CSE 454 – DATA MINING
HOMEWORK #01
GÖKHAN HAS – 161044067

DBSCAN algorithm, two or two data points their neighborhoods with each other in multidimensional space. It is based on uncovering. Database, spatial mostly spatial since it deals with the point of view used in the analysis of data.

Density-based clustering has required a number of new terms.

- The neighborhood of an object within the ϵpsilon diameter is called the-neighborhood of that object.
- If an object contains ϵpsilon - MinPts objects whose neighborhood is at least a minimum number, then this object is called a seed object.

Dataset

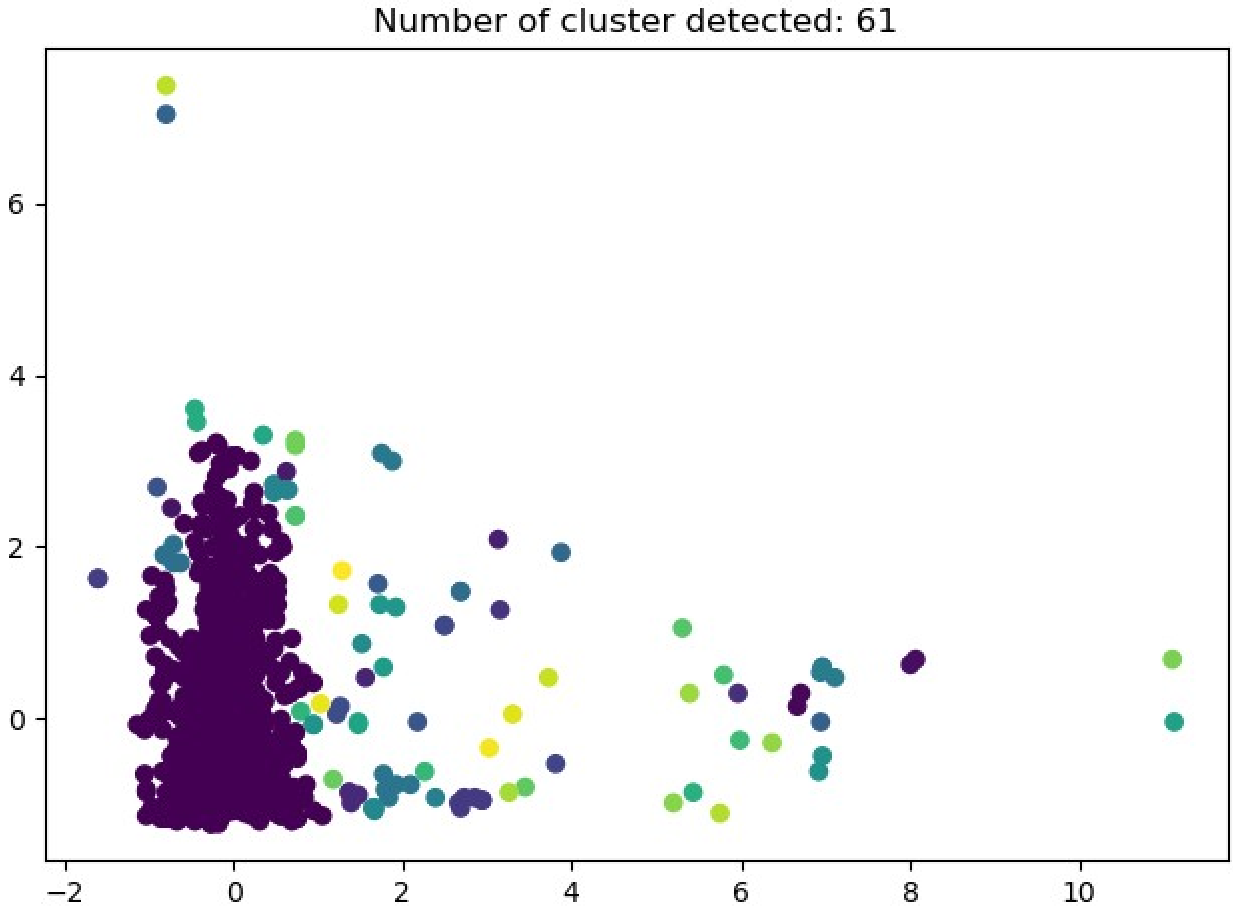
The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. For more details, consult the reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones).

