Quality Assessment in Spatial Clustering of Data Mining

Azimi, A. and M.R. Delavar

Centre of Excellence in Geomatics Engineering and Disaster Management, Dept. of Surveying and Geomatics
Engineering, Engineering Faculty, University of Tehran, Tehran, Iran
ar.azimi@yahoo.com
mdelavar@ut.ac.ir

Abstract

Because of the use of computers and its advances in scientific data handling and advancement of various geo and space borne sensors, we are now faced with a large amount of data. Therefore, the development of new techniques and tools that support the transforming the data into useful knowledge has been the focus of the relatively new and interdisciplinary research area named “knowledge discovery in spatial databases or spatial data mining”. Spatial data mining is a demanding field since huge amounts of spatial data have been collected in various applications such as real-estate marketing, traffic accident analysis, environmental assessment, disaster management and crime analysis. Thus, new and efficient methods are needed to discover knowledge from large databases such as crime databases. Because of the lack of primary knowledge about the data, clustering is one of the most valuable methods in spatial data mining. As there exist a number of methods for clustering, a comparative study to select the best one according to their usage has been done in this research. In this paper we use Self Organization Map (SOM) artificial neural network and K-means methods to evaluate the patterns and clusters resulted from each one. Furthermore, the lack of pattern quality assessment in spatial clustering can lead to meaningless or unknown information. Using compactness and separation criteria, validity of SOM and K-means methods has been examined. Data used in this paper has been divided in two sections. First part contains simulated data contain 2D x,y coordinate and second part of data is real data corresponding to crime investigation. The result of this paper can be used to classify study area, based on property crimes. In this work our study area classified into several classes representing high to low crime locations. Thus, accuracy of region partitioning directly depends on clustering quality.

Keywords: Spatial Data Mining, Quality Assessment, Clustering, Compactness, Separation

1. INTRODUCTION

Data clustering is a useful technique for many applications, such as similarity search, pattern recognition, trend analysis, market analysis, grouping and classification of documents [10]. Clustering is perceived as an unsupervised process since there are no predefined classes and no examples that would show what kind of desirable relations should be valid among the data [11, 15]. Consequently, the final partitions of a data set require some sort of evaluation in most applications. The fundamental clustering problem is to partition a given data set into groups (clusters), such that the data points in a cluster are more similar to each other than points in different clusters [14].

Spatial data itself lies in uncertainty, and on the other hand, any uncertainty reproduced in spatial data mining process, propagated and accumulated, leads to the production of uncertain knowledge [8].

The uncertainties in knowledge discovery and data mining may exist in the process of spatial data selection, spatial data preprocessing, data mining and knowledge representation.

The uncertainty in data mining is the result of uncertainty in data and/or the data mining analysis undertaken. At the same time, a number of uncertainties exist in spatial data mining. The main phase of knowledge discovery in database (KDD) is data mining and it refers to the limitation of models and mining algorithms such as clustering algorithms. Clustering is one of the tasks in the data mining process for discovering groups and identifying particular distributions and patterns in the underlying data [14]. Thus, the main concern in the clustering process is to reveal the organization of patterns into sensible groups, which allows us to discover similarities and differences, as well as to derive useful inferences about them [14].

In this paper we present a clustering validity procedure, which evaluates the results of clustering algorithms on same data sets. We use a validity index, CD, based on well-defined clustering criteria enabling the selection of the optimal number of clusters for a clustering algorithm that result in the best partitioning of a data set.

2. SPATIAL DATA MINING

Huge amounts of data have been collected through the advances in data collection, database technologies and data collection techniques. This explosive growth of data creates the necessity of automated knowledge/information discovery from data, which leads to a promising and emerging field, called data mining or knowledge discovery in databases [16]. Spatial data mining is the discovery of interesting relationships and characteristics that may exist implicitly in spatial databases [1]. KDD follows several stages including data selection, data preprocessing, information extraction or spatial data mining, interpretation and reporting [2].Data mining is a core component of the KDD process.

Spatial data mining techniques are divided into four general groups: spatial association rules, spatial clustering, spatial trend detection and spatial classification [3, 4, 5].

2.1. Spatial Association Rules

Spatial association rules mean the rules of the form "P ==> R", where P and R are sets of predicates, use spatial and non-spatial predicates in order to describe spatial objects using relations with other objects.

