Spatial Data Analysis using DBSCAN Method and KNN classification

I Putu Sugi Almantara[1*], Ni Wayan Sri Ariyani[2], Ida Bagus Alit Swamardika[3]
[1]Department of Electrical and Computer Engineering, Post Graduate Program, Udayana University
[2,3]Department of Electrical and Computer Engineering, Udayana University
*sugik.almantara@gmail.com

Abstract Spatial Data Clustering is one of the most important technical techniques used to obtain information about knowledge about number boundaries in databases from various applications. This technique can determine groups of forms that cannot be arranged and can be used effectively with a budget. Exploring interesting and useful spatial boundary patterns is more difficult to extract traditional and categorical numerical polymers because of the difficulty of species, the relationship between autocorrelation of spatial boundaries. One of the pioneering techniques in the development of facial and technical grouping technologies is DBSCAN. This technique can determine groups of shapes that cannot be arranged and can be arranged in an effective way. the groups that have already received the next classification process are carried out in order to obtain information on the classes already formed. The K-Nearest Neighbour classification technique is based on learning by analogy. When there is new data, K-Nearest Neighbor will look for a class of data from the learning sample that is closest to the new data. This closeness can be defined using the Euclidean Distance calculation method.

***Index Terms* — Spatial Data Clustering, DBSCAN and K-Nearest Neighbour.**

I. Introduction

DATA mining is a step in the Knowledge Discoveryin Database (KDD) process which consists of the application of data analysis and the discovery of algorithms that produce certain enumerations of patterns in data. [1]. known but potentially useful from large spatial data sets. Excavating interesting and useful patterns from spatial data sets is more difficult than extracting traditional and categorical numerical data patterns due to the complexity of the types, relationships and autocorrelation of these spatial data sets [2]. Most recent research on spatial data uses clustering techniques due to the nature of the data. Clustering is the process of grouping large amounts of data into several classes according to their respective characteristics. The most efficient clustering algorithm to determine clusters in data with different densities is the density based clustering algorithm [3]. DBSCAN is one of the examples of the pioneering development of grouping techniques based on density or which is commonly known by the term density-based clustering [4]. Research using the DBSCAN Method has been conducted several times before. Classification is one of the many tasks in Data Mining done and has been implemented in the fields of statistics, pattern recognition, decision making, machine learning, neural networks and others. Classification is a supervised learning method. False one of the most popular classification methods is K-Nearest Neighbor. KNN (K-Nearest Neighbor) is one of many classification algorithm that is often used because of simplicity and ease in its application. This algorithm performs the process classification based on the distance between testing data with training data. This algorithm is trying classifying new data for which class is unknown by selecting the number of k data that is located closest to the new data. The advantage of using an algorithm KNN is the speed of the training process, it's easy studied and used, can be applied to data which has a lot of noise (noisy), and very effective if used in the classification process with large dataset. This study discusses the analysis of spatial data that will be grouped using the DBSCAN method and classifies the results of grouping data into data that already exists in the database using the KNN method.

II. Literature study

A. Data Mining

Data mining is the analysis phase of the knowledge process discovery in database. Data mining is a process from analyzing data from different perspectives and summarize it into information. Data Mining is one of the important techniques in searching knowledge in a collection of digital data (knowledge mining from data). The results of this search can be used to predict the future and find trends based on patterns and data relations. The application of data mining requires a variety of software for analyzing data for searching relationship data and patterns that can be used for make accurate predictions.

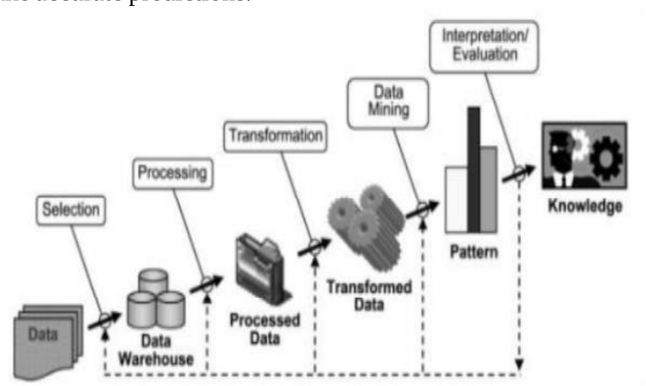

FIG. 1 DATA MINING STAGES

1) Data Selections

Select or segment data according tosome criteria.

2) Pre-Processing and Cleaning Data

Data cleaning phase is done by delete certain information that is deemed not necessary and biased because it slows down the query.