https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009

Parameters Analysis

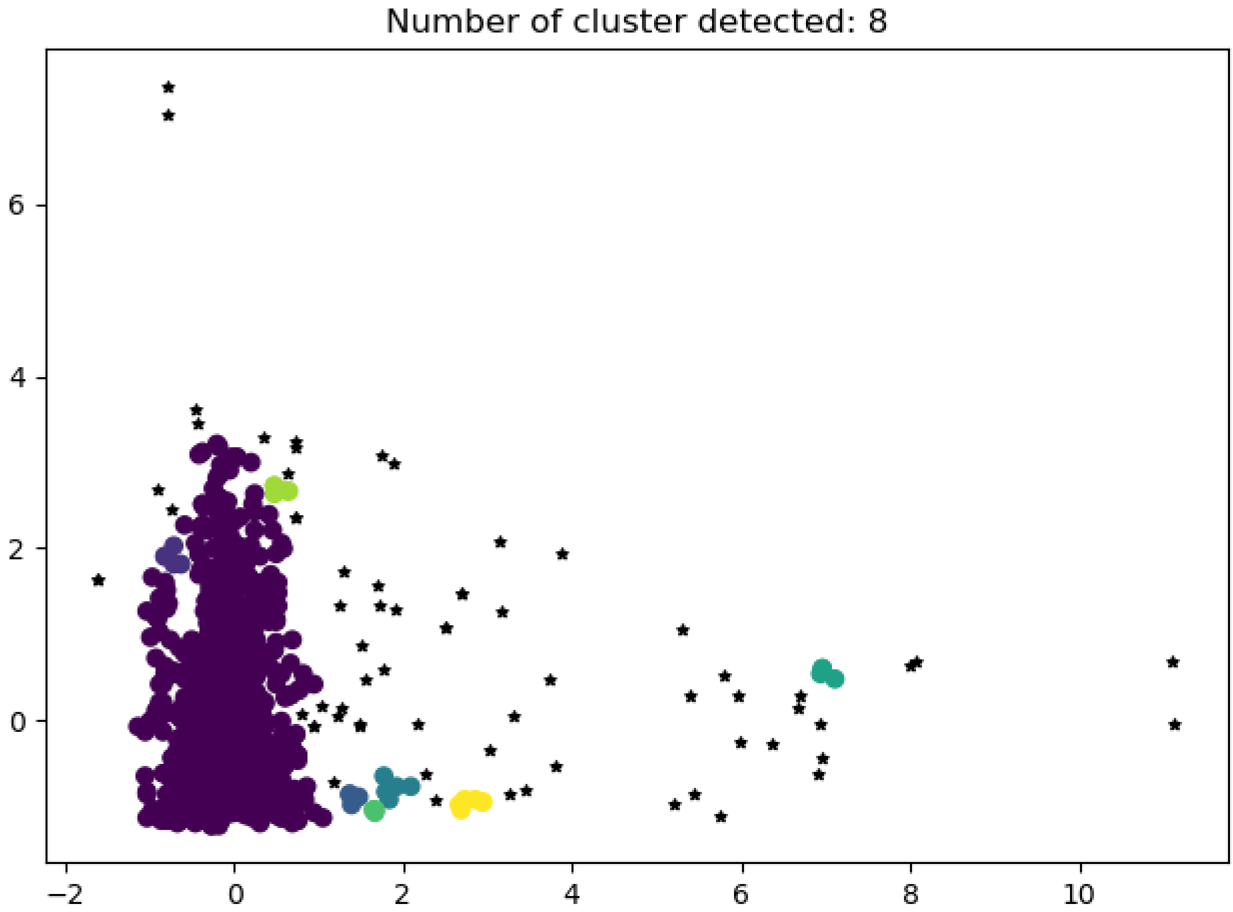
1-) Epsilon = 0.2, Minpoints = 1
# of cluster = 61