2.2. Spatial Clustering

Clustering is the task of grouping the objects of a database into meaningful subclasses (that is, clusters) so that the members of a cluster are as similar as possible whereas the members of different clusters differ as much as possible from each other.

2.3. Spatial Trend Detection

We define a spatial trend as a regular change of one or more non-spatial attributes when moving away from a given object. We use neighborhood paths starting to model the movement and we perform a regression analysis on the respective attribute values for the objects of a neighborhood path to describe the regularity of change [13].

2.4. Spatial Classification

The task of classification is to assign an object to a class from a given set of classes based on the attribute values of the object. In spatial classification the attribute values of neighboring objects may also be relevant for the membership of objects and therefore have to be considered as well.

3. CLUSTERING

The main advantage of using clustering is that interesting structures or clusters can be found directly from the data without using any prior knowledge.

Clustering algorithms can be roughly classified into hierarchical methods and non-hierarchical methods. Non-hierarchical method can also be divided into four categories; partitioning methods, density-based methods, grid-based methods, and model-based methods [6].

Partitioning methods generate initial k clusters and improve the clusters by iteratively reassigning elements among k clusters. The number of “k” and iteration are user inputs. K-means was selected as a partitioning method. Self Organization Map (SOM) as a model-based method is an unsupervised learning neural network that maps an n-dimensional input data to a lower dimensional output map while maintaining the original topological relations [6].

3.1. K-means

The K-means method is probably the most well known clustering algorithms. The algorithm starts with k initial seeds of clustering, one for each cluster. All the n objects are then compared with each seed by means of the Euclidean distance and assigned to the closest cluster seed. The procedure is then repeated over and over again. At each stage the seed of each cluster is recalculated using the average vector of the objects assigned to the cluster. The algorithm stops when the changes in the cluster seeds from one stage to the next are close to zero or smaller than a pre-specified value. Every object is only assigned to one cluster [7].

The accuracy of the K-means procedure is basically dependent upon the choice of the initial seeds. To obtain better performance, the initial seeds should be very different among themselves.

3.2. Self Organizing Map (SOM)

Self Organization Map has the ability to learn unsupervised pattern. The SOM neural network is a very promising tool for clustering and mapping spatial datasets describing nonlinear phenomena [12]. Self-organizing networks modify their connection weights based only on the characteristics of the input patterns. The goal of the learning process is not to make predictions, but to classify data according to their similarity. In the neural network architecture, the classification is done by plotting the data in *n*-dimensions onto a, usually, two-dimensional grid of units in a topology preserving manner. The neural network consists of an input layer and a layer of neurons. The neurons or units are arranged on a rectangular or hexagonal grid and are fully interconnected [12].

4. QUALITY ASSESSMENT IN DATA MINING PROCESS OF KDD

Spatial data itself lies in uncertainty, and on the other hand, a number of uncertainties exist in spatial data mining process [8]. Uncertainties dealing with data mining mainly refer to the limitation of mathematical models, and mining algorithm may further propagate, enlarging the uncertainty during the mining process (Figure 1).

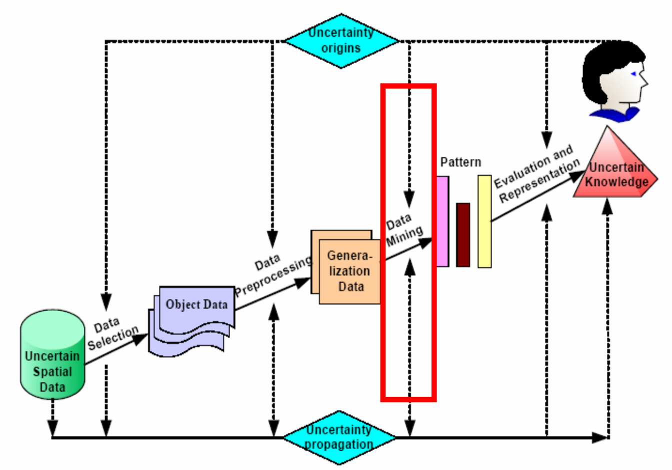


Figure 1: Uncertainties and their propagation in the process of spatial data mining [8]