3) Transformation

Data is not only transferred but transformed in layers can be added like layers demographic commonly used in market research.

4) Data Mining

This stage deals with the extraction of patterns from data a pattern can be defined as a set facts (data).

5) Interpretation

The patterns defined by the system are interpreted become knowledge which can then be used to support decision making.

B. Clustering Method

Clustering is a process of participant grouping the data into classes or cluster the cluster based on a similarity of the attribute -the attribute among groups the data. Clustering technique aims to divide a group of objects into groups whether or not groups can be determined based on the clustering method used later, so that the same object is in the same group based on several similarity functions, identifying the same sub-population in the data. Clustering refers to grouping records / lines, observations or cases into classes of similar objects. Clustering is often done as a first step in the data mining process, with the resulting cluster being used as further input to different stages, such as neural networks. Due to the very large size of the database at this time then applying the first clustering analysis will greatly assist the data mining process. One scalability technique for cluster algorithms is to summarize data in dense regions. Because clusters correspond to dense regions, records in this region can be summarized collectively through summarized representations called cluster features (CF).

C. Density-Based Spatial Clustering of Application with Noise (DBSCAN)

DBSCAN (Density-Based Spatial Clustering of Application with Noise) is one of the pioneers in the development of grouping techniques based on density or commonly known as density based clustering. DBSCAN is a clustering method that builds areas based on connected density (desity-connected). Each object of an area radius (cluster) must contain at least a minimum amount of data. All objects that are not included in the cluster are considered as noise. DBSCAN is an algorithm that builds high density areas into clusters and finds clusters in arbitrary form in spatial databases containing noise in them. The key to complexity-based grouping for each cluster object is the minimum environment radius for the minimum amount of data. DBSCAN will determine its own number of clusters to be generated so that it does not require the number of clusters desired but requires 2 other inputs, Minpts are a minimum of many items / objects in a cluster and EPS is the value for the distance between items on which the formation of a neighborhood from a point item. DBSCAN scans for clusters by checking the neighborhood (eps-neighborhood) of each point in the database. If neighborhood of point x contains more than MinPts, the new cluster is a core object. Then DBSCAN iteratively collects directly density-reachable objects from the core object, which may involve the joining of several density-reachable clusters.

D. Classification Method

Classification is a learning function that maps (classifies) an element (item) of data into one of several classes that have been defined. According to Han and Kamber (2006) in general, classification consists of two stages. The first stage, namely learning (learning process), is a model created to describe a set of classes or data concepts that have been predetermined. The model is built by analyzing records in a database that is described in the form of attributes. Each record is assumed to belong to a predetermined class, called a class attribute. The model itself can be an IF-THEN rule, decisiontree, mathematical formula or neural network. The classification method used to map the object above into several classes that have been defined is the KNN (K-Nearest Neighbor) classification method.

E. K-Nearest Neighbor

The K-Nearest Neighbor classification technique is based on learning by analogy. The learning sample is a group of data with n numeric attribute dimensions. All learning samples are stored segmented based on the characteristics of each sample data. This segmentation is known as a data class. When there is new data, K-Nearest Neighbor will look for class data from the learning sample that is closest to the new data. This closeness can be defined using the Euclidean Distance, City Block, Minkowski distance calculation method, and other distance calculation methods. K-NN is used to identify an object or class that has not been defined because the KNN classification does not recalculate all new classes or existing classes.

F. *Euclidean Distance*

Euclidean distance is a calculation of the distance of 2 points in Euclidean space. Euclidean space was introduced by Euclid, a mathematician from Greece around 300 B.C.E. to study the relationship between angle and distance. Euclidean is related to the Pythagorean Theorem and is usually applied in 1, 2 and 3 dimensions. But it is also simple if applied to higher dimensions.

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

FIG. 2 EUCLIDEAN DISTANCE FORMULA 2 DIMENSIONS

This research will use 2-dimensional euclidean distance calculation which will calculate the new spatial data then cluster and classify it.

III. DISCUSSION AND RESULT

There is a new spatial data that has not yet got a group and the class has not been identified which forms a pattern like the following picture.