In the first analysis, the minpoint variable is given a. Noise values are expected in this analysis. Because it is guaranteed that even at least one point can be taken into a separate cluster. Since the objects in the dataset I use are close to each other, a main cluster stands out with its purple color. Epsilon value was kept constant at 0.2.

2-) Epsilon = 0.2, Minpoints = 3

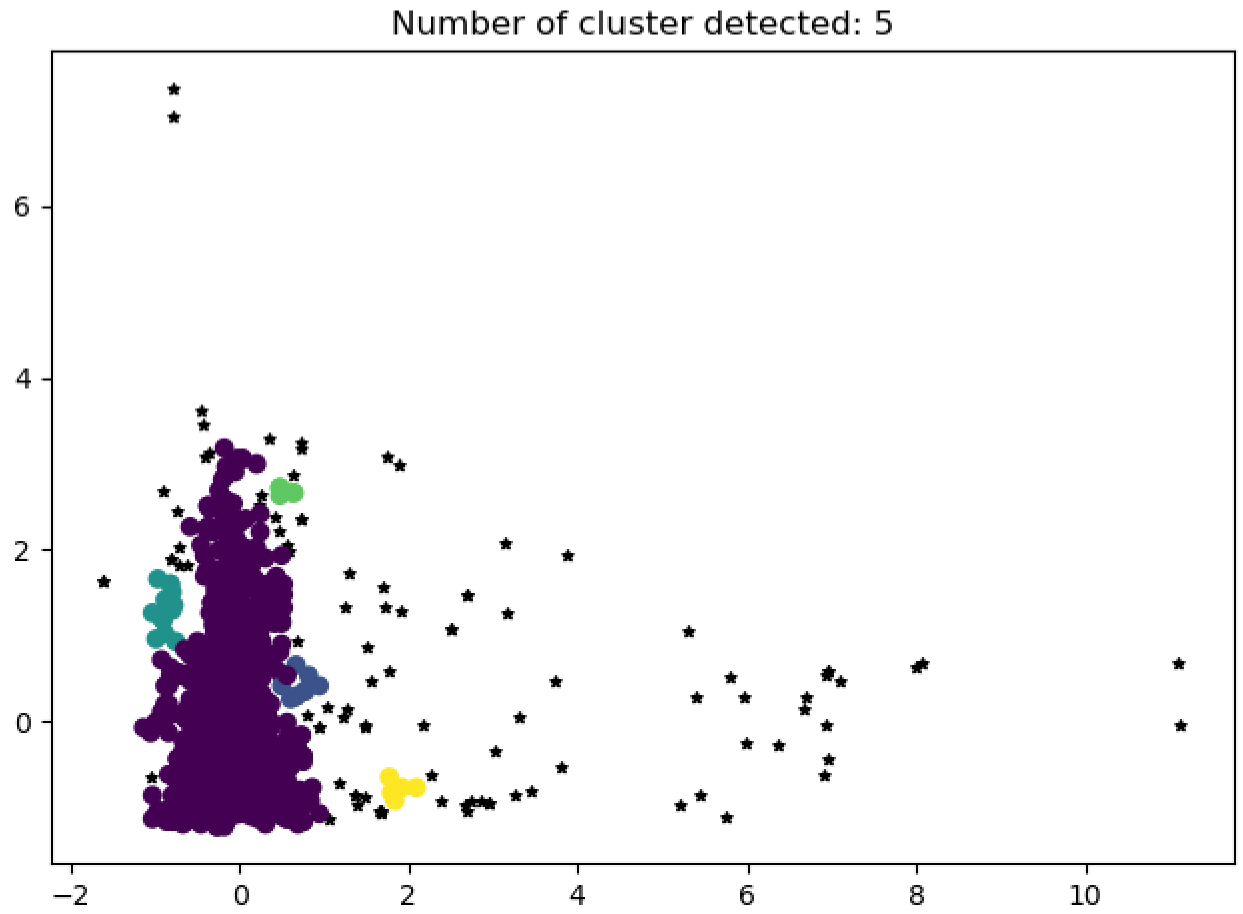
# of cluster = 8



When I increase the minpoint variable from 1 to 3, I expect the number of clusters to decrease. This is already seen in the figures. The number of clusters from 61 has dropped to 8. Because now, at least 3 points must join for a cluster. It is now seen in black on objects that have noise.