4.1. Quality Assessment of Spatial Data Clustering

Clustering is mostly an unsupervised procedure. Obtaining high quality clustering results is very challenging because of the inconsistency of the results of different clustering algorithms. This implies that there are no predefined classes and most of the clustering algorithms depend on assumptions and initial guesses in order to define the best fitting for the specific data set [9]. To decide the number of clusters and evaluation of clustering results have been the subject of several research efforts [8]. The clustering algorithm always tries to find the best fit for a fixed number of clusters. However, this does not mean that even the best fit is meaningful at all. Either the number of clusters might be wrong or the cluster shapes might not correspond to the groups in the data, even if the data can be grouped in a meaningful way at all. The criteria widely accepted for partitioning a data set into a number of clusters are: i. the s*eparation* of the clusters, and ii. their *compactness* [14]. The optimum case implies parameters that lead to partitions that are as close as possible (in terms of similarity) to the real partitions of the data set [14]. Several assessment indices have been introduced, however, in practice; they are not used by most of the clustering methods. A reliable quality assessment index should consider both the compactness and the separation. One of the quality measures that can be used in clustering is described as follows [8]:

The variance of spatial data set *X*, called $\sigma(X)$, the value of the p-th dimension is defined as follows [14, 8]:

$$\sigma_x^p = \frac{1}{n}\sum_{k=1}^{n}(x_k^p - \bar{x}^p)^2$$

where $\bar{x}^p$ is the p-th dimension of

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k, \forall x_k \in X$$

The variance of cluster i is called $\sigma(v_i)$ and its p-th dimension defined as [14,8]:

$$\sigma_{vi}^p = \frac{\sum_{k=1}^{n_i}(x_k^p - v_i^p)^2}{n_i}$$

The total variance of spatial data set with respect to c clusters is:

$$\sigma = \sum_{i=1}^{c} \sigma(v_i)$$

The average compactness of c clusters, *Comp* [14, 8]:

$$Comp = \sigma / c$$

The average scattering of data set compactness, *Scat_Comp* [14,8]:

$$Scat\_Comp = Comp / \|\sigma(x)\|$$

The more compact the clusters are, the smaller the *Scat_Comp* is. Thus, for a given spatial data set, a smaller *Scat_Comp* indicates a better clustering scheme.

The distance between clusters is defined by the average distance between the centers of specified clusters, that is [8]:

$$d = \frac{\sum_{i=1}^{c}\sum_{j=1}^{c} \|v_i - v_j\|}{c(c-1)}$$

The larger d is, the more separated the clusters are. According to above definitions, a quality measure for clustering was defined as follows [8]:

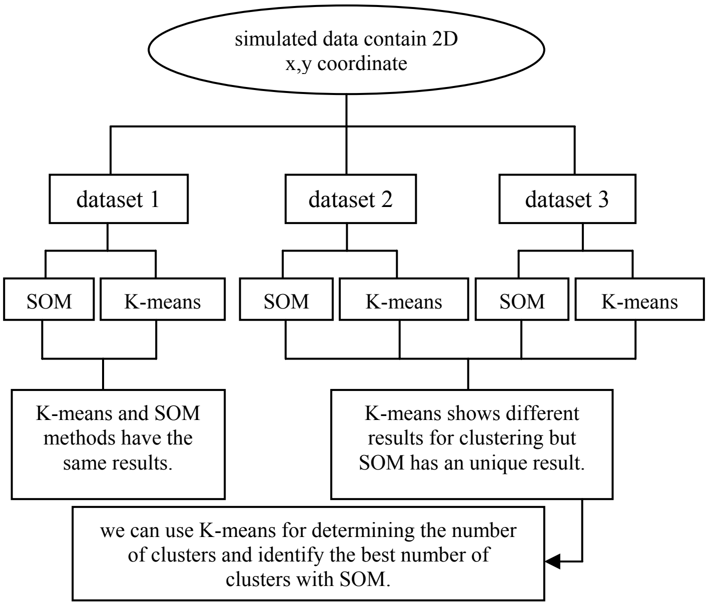
$$CD = Scat\_Comp / d$$

The definition of *CD* indicates that both criteria of "good" clustering (i.e., compactness and separation) are properly combined, enabling reliable evaluation of clustering results [8]. Small values of *CD* indicate all the clusters in clustering scheme are overall compact and separated.

