ISSN No. 0976-5697

Volume 5, No. 2, March 2014 (Special Issue)

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Spatial DataClustering : A Review

G.Chamundeswari
Research Scholar : JNTUK
Kakinada, India
ijarichamu@gmail.com

P. ParthaSaradhi Varma
HoD-IT and TPO : SRKR Engg.College,
Bhimavaram, India
gpsvarma@yahoo.com

Ch. Satyanaraya
Professor in the Dept. of CSE & CE: JNTUK
Kakinada, India
chsatyanarayana@yahoo.com

Abstract: Spatial data mining is the process of discovering interesting patterns from large databases that are useful. Discovering interesting patterns from the spatial information is rather a challenging and complex task when compared to extraction from the traditional datasets. This paper gives an insight into various spatial clustering algorithms, their strengths and weaknesses, and focus on issues like quality measures.

Keywords: Spatial data, clustering algorithms, statistical measures.

I. INTRODUCTION

Clustering is a process of classifying objects in such a way that the objects that belong to the same group are more similar than the objects belonging to other groups. In Data Mining the clustering algorithms play a key and challenging role to handle large amounts of spatial data to effectively retrieve useful information. The vast application of spatial clustering has resulted in the tremendous growth of this field and has attracted researchers to explore this domain.

The applications of spatial clustering is mainly in areas like detection of seismic faults of an earthquake catalogue, extract segments in remotely sensed images, to retrieve information for x-rays, weather and climate data, detection of abnormalities in medical images, creation of thematic geo-maps to name some.

The spatial clustering algorithms can be categorized as given below.

1. Hierarchical Clustering
2. Partitioning based Clustering
3. Density based Clustering
4. Graph based Clustering
5. Other Approaches

In our discussion we shall first focus on the various spatial clustering algorithms which are more appropriate for spatial clustering in section II, followed by section III giving a brief discussion on the challenges in spatial clustering, and finally section IV gives the various measures for evaluating spatial clustering

II. SPATIAL CLUSTERING ALGORITHMS

A. Hierarchical Clustering:

In this data is organized according to the proximity matrix in a hierarchical structure, which is usually represented as a binary tree or a dendogram graphically. The height of the dendogram gives the distance between the objects or clusters i.e., the custer-subcluster relationships. Hierarchical clustering is mainly classified into agglomerative and divisive.

B. Agglomerative Clustering:

These algorithms use an approach in which every point in the data set is considered as a cluster and then slowly merging the closest pair of objects [26]. AGNES and CURE [17] are the traditional algorithms that come under this category.

CURE uses random sampling and partitions. A random sample is first partitioned and partial clusters are formed. In the second pass partial clusters are clustered to yield the desired clusters. CURE is robust to outliers and identifies non-spherical clusters also. It also handles variable sized clusters.

C. Divisive Clustering:

Is one in which the whole of the dataset is taken as a single cluster initially and then successively split or divide the cluster until all the clusters are singleton clusters. MONA, DIANA [4] belongs to this category.

Earlier algorithms like AGNES and DIANA suffer from the use of over simplified measures to split or merge the clusters, hence tend to produce erroneous clusters. Whereas, CURE and CHAMELEON use a more complex better method for splitting and merging [9].

BIRCH [27] is an integrated hierarchical clustering method which initially performs clustering on the data to compress them into clustering features before using iterative relocation to improve clustering quality.

The hierarchical clustering algorithms are applicable to all types of data. They give flexibility at the level of granularity. But these algorithms lack robustness and are sensitive to outliers and nose.

D. Partitioning Based Clustering:

In this approach the dataset containing N objects is partitioned into k-partitions. These objects are assigned to k clusters on some distance measures like Euclidean distance etc., and then use mean or median to calculate the nearest distance from the center. K-MEANS [11], K-MEDIODS [1], CLARANS [5] are some to name in this category.

In K-MEANS algorithm the number of clusters and the centroids are to be initialized. K-means assigns each object to

CONFERENCE PAPER
Two day National Conference on Advanced Trends and Challenges
in Computer Science and Applications
Organized by: Shree Vishnu Engineering College for Women, Bhimavaram A.P.
Schedule: 18-19 March 2014

its nearest center forming a new set of clusters. All the centers of these new clusters are recomputed by taking the mean of all the objects in each cluster [27]. The objective criterion used in this algorithm is a squared error function defined as

$$E = \sum_{i=1}^{k} \sum_{x \in c_i} |x - m_i|^2$$

Where, x is the point in space representing the given object, and m_i, the mean of the cluster C_i.