3-) Epsilon = 0.2, Minpoints = 5

# of cluster = 5

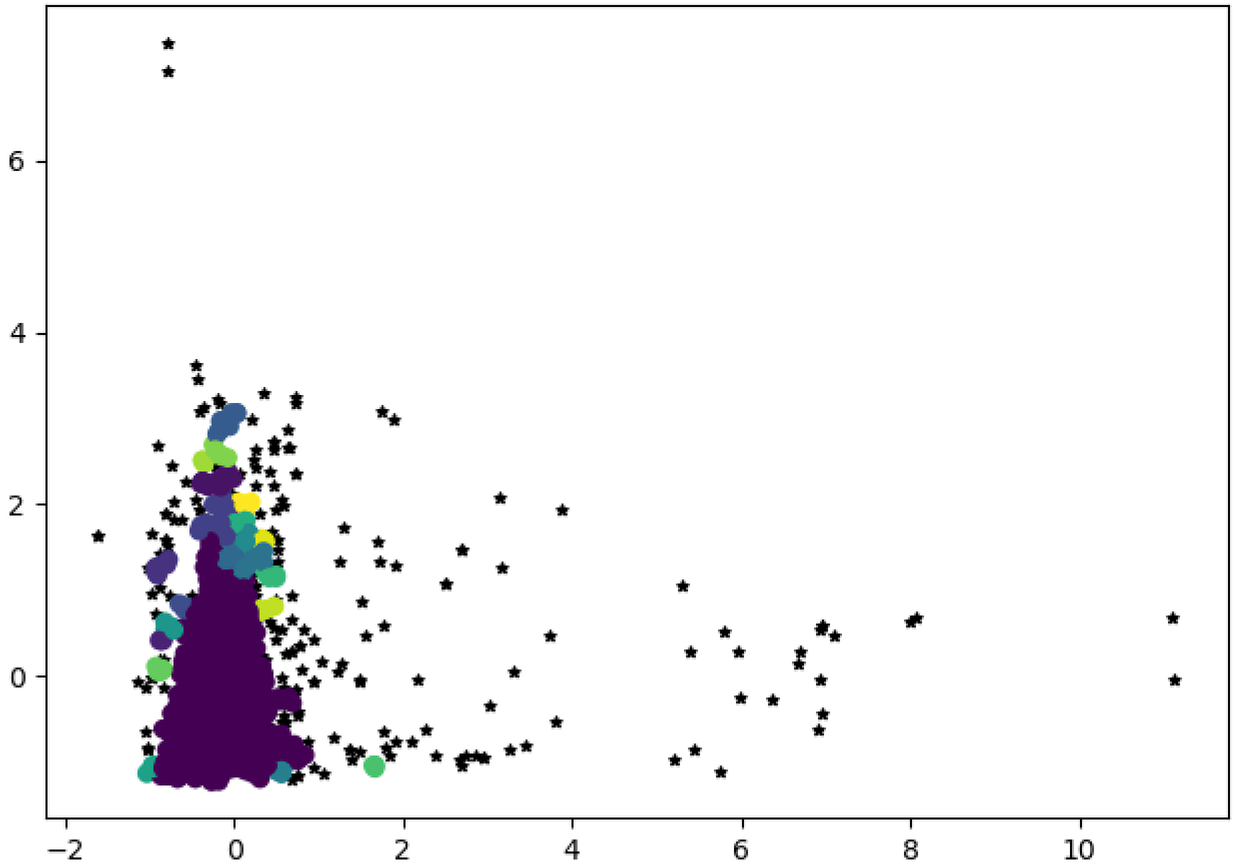


It was seen that when I increased the minpoint variable to 5, the number of clusters decreased further. At least 5 objects are now required for a cluster. Therefore, it was observed that 5 clusters were formed. If the length of the cluster formed by the neighbors of the point is less than the Minpoint value, the number of noise increases because the noise is accepted and the Minpoint value increases compared to the previous one.

