to lower the computational overhead. Moreover, most of the existing algorithms can only find clusters which are spherical in shape.

In addition to the hierarchical and partitioning approaches, there are other clustering methods such as the graph theoretic methods, the density-based methods, the grid-based methods, the neural network methods, the fuzzy sets methods, and the evolutionary methods.

The graph theoretic methods often convert the clustering problem into a combinatorial optimization problem that is solved by graph algorithms or heuristic procedures. The density-based methods generally assume a mixture of distributions, with each cluster belonging to a specific distribution, for the data. Their purpose is to identify the clusters and the associated parameters. The grid-based methods impose a grid data structure on the data space in order to make density-based clustering more efficient. They however suffer from the curse of dimensionality as the number of cells in the grid increases. Neural network models generally perform clustering through a learning process. The self-organizing map, for example, can be treated as an on-line version of k-means with competitive learning. The fuzzy sets methods solve clustering problems where an object can belong to multiple clusters with different degrees of membership. The fuzzy c-means algorithm and fuzzy graph method are typical examples.

4. CLARA ALGORITHM

To eliminate the computational complexity problem of PAM algorithm, another partition based clustering algorithm called LARA was introduced by [23]. The PAM algorithm is given below [24]. This procedure, considers small samples of the actual data as a representatives of the data. PAM algorithm is used to identify the medoids for each of these samples. Then each object of the entire dataset is assigned to the resulting medoids. Similar to PAM, the objective function is computed to select the best set of medoids as output. Experiments described in [23] indicated that 5 samples of size 40+ 2k give satisfactory results. The computational complexity of each iteration of CLARA is of $O(ks^2 + k(n-k))$, where s is the size of the sample.

4.1 PAM Algorithm:

1. Build Phase: Randomly select two initial data points as medoids. The selection is made in such a way that the dissimilarity to all other data objects is minimal. The main objective of this step is to decrease the objective function.

2. **Swap Phase**: The Swap phase computes the total cost 'T' for all pairs of objects ri and sh, where ri € R is currently selected and sh € S is not.

3. **Selection Phase:** This phase selects the pair (ri, sh) which minimizes 'T'. If the minimum T is negative, the swap is carried out and the algorithm reiterates Step 2. Otherwise, for each non-selected object, the most similar medoid is found and the algorithm stops.

4.2 CLARA Algorithm

1. For i=1 to 5, repeat Steps 2 to 5.
2. Draw a sample of 40 + 2k objects randomly from the entire data set and call PAM algorithm to find k medoids of the sample.
3. For each object O in the entire data set, determine k-medoids which is most similar to O.
4. Calculate average dissimilarity of the clusters obtained from Step 3. If this value is less than current minimum, use the new value as current minimum and retain the k medoids found in Step 2 as the best set of medoids obtained so far.

5. Return to Step 1 to start the next iteration.

5. PROPOSED METHOD

As understood from the literature study, clustering algorithms consider outlier detection but only to the point they do not interfere with the clustering process. In these algorithms, outliers are only by-products of clustering algorithms and they cannot rank the priority of outliers. Furthermore, algorithms that combine and compare the performance of using partition clustering algorithm combined with distance based outlier detection is not available. In this study, the CLARA algorithm is modified and combined with mediod distance is proposed and tested on large and small datasets. The proposed methodology is given below.

Modified CLARA Algorithm

1.Perform clustering PAM/CLARA/CLARANS.

2. Calculate the average number of points in 'k' cluster (AKN).

3.Clusters are segregated as small and large clusters. All those clusters which have less than half of AKN are declared as small cluster.

4. Small clusters are removed from the datasets as outliers. The outliers in the large clusters are then detected using the following procedure.

5. The Absolute Distances between the Medoid (ADMP) of the current cluster and each one of the points (pi) is calculated using Equation 1. A threshold value is calculated as the average of all ADMP values of the same cluster multiplied by 1.5. When the ADMP value of a cluster is greater than T, then it is an outlier, else it is an inlier.

$$ADMP = |pi - \mu| \quad ---(1)$$

6. RESULTS

In this paper experiments are performed using PAM algorithm, CLARA algorithm and proposed algorithm. Proposed algorithm which is modified version of CLARA algorithm is explained in section 5. Experiments are performed on standard data sets i.e. iris and bupa data sets. Our results show that proposed CLARA algorithm with mediod approach outperforms in terms of time efficiency over other algorithms.