Partitioning algorithms are general in usage, simple and have less time complexity. On the other hand the efficiency of the algorithms depends on the number of clusters and the centroids initialization. Efficiency of these algorithms is low and they are sensitive to outliers and noise.

E. Density Based Clustering:

The dense regions of objects that are surrounded by low density regions are clustered [5]. The idea is that for every point of cluster, the neighborhood of the given radius has to contain at least a minimum number of points [1]. The grid based density clusters divide the data space into grid cells and then clusters are formed according to the densities. DBSCAN, DENCLUE, CLIQUE, STING [12], Wave clusters belong to this category.

DBSCAN makes use of the density of the data points within a region to discover clusters. It is very sensitive to two input parameters *minpts* and *Eps* the radius. OPTICS overcomes this difficulty by computing and augmented cluster ordering for a wide range of parameter settings which gives users the flexibility of interactive cluster analysis [27]. However DBSCAN and OPTICS cannot handle high dimensional clustering. STING [19] explores statistical information stored in the grid cells, whereas Wave cluster clusters objects using a wavelet transform method. CLIQUE represents a grid based approach for clustering high dimensional data space [4].

The density based algorithms can handle different sizes and arbitrary shaped clusters very efficiently with less time complexity. But cannot handle varied density clusters. These algorithms are dependent on the choice of the threshold values. The grid cells in grid based clustering are rectangular and hence find difficulty in detecting clusters which are spherical in shape.

F. Graph Based Algorithms:

In these algorithms the objects are represented as nodes in the graph and the relationship between the objects are represented by weights of the edges. These algorithms use approaches [5] like

a. Sparsification – in which the proximity graph is sparsified to retain the connections of an object that are its nearest neighbors.
b. Defining the similarity measures between the objects based on the nearest neighbors.
c. Defining the core objects, so as to cluster the objects around a core, so that clusters of different sizes and shapes can be found.
d. Providing more information in the graph, so that decision of whether two similar clusters can be merged or not.

Shared Nearest Neighbor (SNN), CHAMELEON [27], KNN are some that can be used for spatial clustering.

The strength of these algorithms is they can deal with noise and outliers [3]. They can also detect clusters of different sizes and shapes. The core object concept adds flexibility though these algorithms cannot cluster all the objects.

G. Other Approaches:

There are other approaches that are more popularly used in spatial clustering

H. Mixture Model Based Clustering:

Many algorithms use a model for the clusters and find the best fit model for the data. Not all clustering algorithms are suitable for all types of data. Hence model fitting is a critical and an effective method which is finding much scope in today's real time applications.

Mixture models are statistical models that model the data using a number of statistical distributions [5]. This uses an approach known as the maximum likelihood estimation (MLE) to estimate the parameters.

One of the best algorithms for spatial clustering is the EM (Expectation Maximization) algorithm which uses the MLE. A mixture model in terms of Gaussian distribution is an effective algorithm for spatial clustering [16]. A d-dimensional Gaussian distribution representing a cluster C_i is clustered by the mean of that cluster and a $d \times d$ covariance matrix [27]. An iterative process performs the expectation step, where in the probability of each object belonging to each distribution (class) is calculated. Then in the maximization step the new estimates of the parameters that maximize the expected likelihood are found until the parameters do not change. The EM clustering algorithm tries to maximize the log likelihood of the mixture model as

$$E = \sum_{x \in D} \log P(x)$$

Where, P(x) is the probability density function.

These algorithms find clusters of various sizes and of elliptical shapes. The complexity of these algorithms is very less. On the other side, they cannot handle clusters with few objects and points that are collinear.

I. Fuzzy Clustering:

Unlike the other clustering methods, here the fuzzy set theory is used, where the uncertainty in the spatial clustering is overcome.