4-) Epsilon = 0.1, Minpoints = 4

# of cluster = 22

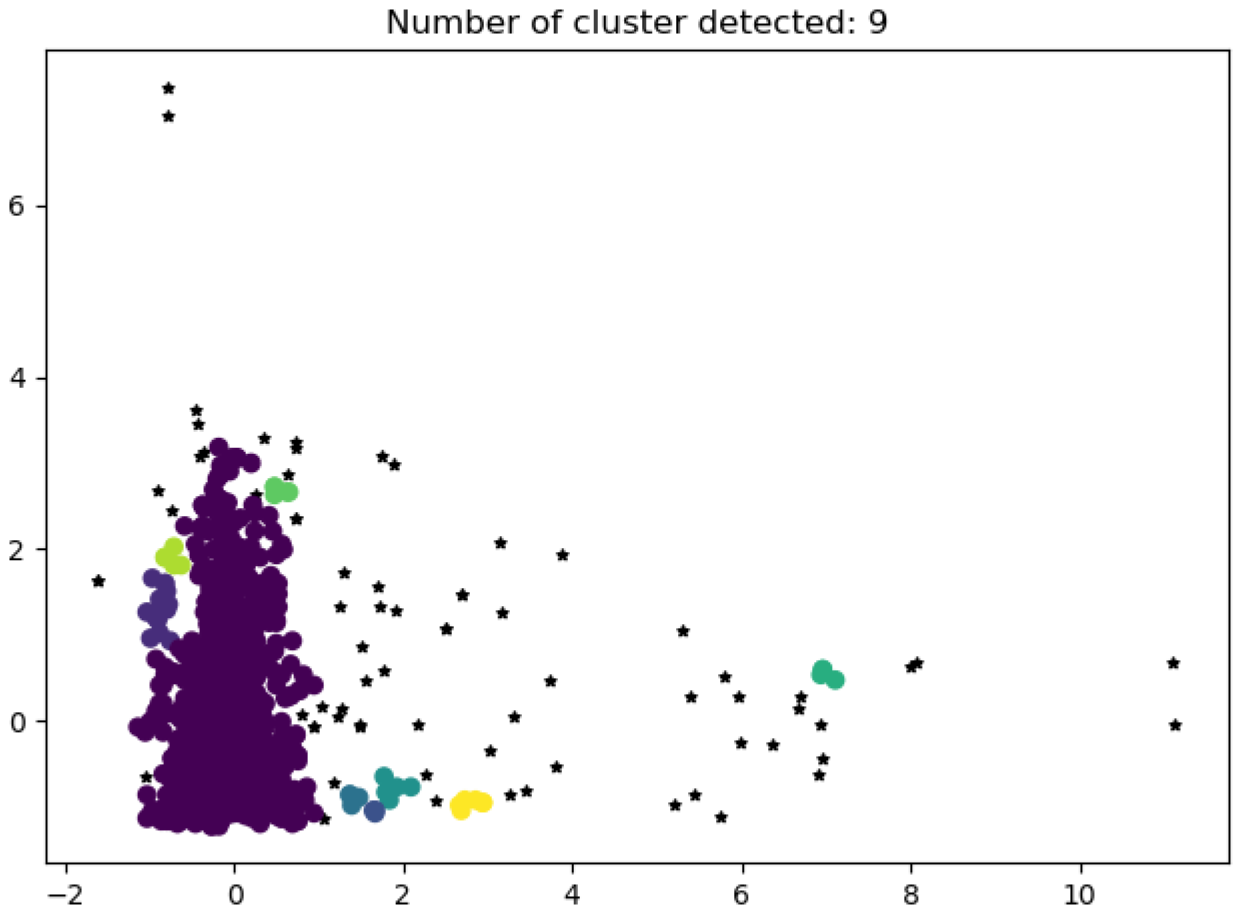
Number of cluster detected: 22



Now the minpoint variable will be kept constant at 4 and the epsilon value will be increased. First, the value 0.1 was randomly given in the plot above. As I value the Epsilon value relatively low, 22 separate clusters were formed as seen in the plot. In the dataset I use, the data are collected on the left side of the plot, so the clusters are on that side. Noise values have emerged on the right, where there is less data. And the areas where objects are relatively far from each other are these areas.