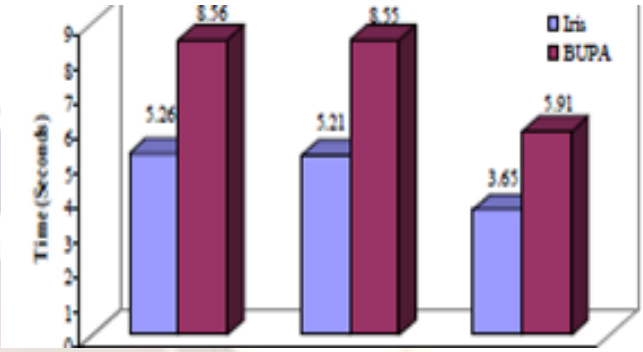

Fig 1 clustering algorithms performance

7. CONCLUSION

In this paper we have addressed the problem of clustering spatial data. Finally our experimental results showed that the performance of the proposed algorithm is much better than the existing algorithm. This decreases the time taken for the detection process. The experimental results prove that CLARA clustering algorithm when combined with medoid improves the accuracy of detection and increases the time efficiency.

8. REFERENCES

1. Fayyad, U. M., .J., Piatetsky-Shapiro, G., Smyth, P. 1996: *"From Data Mining to Knowledge Discovery: An Overview"*, in: Advances in Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, pp.1-34.

2. K Zeitouni "A survey of spatial data mining methods databases and statistics point of views", Data warehousing and web engineering, 2002 - books.google.com.

3. Longley P. A., Goodchild M. F., Maguire D. J., Rhind D. W., Geographical Information Systems - Principles and Technical Issues, John Wiley & Sons, Inc., Second Edition, 1999.

4. G. Kiran Kumar, T.Venu gopal and P.Premchand "A Novel method of modeling Spatial Co-location patterns on spatial Database", 2nd International conference ICFOCS 2011 held at IISc Bangalore,India Aug 7-9, 2011.

5. Manuel Alfredo PECHPALACIO, "*Spatial Data Modeling and Mining using a Graph-based Representation*", PhD Thesis.

6. Martin Ester, Hans-Peter Kriegel, Jörg Sander "*Algorithms and Applications for Spatial Data Mining*", Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS, Taylor and Francis, 2001.

7. M. S. Chen, J. Han, P. S. Yu. "Data mining, an overview from database perspective", IEEE Transactions on Knowledge and data Engineering, 1997.

8. Li D.R., Wang S.L., Li D.Y. and Wang X.Z., 2002, Theories and technologies of spatial data knowledge discovery. Geomatics and Information Science of Wuhan University 27(3), 221-233.

9. Spielman, S. E., & Thill, J. C. , 2008, " *Social area analysis, data mining and GIS*" . Computers Environment and Urban Systems, 32(2), 110–122.

10. Ester, M., Kriegel, H. P., & Sander, J. ,1997. "*Spatial data mining: A database approach* ". In Advances in spatial databases (pp. 47–66). Berlin: Springer-Verlag Berlin.

11. Yao, X., & Thill, J.-C. 2007. "Neurofuzzy modeling of context–contingent proximity relations" Geographical Analysis, 39(2), 169–194.

12. Kazar, B. e tal . 2004. Comparing exact and approximate spatial auto-regression model solutions for spatial data analysis (pp. 140–161). Berlin: Springer-Verlag.

13. K Zeitouni "A survey of spatial data mining methods databases and statistics point of views", Data warehousing and web engineering, 2002 - books.google.com.

14. Han, J., Kamber, M., 2001, Data Mining: Concepts and Techniques (San Francisco: Academic Press)

15. S Shekhar, P Zhang. Data Mining and Knowledge Discovery 2010 – Springer.

16. Shekhar, S., & Huang, Y. (2001). *Discovering spatial co-location patterns: A summary of results*. In C. Jensen, M. Schneider, B. Seeger, & V. Tsotras (Eds.), Advances in spatial and temporal databases, proceedings, lecture notes in computer science (pp. 236–256). Berlin: Springer-Verlag.

17. Gordon, A. D. 1996. "Hierarchical classification. Clustering and classification" (pp. 65–122). River Edge, NJ, USA: World Scientific Publisher.

18. Diansheng Guo, Jeremy Mennis ,2009 " *Spatial data mining and geographic knowledge discovery—An introduction*", Computers, Environment and Urban Systems, Elsevier.

[19] J. Han and M. Kamber, *Data Mining Concepts and Techniques*. Morgan Kaufman, 2006.

[20] R. Xu, "Survey of Clustering Algorithms," *IEEE Transaction on Neural Networks*, vol. 16, no. 3, May 2005.

[21] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, " A density based algorithm for discovering clusters in large spatial data sets with noise," in *2nd International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 226– 231.

[22] A. Hinneburg and D. Keim, "An efficient approach to clustering in large multimedia data sets with noise," in *4th International Conference on Knowledge Discovery and Data Mining*, 1998, pp. 58–65.

[23] Kaufman, L. and Rousseeuw, P. (1990) Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, New York.

[24] Ng, R. and Han, J. (2002) CLARANS: A Method for Clustering Objects for Spatial Data Mining, IEEE Transactions on Knowledge and Data Engineering. Vol.14, No.5.