Analysis of DBSCAN Algorithm in Determining Epsilon Parameters Numerical Data Clustering

Herwin Simanjuntak[1], Sawaluddin[1] and Muhammad Zarlis[2]

Graduate Program of Computer Science, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia

[1] *Department of Computer Science, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia*

[2] *Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia*

Keywords: DBSCAN, Clustering, Euclidean Distance, Epsilon.

Abstract: DBSCAN algorithm (Density-Based Spatial Clustering of Applications with Noise) is one of the numerical based clustering algorithms, numerical data is used as the test for this algorithm. The DBSCAN algorithm has the disadvantage of being difficult to determine the appropriate Epsilon value in order to obtain good clustering results. In the DBSCAN algorithm, the value of epsilon is calculated based on a lot of data from the entire data that is captured. In this study a modification of the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm was carried out by determining the value of epsilon, the results obtained in the study of Euclidean Distance obtained better than the results obtained from the DBSCAN.

1 INTRODUCTION

The two main approaches are used to develop clustering methods through a partitioning approach and clustering on a hierarchical approach. Clustering partitions is data that is grouped by means of data sorted by analyzing into available clusters. Because clusters group data into several groups to have many similarities and have little in common. (Poteras, et al. 2014).

Clustering is part of an unsupervised learning method because it does not require cluster definitions first (Nisha and Kaur, 2015). In clustering, the measurement of similarity between objects is done by measuring the distance for each pair of objects. This measurement can be done using the Euclidean Distance, Manhattan Distance and Minkowski Distance methods.

Almost all well-known grouping algorithms need input parameters that are difficult to determine but possess a crucial effect on the results of grouping. In addition, for some data sets not make large parameter settings to get accurate grouping algorithms in group structures.

DBSCAN is the basic algorithm for density-based clustering techniques. One of the advantages of using these techniques is that the method doesn’t require the number of clusters to be given before or they do not make assumptions about densities or variants in clusters that might be in the data set. The basic algorithm for density-based clustering techniques is DBSCAN. This can detect groups of various shapes and sizes from large amounts of data containing noise and outliers (Shah, 2012).

2 RESEARCH METHODS

Clustering is very important for research topics of machine learning and datamining. Clustering has developed into a technique that is popular in the area of pattern identification, image processing and data mining (Aranganayagi, 2007). Classical clustering techniques such as the k-means method, partition data into k clusters (MacQueen, 1967) and are very sensitive to the initial values of each cluster center (Cuietal, 2015).

According to Tan (2006) clustering is grouping objects (data) that are based only the information

Simanjutak, H., Sawaluddin, . and Zarlis, M.
Analysis of DBSCAN Algorithm in Determining Epsilon Parameters Numerical Data Clustering.
DOI: 10.5220/0008552002200222
In *Proceedings of the International Conference on Natural Resources and Technology (ICONART 2019)*, pages 220-222
ISBN: 978-989-758-404-6

contained in the object and the relationship between these objects. This grouping of data is usually done based on the similarity of values between data (Xia et al. 2008).

Distance measurement or similarity between objects and databases including the basic principles of data grouping that carry out unsupervised learning. (Nisha, 2015). Clustering aims to make objects in one group only consist of objects that have similarities to each other and are different from objects in other groups.

2.1 DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*)

DBSCAN is one of the density-based clustering algorithms. The algorithm extends high-density regions into clusters and places irregular clusters in spatial databases with noise. This method defines clusters as the maximum set of density-connected points. DBSCAN has 2 parameters, namely Eps (maximum radius of a neighborhood) and MinPts (minimum number of points in the Eps-neighborhood of a point). The basic idea of density-based clustering is related to several new definitions:

1. The neighbor hood with the Eps radius ofan object is called the Eps neighbor of an object
2. If the Eps-neighbor hood of an object contains a minimum number of minimum points, Min Pts, then an object is called the core object
3. Given the set of objects D, theobject p is said to be directly density - reachable from the object q if p is included in the Eps - neighborhood of q and q is the object's core.

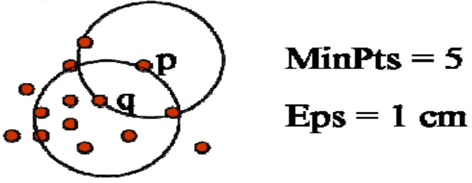

Figure 1: *Eps-neighborhood* Arthur (2010)

2.2 Euclidean Distance Method

Euclidean distance is the distance measured straight from one coordinate point to another. Although this method is less realistic, it is generally used because this method is easy to understand and easy to model. We can find applications from Euclidean distances on several conveyor models, transportation and distribution systems.

3 PROBLEMS IDENTIFICATION

From the background described above, almost all clustering algorithms need input calculations that are hard to analyze however, have an important impact on results, finding cluster size and efficiency even for large data sets. DBSCAN will detect clusters and determine how to determine epsilon parameters automatically in an accurate way to determine input parameters and find clusters with different densities on the Iris dataset by comparing measurements to that method.