5-) Epsilon = 0.2, Minpoints = 4

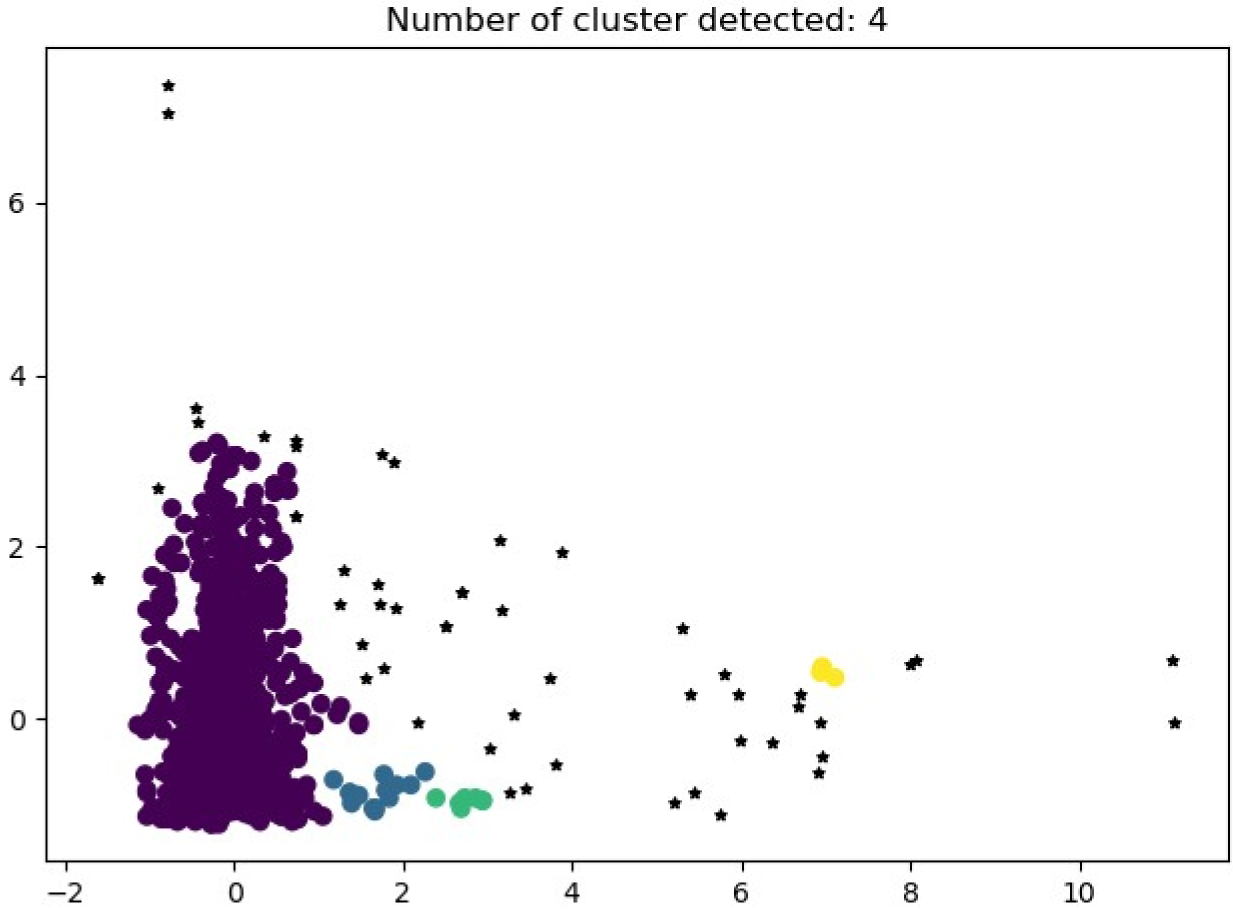
# of cluster = 9



When I increased the Epsilon value from 0.1 to 0.2, the number of clusters decreased from 22 to 9. Many clusters merged. That's why it decreased. Because now, theoretically, it is provided that objects that have more distances to each other are in the same set. Likewise, it was observed that new clusters were formed on the right side where data was relatively less (distance between objects is relatively greater). Where objects are relatively close to each other, clusters are merged, the number of clusters has decreased.

6-) Epsilon = 0.3, Minpoints = 4

# of cluster = 4



When we made the Epsilon value 0.3 this time, the number of clusters decreased further. This is a situation that changes depending on the data set. Where the objects are relatively close to each other, small clusters merge to form large clusters. The large main cluster is visible to the left of the plot. Because now the circle drawn over an object has a larger radius. This will cause objects that are relatively close to each other to fall into the same and larger cluster. Objects on the right side of the plot are quite likely to be noise. It looks the same in the graph. Of course, it is also effective if the minpoints variable is 4. However, better analysis can be made since it is already held constant in the last two analyzes.

Automatically Decide Parameters

I have read some articles for automatic determination of parameters. The articles caused some confusion as they were used more as an academic language. I would like to briefly summarize the articles on this subject.