The Fuzzy C-Means (FCM) is one of the most popular one of this kind. It was first introduced by Dunn [15]. The iterative process produces clusters by minimizing the weighted within group sum of objective function. It works well with images. The major drawback is it is sensitive to noise and does not incorporate any kind of information regarding the spatial context. Many enhanced algorithms were proposed to overcome these drawbacks keeping in view the spatial structures. A geometrically guided FCM was proposed [14] based on semi-supervised FCM for multivariate image segmentation (GG-FCM). Similarly [13] proposed an algorithm called penalized FCM (PFCM) for image segmentation inspired by the neighboring EM (NEM) [16]. Other approaches were proposed to incorporate regularization terms in FCM called adaptive FCM (AFCM) to overcome noise sensitivity.

J. ANN Based Clustering:

ANNs are biologically inspired networks. In NN clustering the active neurons reinforce the neighboring regions by suppressing the activities of all other neurons. Algorithms like SOFM, ART, MLPs are efficient [22]. They

CONFERENCE PAPER
Two day National Conference on Advanced Trends and Challenges in Computer Science and Applications
Organized by: Shree Vishnu Engineering College for Women, Bhimavaram A.P.
Schedule: 18-19 March 2014

are used in various applications like medical diagnosis, segmentation of images, sonar/radar detection etc.,

MLPs (Multilayer Perceptrons) are popular for practical applications like pattern recognition. This topology uses back propagation algorithm, based on steepest descent method in upgrading connection weights.

SOFMs represent high dimensional input patterns which can be visualized as a two-dimensional lattice structure [23][24]. In this topology the input patterns are connected to the other layers via adaptable weights during the training process. The major disadvantage being over training.

K. Evolutionary algorithms for Clustering:

EAs fall into the category of "generate and test" algorithms. They are stochastic, population-based algorithms. Variation operators (recombination and mutation) create the necessary diversity and thereby facilitate novelty. Genetic algorithms (GA) [8] are one such category. Initially, the population of randomly generated chromosomes contains genes with allele values are initialized with random candidate solution. Evaluate each candidate. An iterative process starts with selecting the parents, recombine pairs of parents, mutate the resulting offspring, evaluate new candidates and finally select individuals for the next generation until the termination condition is satisfied [2].

GKA is one which is combination of k-means and GA that could find global optimum [10].

III. CHALLENGES IN SPATIAL CLUSTERING

A spatial object has some special characteristic of having spatial relationship, it occupies some space. These objects possess different shapes like points, lines, polygons in mathematical sense or latitudes, longitudes for rivers, roads in the geographical sense. The relationships like similarity or dissimilarity is a complex task for evaluation.

The efficiency of the spatial clustering depends on the large dataset. The characteristics of the spatial data are the types of attributes, the number of dimensions and, the amount of noise and outliers [27].

The major challenges in spatial clustering being

a. The uncertainty in boundaries, regions, structure, density or shapes of the spatial data.
b. The clustering process must be more reliable and efficient to handle large sized data and dimensionality.
c. Preprocessing of spatial data like cleaning, feature selection, data transformation, etc.,
d. Improving the computational efficiency.
e. Incorporate effective visualization techniques.

IV. EVALUATION OF SPATIAL CLUSTERS

Spatial cluster evaluation measures are classified into Supervised and Unsupervised .

a. ***Supervised***: Supervised measure (External Indices) is the discovery of cluster structure that matches some external structures. These are of two kinds classification oriented and similarity oriented [5].

In the *classification oriented* measures entropy, recall, precision, purity, F-measure are some commonly used ones.

(a). Entropy: For each cluster, the class distribution of the data is calculated first, we compute the probability that a member of cluster i belongs to cluster j as $p_{ij} = m_{ij} / m_i$, where m_i is the number of objects in cluster *i* and m_{ij} is the number of objects of class *j* in cluster *i*. The entropy of each cluster *i* is given by $e_i = -\sum_{j=1}^{L} pij \log_2 pij$, where L is the number of classes. The total entropy $\boldsymbol{e} = \sum_{i=1}^{K} \frac{m_i}{m} \boldsymbol{e_i}$, where K is the number of clusters and *m* is the total number of data points.

(b). Recall: It is the measure of a cluster containing all the objects of a specified class. The recall of cluster *i* with respect to class *j* is ***recall (i,j) = m_{ij} / m_j***, where, m_j is the number of objects in class *j*.

(c). Precision: It is the fraction of a cluster that consists of objects of a specified class. The precision of cluster *i* with respect to class *j* is ***precision (i,j) = p_{ij}***.

