Overview for DBSCAN

In Machine learning having grouped data is critical In improving their efficiency. Considering a given dataset holds data points with close similarity and matching patterns helps a model to understand better and make relationships within the data. This grouping is achieved through different mechanisms and these mechanisms are referred to as clustering algorithms. Clustering is an unsupervised machine learning algorithm that divides data in meaningful groups called clusters. Without a human agent posing a labeled distinction between data points, clustering provides a membership to a certain subset based on the distribution of the data.

The main advantage of clustering is mentioned in cluster hypothesis - Documents in the same cluster behave similarly with respect to relevance to information needs(Manning et.al , 2008). This means the members will respond to input variations in a close similarity. For example if the clustering for users were done based on their geographical location, the variation in products will get similar responses for people located in the same region due to the social , cultural and economic similarity.This behaviour helps machines in identifying targeted audience and data points that need adjustments.

Based on the approach they take to form clusters, these algorithms are divided into four algorithms - hierarchy based, partitioning based, density based and grid based(Bharti, n.d). DBSCAN(Density Based Spatial Clustering of Applications with Noise) is one of the density based clustering algorithms. The main feature that distincts density based clustering algorithms is their ability to form arbitrarily shaped areas to include points within a given cluster. On the contrary the other classifications focus on forming a spherically shaped area to find points with proximal distance.

DBSCAN aims to address three shortcomings posed by other types of clustering algorithms. As it depends on the dataset itself in determining its parameters -namely Epsilon (radius of the neighbourhood region) and minPoints(minimum number of points that should be contained within that neighborhood), the requirements of domain knowledge to determine the input parameters is minimal. Secondly, it addresses the issue of finding clusters with arbitrary dense regions. Most partitioning based clusters have spherical shapes. However, this approach neglects the fact that data points within a dataset lack uniform distribution. In comparison to density based clustering algorithms, the spherically clustered sets result in correspondingly poor behaviour. When it comes to DBSCAN and other density based clustering algorithms, they tend to follow distribution based shape.

It has two main parameters in the implementation of the algorithm. These are Epsilon and minPoints. The former serves as the radius of the neighbourhood around a given point within the dataset while the latter defines the minimum number of points that can form a cluster. The approach assigns different roles for the points around arbitrarily picked points to progress in including them within the initialized cluster. The core point is taken as a centroid around which other neighbour points form a cluster. It has at least neighbour points of at least minPoints in its Epsilon-neighbourhood.Another assignment of point is Direct Density Reachable(DDR) points. These are points within the range of radius ε. There is also a definition that links two arbitrary

points in the dataset. It is called Density Reachable(DR) points and it holds two points for which there exists a path through concatenation of Direct Density Reachable points.

Every clustering method is evaluated based on its strength, attaining high intra-cluster similarity and low inter-cluster similarity(Manning et.al , 2008). Its ability to find a dense region and making points around that dense region a member of the same cluster shows the DBSCAN's efficiency in ensuring data points with the same behaviour will find their way in the same cluster. The advantage of having non-spherical shapes can be seen in others as having data points that seem close,but have more similar behaviour to others to be classified within different groups, causing less inter-cluster similarity.

In conclusion, DBSCAN addresses low performance behaviours resulting from partitioning based and other clustering algorithms in its ability to select points based on their intrinsic property rather than explicitly expressed parameters. It also showcases its performance through strong intra-cluster similarity and low inter-cluster similarity. Its behaviour in creating clear and distinct dense regions that are separated with low dense regions give an intuition for human agent and also a powerful predictive behaviour for a Machine learning model. Its ability to find arbitrarily located and shaped clusters in general makes it perform better on a large dataset in comparison to other clustering algorithms.

References

Martin Ester, Hans-Peter Kreigel , Jorg Sandar, Xiaowei Xu: *A density based algorithm for discovering clusters.* A density based algorithm for discovering clusters in large spatial databases with noise. 1996.

Christopher D. Manning, Prabhakar Raghavan , Hinrich Schütze*:Introduction to Information Retrieval. 2008*

Vishal Bharti*: Clustering Algorithm*