In the first article, he proposes a simple and effective method for automatically specifying the input parameter of DBSCAN, epsilon. The study remains with the original idea of the DBSCAN algorithm and only tries to skip the user interaction needed and allows the algorithm to determine the appropriate value itself. This is done using some basic statistical techniques for outlier detection. Here, we discuss two different approaches that apply the standard deviation concept to the problem of outlier detection: the empirical rule for normal distributions and Chebyshev's inequality for non-normal distributions.

Another article proposes a new hybrid approach consisting of Binary Differential Evolution (BDE) and DBSCAN clustering algorithm as BDE DBSCAN to quickly and automatically select very suitable Eps and MinPts parameters for the DBSCAN algorithm. BDEDBSCAN performance is evaluated using various datasets with different densities and shapes. As shown in the results shown on the applied datasets, the proposed algorithm provides optimum accuracy with purity ranging from approximately 99.4% to 100%.

In the last article I could find, a new approach is proposed to calculate the eps and MinPts parameters of the DBSCAN algorithm. It is based on the kdist function that calculates the distances between the points of a data set and their second closest neighbors. As stated above, the MinPts parameter is very difficult to determine, so it is often chosen empirically depending on the data sets being investigated. In the case of the eps parameter, the main issue is to accurately determine the sharp increments.

https://core.ac.uk/download/pdf/219373759.pdf

https://hal.inria.fr/hal-01557638/document

https://www.researchgate.net/publication/342604119_A_New_Method_for_Automatic_Determining_of_the_DBSCAN_Parameters