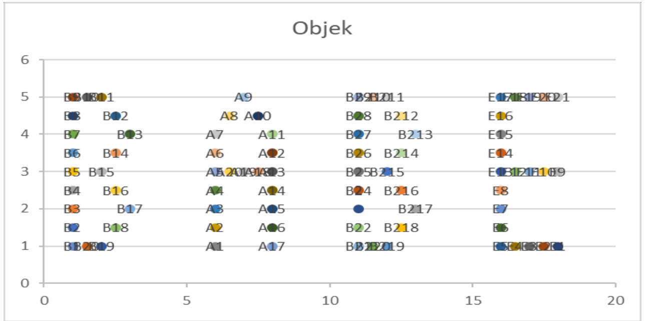

FIG. 3 SPATIAL DATA OBJECT

Figure 3 above is the objects that have not been grouped into several groups that have the same similarity in each group. The clustering method used to divide the group is the DBSCAN method, where the Minpts parameter used is 5 and the EPS used is 3. To calculate the distance of the core point to another point, the Euclidean Distance formula can be used. The calculation in the first iteration chooses point B20 (1.5.1) as the core point to be the calculation. And so on for each point there is calculated the distance from the core point B20 (1.5.1).

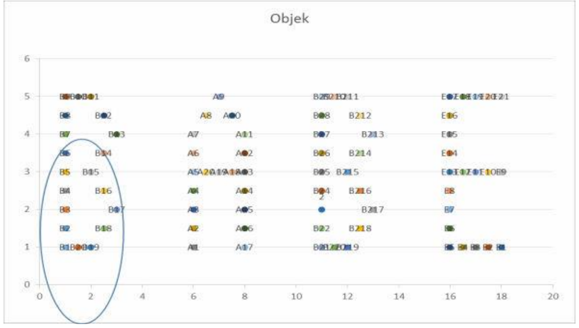

FIG. 4 RESULTS OF THE FIRST ITERATION

The points that fulfill the requirements are 13 points, this number also meets the number of members because the Minpts of this clustering parameter is 5. And to select the next core point, choose the greatest distance or equal to Eps but if there is a point that has become a core point in the previous iteration then it cannot be used as a core point anymore and look for the next largest sequence. From the result of the first iteration that is circled in blue with the cluster members in it, the second is the calculation of the second iteration where the results point that meets the requirements there is point B14 (2.5, 3.5) which has the largest distance and point B14 will be used as the next core point.

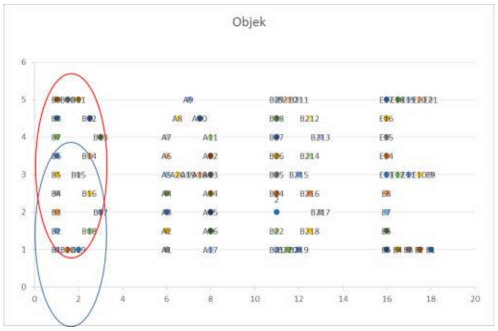

FIG. 5 RESULTS OF THE SECOND ITERATION

From the graph in Figure 5 iteration 2 results above obtained group results of cluster 1 because there are no cluster members in the Eps 3 area then the next will be followed by selecting new core points because there are still object points that do not get the cluster yet. The selected core point is A1 (6.1).

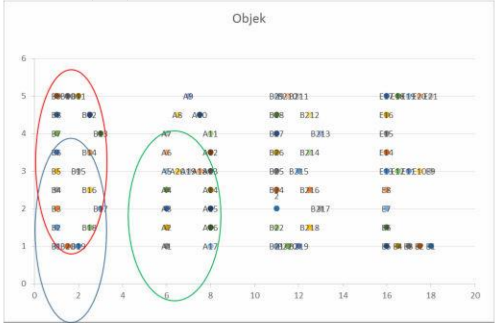

FIG. 6 RESULTS OF THE THIRD ITERATION

From the result of the third iteration in figure 6 circled in green with the cluster members in it, the next is the fourth iteration calculation where the results point that meets the requirements is point A7 (6.4) which has the largest distance and point A7 will be used as a core point next. Iteration continues until all spatial data gets a group.

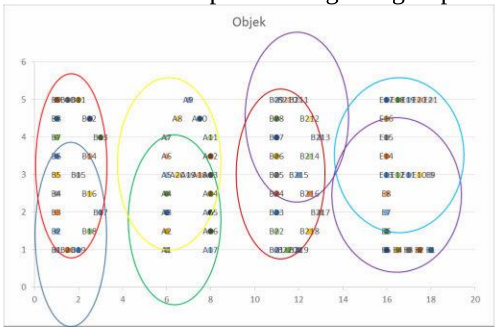

FIG. 7 RESULTS OF THE END ITERATION

From the graph in Figure 7 end iteration results obtained group 4 cluster results because there are no cluster members in the Eps 3 area and all points / objects have their respective groups / clusters and in this object data there are no members who have groups or noise. Then define the relationship between the core points to be used as a group / cluster with the condition that several core points are density-connected.