4 RESULT AND DISCUSSION

At this stage, the performance of the DBSCAN is testing, Figure 2 shows a graph of clustering results measured based on Euclidean Distance values obtained from the DBSCAN algorithm.

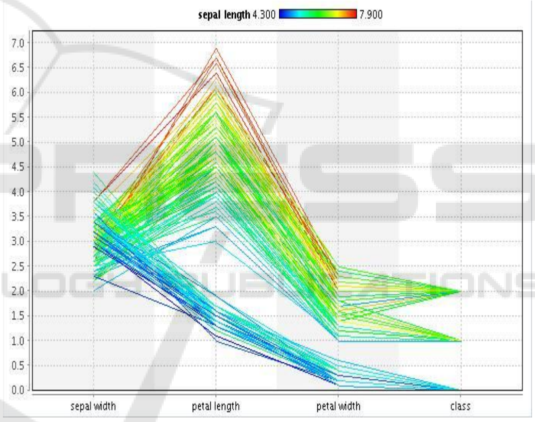

Figure 2: Graph of Test Results on The Iris Dataset

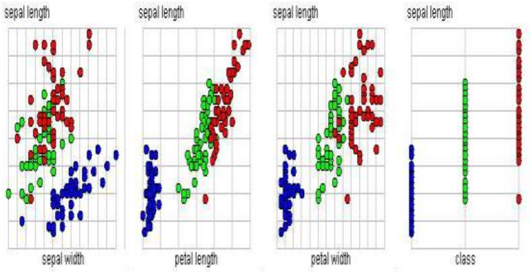

Figure 3: Test Results Cluster Plot on The Iris Dataset

At this stage, the performance of the DBSCAN algorithm is tested on the cluster distribution that can be seen in Figure 4.2 shows the plot of clustering results measured based on Euclidean

Distance values obtained from the DBSCAN algorithm.

5 CONCLUSION

The method for determining and applying Epslilon and MinPts values in the DBSCAN algorithm can be done to obtain better clustering results.The method of determining and applying the Epslilon and MinPts values in the DBSCAN algorithm will directly affect the number of clustering produced. The cluster results of the Camberra Distance method are very good, by producing a large number of clusters and the cluster neutralization produced very well by the Distance.

REFERENCES

Aranganayagi, S., Thangavel, T. 2007. Clustering categorical data using silhouette coefficient as a relocating measure. In *Proceedings of 2007 Internasional Conference on Computational Intelligence and Multimedia Application*. pp.13-17

Cui, X., Wang, F. 2015. An Improved Method for K-Means Clustering. *Proceedings of 2015 International Conference on Computational Intelligence and Communication Networks (CICN)*. pp.756-759

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, A Density-Based Algorithm for Discovering Clusters. In *2nd International Conference on Knowledge Discovery and Data Mining* (KDD-96), 1996

Fayyad, U., Shapiro, G. P., Smyth, P. 1996. From Data Mining to Knowledge Discovery in Database. *AI Magazine*: pp. 37-53

Gorunescu, F. 2011. *Data Mining: Concepts, Models and Techniques*. Springer: Berlin.

Gothai, E., Balasubramanie, P. 2010. Performance evaluation of hierarchical clustering algorithms. *Proceedings of The Internasional Conference on Communication and Computational Intelligence - 2010*. pp. 457-460.

Han, J. Kamber, M. 2006. *Data Mining: Concepts and Techniques*. 2nd Edition. Elsevier: San Francisco.

Hu, X., Liu, L., Qiu, N., Yang, D., Li, M. 2018. A MapReduce-based improvement algorithm for DBSCAN. *Journal of Algorithms and Computational Technology*, *12*(1), 53-61

MacQueen, J. 1967. *Some Methods for Classification and Analysis of Multivariate Statistics and Probability*. University of California Press, Berkeley. California.

Gaonkar, M. N., Sawant, K. 2013. AutoEpsDBSCAN: DBSCAN with Eps automatic for large dataset. *International Journal on Advanced Computer Theory and Engineering*, 2(2), 11-16.

Nisha, Kaur, P. J. 2015. Cluster Quality Based Performance Evaluation of Hierarchical Clustering Method. *Proceedings of 2015 1st International Conference on Next Computing Technologies*. pp. 649-653.

Poteras, C. M., Mihăescu, M. C., Mocanu, M. 2014. An optimized version of the k-means clustering algorithm. *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems*, pp. 695–699.

Rokach, L., Maimon, O. 2005. *Data Mining and Knowledge Discovery Handbook*. Springer: Tel Aviv.

Shah, G. H., Bhensdadia, C. K., Ganatra, A. P. 2012. An empirical evaluation of density-based clustering techniques. *International Journal of Soft Computing and Engineering (IJSCE) ISSN, 22312307*, 216-223.

Tan, P. N., Steinbach, M., Kumar, V. 2006, *Introduction to Data Mining (Vol. 1)*, Pearson Addison Wesley: Boston.