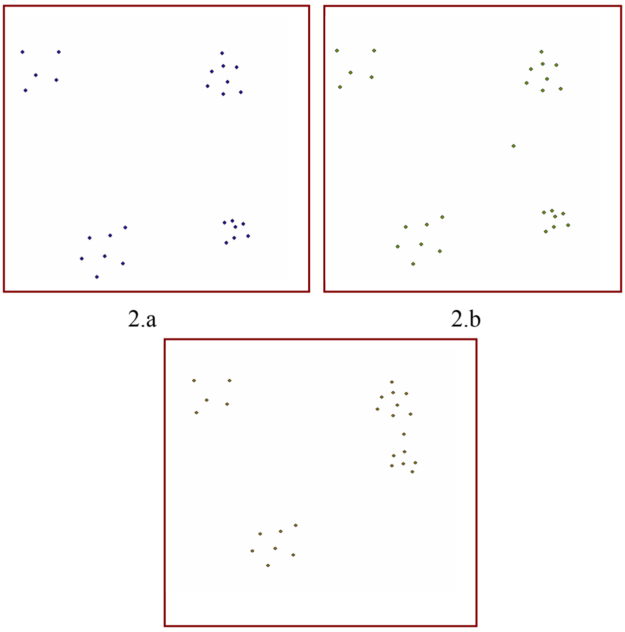
5. IMPLEMENTATION

Data used in this paper divided in two sections. First part contains simulated data in 2-dimension which include three datasets contain 2D x,y coordinate. Second part of data is real data corresponding to crime investigation which include n-dimension[1] in distance.

Our real data are related to property crime. In section of real data, first, some raw raster layers were added to existing data based on the number of crime locations. Then for each cell of raster layer, the distance of crime committed cell for each cell of raster layer was computed, here means of distance is Euclidean distance. These distances, will be assigned to corresponding raster layer value. These operations are repeated for each crime location and each raster layer. In order to profit all created rasters for data mining, we made a multispectral raw raster data layer. After these operations we made our database (relational database) for data mining process, in this part each row of table illustrates each cell of final raster and each field show distances. Our real data has higher dimension and numbers than the simulated data. About simulated data we have a background of systematic division of items into parts but in real data we do not have any priori knowledge about distributions. Figure 2 shows various ways to partition simulated data and assume the optimal partitioning of data set in four clusters. In this section we use the proposed validity index '*CD*' experimentally.


simulated data contain 2D x,y coordinate
dataset 1
dataset 2
dataset 3
SOM
K-means
SOM
K-means
SOM
K-means
K-means and SOM methods have the same results.
K-means shows different results for clustering but SOM has an unique result.
we can use K-means for determining the number of clusters and identify the best number of clusters with SOM.

Flowchart 1. Process to find the optimal number of clustering



2.a 2.b

2.c

Figure 2:

a- Simulated dataset 1: contain full separated data.

b- Simulated dataset 2: contain separated data and one departed member from other data.

c- Simulated dataset 3: contains not well separated data.

At first two methods including K-means and SOM, to analysis the simulated data were implemented. Numbers of selected clusters for two methods were 2, 3, 4 and 5. Criteria for compacted and separated clusters achieved from *CD* index.

Although experimental datasets used in this paper have small dimension and number, but they have complexity in distribution. First dataset contains separable group and are clustered, however, the second dataset has some outliers which cause some problems in clustering, these problems are evident in clustering schema (refer to Figure, 4 and 5). In the

[1] n refers to number of crimes

third one, two groups are so close which is the major problem of cluster analysis. Thus consideration of the second and third datasets is so important in clustering qualification and will be useful especially for datasets which their distribution is unknown.

Table 1 presents *CD* index values for the resulting clustering schema for dataset 1 found by K-means and SOM, respectively. The clustering schemes and values of index for both methods are same (Figure 3).