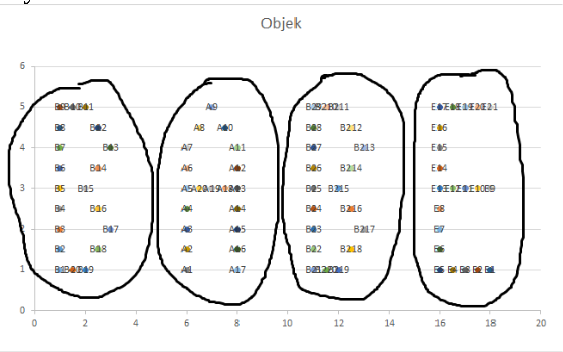

FIG. 7 RESULTS OF THE END ITERATION

Core points B20 with B14 will be the first group, A1 with A7 will be the second group, B211 with B24 will be the third group and E6 with E16 will be the fourth group. New data is classified into classes with the largest number of closest neighbors (majority). In this process is to recognize patterns that are formed by comparing the distance of each cluster with a database of letters alphabetically a to z that has been available. calculations on KNN here use the Euclidean Distance method as well as the clustering process performed on the DBSCAN method. Calculations are performed on the first and fourth clusters of each class in the database. All classes of data that have not been previously defined have been successfully classified with the KNN (K-Nearest Neighbor) method and each of these new classes can be recognized as letters according to the classification performed based on the data in the database.

IV. CONCLUSION

During the process of clustering and classification using the K-NN method, it is good to recognize an object or class that has not been defined because the KNN classification does not recalculate all new classes or existing classes. It has also been explained previously that the KNN method is a lazy learner because only new objects / classes are tested with pre-existing objects / classes and this method is also suitable for classifying objects / classes that are universal such as classification of numbers, letters, plant types, types animals and others. Maybe for further analysis can use a combination of various techniques to get a classification that is faster, efficient and accurate.

V. REFERENCES

[1] L. G. P. Suardani, I. M. A. Bhaskara and M. Sudarma, "Optimization of Feature Selection Using Genetic Algorithm with Naïve Bayes Classification for Home Improvement Recipients," *International Journal of Engineering and Emerging Technology*, pp. 66-70, 2018.

[2] I. W. S. Pramana, P. R. Iswardani and N. W. S. Aryani, "Application of Data Mining in Optimization of Hotel's Food and Beverage Costs," *International Journal of Engineering and Emerging Technology*, pp. 1-5, 2019.

[3] D. Ardiada, P. A. Ariawan and M. Sudarma, "Evaluation of Supporting Work Quality Using K- Means Algorithm," *International Journal of Engineering and Emerging Technology*, pp. 52-55, 2018.

[4] I. C. Dewi, B. Y. Gautama and P. A. Mertasana, "Analysis of Clustering for Grouping of Productive Industry by K-Medoid Method," *International Journal of Engineering and Emerging Technology*, pp. 26-30, 2017.

[5] N. M. A. Santika Devi, I. K. G. Darma Putra And I. M. Sukarsa, "Implementasi Metode Clustering Dbscan Pada Proses Pengambilan Keputusan," *Lontar Komputer*, Pp. 185-191, 2015.

[6] I. M. S. Putra, "Algoritma Dbscan (Density- Based Spatial Clustering Of Applications With Noise)," Program Studi Teknologi Informasi Fakultas Teknik Universitas Udayana, Denpasar, 2018.

[7] E. N. Suhendra, I. W. Santiyasa And A. Muliantara, "Pengaruh Pca (Principle Component Analysis) Terhadap Klasifikasi Prakiraan Hujan Harian Di Daerah Kuta Selatan Menggunakan Algoritma Knn (K-Nearest Neighbor)," *Ilmu Komputer*, Pp. 8-14, 2015.

[8] I. G. Harsemadi, M. Sudarma and N. Pramaita, "Implementasi Algoritma K-Nearest Neighbor pada Perangkat Lunak Pengelompokan Musik untuk Menentukan Suasana Hati," *Majalah Ilmiah Teknik Elektro*, pp. 15-19, 2016.

[9] I. G. A. S. Melati, L. and I. A. D. Giriantari, "Knowledge Discovery Data Akademik Untuk Prediksi Pengunduran DIri Calon Mahasiswa," *Majalah Ilmiah Teknologi Elektro*, pp. 325-331, 2018.