(d). Purity: It is the measure of a cluster containing objects of single class. ***Purity*** = $\sum_{i=1}^{K} \frac{m_i}{m} \boldsymbol{p_i}$

(e). F-measure: It is a combination of both precision and recall, which measures the extent to which a cluster contains only objects of a particular class and all objects of that cluster.

F(i,j)=(2 X precision(i,j) X recall(i,j))/ (precision (i,j)+recall(i,j).

b. Unsupervised: Unsupervised is the goodness of a clustering structure without respect to external information [5]. The unsupervised measure (Internal Indices) of cluster validity is of two kinds, measure of cluster cohesion (compactness) and measure of cluster separation (isolation).

The general expression for the overall cluster validity for a set of k-clusters is the weighted sum of the validity of individual cluster.

$$\textbf{Overall validity} = \sum_{i=1}^{k} W_i . \boldsymbol{Validity}(C_i)$$

Where, Validity (C_i) can be cohesion, separation of both.

$$\textbf{Cohesion } (C_i) = \sum_{x \in C_i, y \in C_i} \boldsymbol{Proximity}(x, C_i)$$

$$\textbf{Separation } (c_i, c_j) = \sum_{x \in C_i, y \in C_j} \boldsymbol{Proximity}(x, y)$$

Where, the proximity can be any of the distance measures. The Squared Euclidean distance, Minkowski distance, Mahalanobis, distance are often used metrics.

The Silhouette co-efficient combines both Cohesion and Separation.

$$\boldsymbol{Si = (b_i - a_i) / max (a_i - b_i)}$$ for the i[th] object

Where a_i, is the average distance from i to all other objects in its cluster and b_i is the average minimum distance from i to all the objects in another given cluster.

The other quality measures [1] are :

Variance of the spatial data set X

$$\sigma_x^p = \frac{1}{n} \sum_{k=1}^{n} (x_x^p - \bar{x}^p)^2$$ of the p[th] dimension

Where $\bar{x}^p = \frac{1}{n} \sum_{k=1}^{n} x_k , \forall x_k \in X$ of the p[th] dimension

Variance of the cluster is defined by

$$\sigma_x^p = \frac{\sum_{k=1}^{n_i} (x_x^p - v_i^p)^2}{n_i}$$

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

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

The average Compactness of C clusters is given by

$$\textbf{Comp} = \frac{\sigma}{c}$$

The average compactness of C clusters

$$scat\_comp = \textbf{comp} / ||\sigma (x)||$$

The more compact the clusters are the smaller is the *scat_comp.*

CONFERENCE PAPER
Two day National Conference on Advanced Trends and Challenges in Computer Science and Applications
Organized by: Shree Vishnu Engineering College for Women, Bhimavaram A.P.
Schedule: 18-19 March 2014

The distance between the clusters is the average distance between the center of the specified clusters.

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

The larger d is, the more are the clusters separated.

CD = ***scat_comp/d***

Small values of CD indicate all the clusters in the clustering scheme are overall compact and well separated [1].

V. CONCLUSION

In this survey paper we have presented an overview of the various useful spatial clustering algorithms. We have initially categorized the spatial clustering algorithms into Hierarchical, Partitioning, Density based, Graph based, and other approaches. We have also discussed the general uses and drawbacks of some algorithms in each of these categories.

We then focused on the challenging issues and discussed about the various validity measures which will be very useful while selecting an algorithm for better quality clustering in a particular application or to develop algorithms and testing their quality with these validity measures.

VI. REFERENCES

[1] Jingke Xi, "Spatial Clustering Algorithms and Quality Assessment," Proc. 2009 International Conf. on AI, DOI 10.1109/JCAI.2009.162.

[2] Rui Xu, "Survey on Clustering Algorithms," IEEE Transactions on Neural Networks, vol.16, No.3, 2005.

[3] Harleen Kaur, Ritu Chauhan and M. Afshar Alam, "Spatial Clustering Algorithm using R-tree," Journal of Computing, vol.3, issue2 Feb 2011, ISSN 2151-9617.

[4] Martin Ester, Hans-Peter Kriegel, Jorg.Sander, "Algorithms and Applications for Spatial Data Mining," Geographic Data Mining and Knowledge Recovery, Research Monogram in GIS, Taylor & Francis, 2001.