Table 1: Results of K-means and SOM for dataset 1

		K=2	K=3	K=4	K=5
For dataset 1	K-means	0.7274	0.4993	0.0386	0.1966
	SOM[1]	0.7274	0.5274	0.0386	0.2852

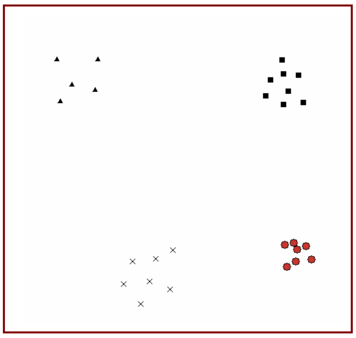


Figure 3: Perspective of results for dataset 1 (K-means and SOM) K=4

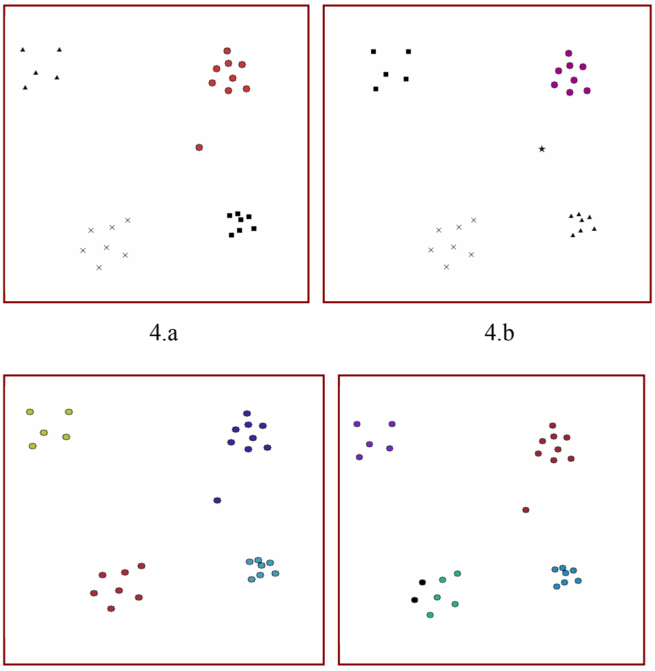
In Tables 2 and 3 differences in values of *CD*, due to different input values for clustering algorithms have been applied to a dataset resulting in creation of different partitioning schemes. Here we note the *CD* index for clustering methods is independent of the algorithms [14].

Tables 2 and 3 present different *CD* index for several iterations for K-means and SOM.

Figures 4 and 5 present the partitioning of dataset 2 and 3. K-means apportion data into four and five clusters for dataset 2 and it divides data into three and four clusters for dataset 3. The reason of unusual result for K-means clustering is achieved due to distributions of the data. However, we can introduce optimum clustering by SOM method in each dataset.

Table 2: Results of K-means and SOM for dataset 2

		K=2	K=3	K=4	K=5
For dataset 2	K-means	0.6104	0.4592	0.831	0.0697
		0.7174	0.5896	0.831	0.2240
	SOM	1.2104	1.2560	1.1000	1.1964



4.a

4.b

4.c

4.d

Figure 4:

a- Perspective of results for dataset 2 (K-means) k=4

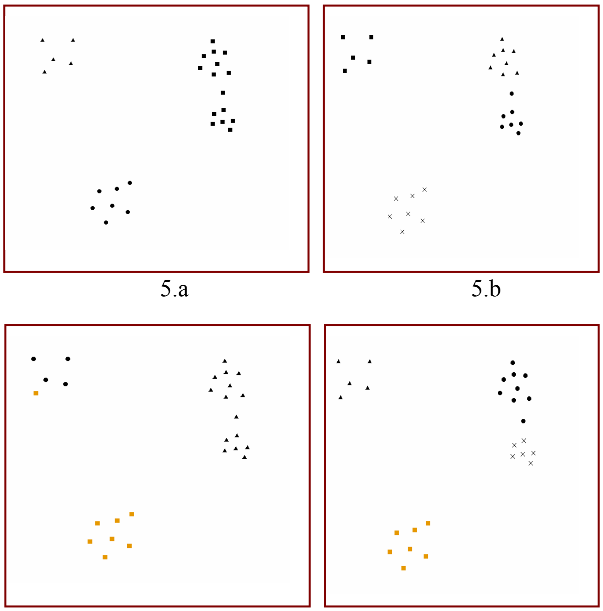
b- Perspective of results for dataset 2 (K-means) k=5

c- Perspective of results for dataset 2 (SOM) k=4

d- Perspective of results for dataset 2 (SOM) k=5

Table 3: Results of K-means and SOM for dataset 3

		K=2	K=3	K=4	K=5
For dataset 3	K-means	1.2416	1.3723	1.1321	1.5406
		1.5874	1.0778	1.1132	1.2693
	SOM	1.4035	1.2105	1.1458	1.2706



