Microclustering with Outlier Detection for DADC

Aswathy Priya M.

Computer Science & Engg, Rajiv Gandhi Institute of Technology / Kerala Technical University, India
Corresponding Author: aswathypriyam[at]gmail.com

Abstract: ***Cluster analysis is a machine learning technique for categorizing unlabeled data. The data points are grouped into different clusters based on how similar they are. The objects that may be comparable are grouped together in a group with few or no similarities. Density based clustering algorithms, which can locate clusters of any shape while avoiding outliers, are used in many applications. Density based clustering algorithms consider dense sections of objects in the data space to be clusters, separated by low density areas (noise). The Domain Adaptive Density Clustering (DADC) technique was created to point out the issues of scattered cluster loss and cluster fragmentation. Micro clustering is a stream clustering technique that preserves compact data item information. Micro clusters estimate local density by combining data from several data points in a specific area. Micro-cluster is a time-based improvement to the cluster function that effectively compresses data. Incorrect data might appear in a database for a variety of reasons. Outlier identification is a technique for filtering irregularities generated in a database. In this work, we intend to put forward a method for micro clustering technique with outlier removal for Domain Adaptive Density Clustering.***

Keywords: Density Clustering, Micro Clustering, Outlier Removal

1. Introduction

The significance of data is well acknowledged, and with the rapid growth of contemporary information science and technology, there is no doubt that the amount of data in all parts of life is continually expanding and reaching new heights; these enormous data come from a variety of sources. It's crucial to figure out how to extract meaningful and intelligible data from big datasets. In data mining, clustering technique is used to find links between data items and anticipate data distribution. Clustering is the practice of grouping together data points that have similar characteristics into a category or cluster with the least amount of overlap as possible. In a variety of data analysis domains, clustering techniques are commonly employed. There are various types of clustering techniques. Clustering is unsupervised, which implies that no prior knowledge about how to separate categories or how many to divide into is required. Clustering algorithms try to analyze data without prior knowledge by uncovering its underlying structure and classifying it into different categories based on internal homogeneity and outward bifurcation. By recognizing high-density regions in the high-dimensional data space, density-based clustering algorithms have been widely employed to generate arbitrary shape clusters. A cluster is a data space region with a high density of data points, or a collection of densely connected data points, they're density based clustering algorithms. The density based clustering technique is capable to plot clusters of any size and shape, as well as to meet certain requirements.

DADC [11] is a clustering algorithm that was developed to mention the problems of scattered cluster loss and fragmentation. This DADC consists of 3 steps: (1) Calculation of Domain Adaptive Density (2) Identification of Cluster Center (3) Cluster Self Ensemble. Using a domain adaptive density measurement approach based on K Nearest Neighbours, the density peaks of distinct density regions may be adaptively recognised (KNN). Methods for autonomously retrieving the original cluster centres and unifying fractured clusters, such as cluster centre self-identification and cluster self-ensemble, are presented on this basis. They're density-based clustering approaches. The ability of density-based clustering can discover clusters of arbitrary size and shape, as well as the fact that it meets they're density-based clustering approaches. The ability of density based clustering can locate clusters of any size and shape, as well as the fact that it fits the requirements Large data clustering and classification is a demanding task in data mining. Micro clustering is a clustering approach for data streams that preserves compact data item information. Micro-cluster is a time-based enhancement to the cluster function that compresses data effectively. Micro-clustering refers to clustering models that produce small clusters or cluster sizes that grow in a nonlinear connection with the amount of samples. For density-based clustering, micro-clustering was used. Keeping track of the clusters in a logical and efficient manner. Micro clusters are a popular stream clustering approach that preserves the clustering's compact form. Outlier detection is a method of detecting and filtering abnormalities in a database. Outliers are values that differ significantly from the norm. Other data points could be used to reflect measurement variability, experimental flaws, or uniqueness. The database contains anomalies and outliers. They're often the result of measurement errors or unusual system conditions, and they don't show how the underlying system functions in a larger context. The most common basis of outliers are data entry errors, experimental errors, data processing errors, sampling errors, intentional ones, and natural causes.