[5] Pand-Ning Tan, Vipin Kumar, Introduction to Data Mining, Pearson Education.Inc.

[6] L.Kaufman and P.Rousseeuw, "Finding Goups in Data: An introduction to cluster analysis," Wiley series in probability and statistics.

[7] P.J.Angeline, "Adaptive and Self-Adaptive Evolutionary Computations," Computaional Intelligence, A Dynamic System Prospective, Piscataway, IEEE Press, 1995.

[8] L.O.Hall, I.B.Ozyart, J.C.Bexdek, "Clustering with a Genetically Optimized Approach," IEEE Transactions on Evolutionary Computation 3(2),1999.

[9] J.Han and M.Kamber, Data Mining: Concepts and Techniques, Morgan Kauffman Publishers,2000.

[10] G.Babu and M.Murthy, " A Near Optimal Initial Seed Value Selection in K-Means Algorithm using a Genetic Algorithm," Pattern Recognition Lett., vol14,no.10.1993.

[11] Tapas Konungo et.al, "An efficient K-Means Clustering Algorithm : Analysis of Machine Intelligence," vol24,n0.7,July 2002.

[12] J.C.Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, New York, Plenum Press 1981.

[13] Y.Yang, Ch.Zheng and P.Lin, "Fuzzy C Means Clustering Algorithm with a Novel Penalty Term for Image Segmentation," Opto-Electronics Review 13,No.4,2005.

[14] J.C.Noordam, W.H.A.M.Vanden Broek, and L.M.C.Buyden, "Geometrically Guided Fuzzy C-Means Clustering for Multivariate Image Segmentation," Proc. Int.Conf. on Pattern Recognition1,2000.

[15] J.C.Duan, "A Fuzzy Relative of the ISODATA Process and its use in Detecting Compact Well Seperated Clusters," J.Cybernetics 3,1974.

[16] C.Ambroise and G.Govert, "Convergence of an EM-Type Algorithm for Spatial Clustering," Pattern Recognition Letters 9, 1998.

[17] S.Gupta, R.Rastogi, and K.Shim, "CURE : An Efficient Clustering Algorithm for Large Databases", Proc. 1998, ACM Spatial Interest Group on Management of Data, 1998.

[18] Xiang Lai-Sheng, Guo Ya-jun, Lan Tian, "Topological Cluster : A Generalized View for Density-based Sptial Clustering," ICMSE 2007.

[19] W.Wang and J.Yang, R.Muntx, "STING : A Statistical Information Grid Approach to Spatial Data Minin," Proc. 23rd Conf. Very Large Databases,1997.

[20] Hiuneburg.A, Kein A.D, " An Effiecient Approach to Clustering in Large Multimedia Databases with Knowledge Discovery and Data Mining (KDD98), NewYork, NY,USA,1998.

[21] Samet.H, The Design and Analysis of Spatial Data Structures, Addison-Wesley.1990.

[22] Quintanilla Dominguez, "Improvement for Detection of Microcalcifications Through Clustering Algorithms and Artificial Neural Networks," EURASIP Journal on Advances in Signal Processing 2011, Springer.

[23] T.Kohonen, "The Self-Organizing Maps," 3rd ed., New York, Springer-Verlag, 2001.

[24] T.Kohonen, "The Self-Organizing Map," Proc. IEEE, Vol 78,no.9.1990.

[25] S.Ramaswamu, R.Rastogi, and K.Shim, "Efficient Algorithm for Mining Outliers from Large Datasets," Proc. of ACM-SIGMOD Int.Conf. on Management of Data, ACM Press,2000.

[26] A.K.Jain, M.N.Murthy, P.J.Flynn, "Data Clustering : A Review," ACM Computing Surveys, vol.31,No.3,1999.

[27] J.Han, M.Kamber and A.K.H.Tung, "Spatial Clustering Methods in Data Mining : A Survey," H.Miller and J.Han (eds.), Geographic Data Mining and Knowledge Discovery, Taylor and Francis, 2001.

CONFERENCE PAPER
Two day National Conference on Advanced Trends and Challenges in Computer Science and Applications
Organized by: Shree Vishnu Engineering College for Women, Bhimavaram A.P.
Schedule: 18-19 March 2014