5.a

5.b

5.c

5.d

Figure 5:

a- Perspective of results for dataset 3 (K-means) k=3

b- Perspective of results for dataset 3 (K-means) k=4

c- Perspective of results for dataset 3 (SOM) k=3

d- Perspective of results for dataset 3 (SOM) k=4

[1] Cause of distrbution of points, selected topology function for SOM was used GRIDTOP.

In the following, two methods including K-means and SOM, to analysis the property crime data were implemented (Figure 6).

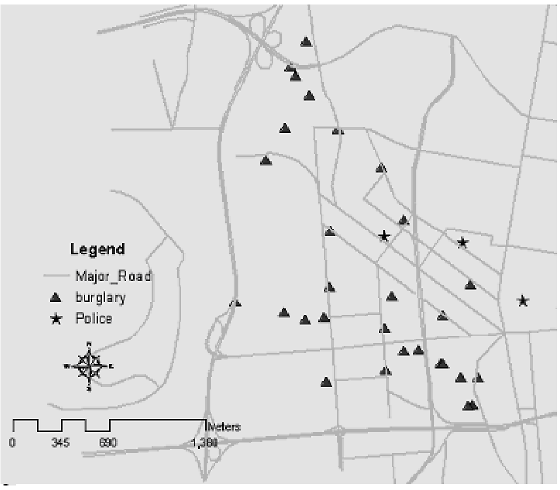
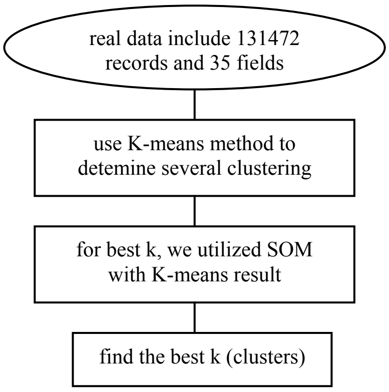


Figure 6: Our real data are related to property crime

Due to the achieved results in the Tables, K-means method can not represent appropriate clustering by itself. In datasets 2, and 3, data does not have well defined condition and K-means shows different results for clustering. K-means method has been used to determine the number of clusters. In the next step we use K-means and SOM methods to determine number of clusters for crime datasets. To identify the best number of clusters we apply the K-means because of speed and simplicity, and then results have been used for SOM. Extracting the best number of clusters from SOM is time consuming. The achieved results from K-means helps us to find number of clusters form SOM method in a much shorter time (Flowchart 2).

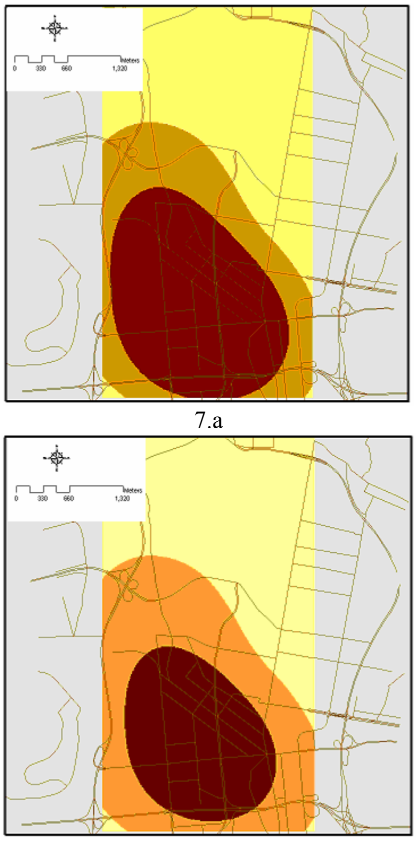


Flowchart 2. Process to find the optimum number of clustering for real data in the best clustering way with SOM.

Table 4: Results of the K-means and SOM for crime dataset

			K=2	K=3	K=4	K=5
For crime dataset	K-means		0.2873	0.2546	0.2676	0.2663
			0.2892	0.2546	0.2676	0.2746
			0.2892	0.2586	0.2572	0.2750
	SOM	GRIDTOP		0.2079	0.2286	
		HEXTOP		0.2532	0.2647	

The result of this paper using real data section can be used to classifying study area, based on property crimes. In this work our study area classified into several classes representing high to low crime locations. Thus, accuracy of region partitioning directly depends on clustering quality.



7.a

7.b

Figure 7:
a- Perspective of study area by using K-means
b- Perspective of study area by using SOM

6. CONCLUSION

It is important to note that there are certain conditions that must be considered in order to render robust performances from SOM. In SOM, if network does not have correct number of clusters, we do not get good results because of high dependence to the shape of point distribution. SOM is too sensitive to outliers and does not give correct clustering results with respect to effect of predefined topology function and point distribution (Table 2). However, during our tests it is quite evident that clusters are better explored by SOM (Figure 7.b). This is due to the effect of the topology which forces units to move with respect to each other in the early stages of the process. Due to simplicity, K-means is faster than SOM. K-means can be utilized as a pre-clustering method to identify accurate number of clusters. The results are affected by predefined cluster centers so the algorithm should have several iterations to achieve best choices.

References

[1] T. Ng. Raymond and J. Han. Efficient and effective clustering methods for spatial data mining. VLDB Conference. Santiago, Chile, 1994.

[2] F. Karimipour, M.R. Delavar and M. Kianie. Water quality management using GIS data mining. Journal of Environmental Informatics, Vol.5, No.2, pp.61-71, 2005.

[3] S. Shekhar, P. Zhang, R. R. Vatsavai and Y. Huang. Trend in spatial data mining, Data mining: next generation challenges and future directions. AAAI/MIT Press, 2003.

[4] M. Ester, H.P Kriegel, and J. Sander. Algorithms and Applications for Spatial Data Mining in Geographic Data Mining and Knowledge discovery. Taylor & Francis, 2001.

[5] K. Koperski, E. Clementini and P. D. Felice. Mining multiple-level spatial association rules for objects with a broad boundary. Elsevier, Data & Knowledge Engineering Vol.34, 251-270, 2000.

[6] X. Hu and I. Yoo. Cluster ensemble and its application in gene expression analysis. Proceedings, The Second Conference on Asia-Pacific Bioinformatics. Vol.29, Dunedin, New Zealand, pp. 297-302, 2004.

[7] R.A. Johnson and D.W. Wichern. Applied Multivariate Statistical Analysis. Prentice-Hall, New Jersey, 2002.

[8] He. Binbin, T. Fang and D. Guo. Quality assessment and uncertainty handling in spatial data mining. Proc. 12th Conference on Geoinformatics. Sweden, 2004.

[9] M. Halkidi. Quality assessment and uncertainty handling in data mining process. Proc, EDBT Conference, Konstanz, Germany. 2000.

[10] M. S. Chen, J. Han, and P. S. Yu. Data mining: an overview from database perspective. IEEE Trans. On Knowledge and Data Engineering, 5(1):866—883, Dec.1996.

[11] R. Rezaee, B. P. F. Lelieveldt and J. H. C. Reiber. A new cluster validity index for the fuzzy c-mean. Pattern recognition letters, Vol(19), pp. 237-246, 1998.

[12] I. Reljin, B. Reljin and G. Jovanovic. Datasets using SOM neural networks. Journal of Automatic Control, University of Belgarde, Vol.(13), pp. 55-60, 2003.

[13] M. Ester, H.P Kriegel, and J. Sander. Algorithms for characterizaion and trend detection in spatial databases. Proceedings, 4th International Conference on Knowledge Discovery and Data Mining, NY, USA, pp. 44-50, 1998.

[14] M. Halkidi and M, Vazirgiannis. Cluster validity assessment: finding the optimal partitioning of a data set. Proc, ICDM Conference, San Jose, California, USA. pp. 187-194, 2001.

[15] Michael J. A. Berry and G. Linoff. Data mining techniques for marketing, Sales and Customer Support. John Willey & Sons, Inc, 1996.

[16] U.M. Fayyad and P. Smyth. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, Menlo Park, CA, 1996.