The following is a list of the paper's contributions.

1) The value of K is calculated as a percentage of the dataset, to consider the variation in the size of each datasets.
2) Micro-clusters are created from the Initial clusters, which are eligible to merge with other based on the similarity calculations of domain densities.
3) Micro-clusters thus formed will be checked against the similarity and threshold matching. And those which satisfy the condition will be merged together.
4) Micro-clusters which left alone from merging with other clusters are due to not meeting the similarity check.

5) The outliers/noise present in these micro-clusters limits them from satisfying the merging criteria. So the outliers from these microclusters needs to be detected and/or removed effectively.
6) After outlier removal, merge the micro-clusters which satisfies the conditions.

2. Literature Survey

Alex Rodriguez and Alessandro Laio concentrated on cluster analysis, which divides components into categories based on their similarity [1]. Among the applications are astronomy, bioinformatics, bibliometrics, and pattern recognition, to name a few. We suggest an approach based on the assumption that cluster centres have a higher density than their neighbours and are isolated from places with higher densities by a significant distance. This concept underpins a clustering method in which the number of clusters is intuitively estimated, outliers are automatically identified and removed from the analysis, and clusters are recognised regardless of the shape or dimensionality of the space in which they are embedded. We use a variety of test examples to demonstrate the algorithm's utility.

Clustering is a centre for research in data mining, with a wide range of practical applications. Rodriguez and Laio provide a clustering algorithm in Science that automatically discovers clustering centres and clusters things efficiently and effectively. On the other side, the method is reliant on a single parameter and has difficulties calculating the "optimal" number of clusters. To address these difficulties, Guangtao Wang and Qinbao Song [2] propose new clustering methods that use statistical testing to automatically select clustering centres. The suggested technique supplies a metric to evaluate the centrality of each item after first generating a new metric to examine the density of an object that is more resistive to the pre-assigned parameter. The system then uses an external statistical testing mechanism to find clustering centres among items with exceptionally high centrality metrics. Finally, it divides the remaining items into clusters based on the density of their nearest neighbours. Extensive experiments on various types of clustering data sets are conducted in order to evaluate the suggested method's performance and compare it to one published in Science. The results indicate that the recommended technique is both effective and reliable.

Cluster analysis tries to uncover hidden relationships between objects in a dataset, allowing like objects to be grouped together and different objects to be separated. A cluster is a connected graph [3] in which the similarity between any two adjacent neighbours is greater than or equal to a threshold; a cluster is a connected graph in which the similarity between any two adjacent neighbours is greater than or equal to a threshold; a cluster is a connected graph in which the similarity between any two adjacent neighbours is greater than or equal to a threshold. Similarity is determined using object local density, which is defined as the sum of the distances between an item and its k-nearest neighbours; a cluster is a connected graph in which the similarity between any two adjacent neighbours is greater than or equal to a threshold.

Density peaks clustering are a density-based clustering technique that excels in accuracy, detecting the number of clusters automatically, and locating the centre points. Density peaks clustering, on the other hand, have a number of issues that must be addressed before it can be extensively adopted. A sensitive predened parameter, for example, is ineffective for large datasets, and a decision graph frequently gives wrong centre points. To address these issues, Lina Liu and Donghua Yu [4] present a novel Density peaks clustering technique based on weighted k-nearest neighbours and geodesic distance to improve clustering performance for both manifold and non-manifold datasets. Weighted k-nearest neighbours based on Euclidean distance are incorporated into the Density peaks clustering technique based on weighted k-nearest neighbours and geodesic distance to optimise local density. According to experimental results on artificial and real-world datasets, including picture datasets, the Density peaks clustering technique based on weighted k-nearest neighbours and geodesic distance outperforms state-of-the-art comparison techniques on both manifold and non-manifold datasets. Furthermore, the DPCGD problem, which results in erroneous decision graph centre points, is solved using the Density peaks clustering technique, which is based on weighted k-nearest neighbours and geodesic distance.

Ensemble clustering is becoming increasingly popular for solving traditional clustering problems. Ensemble clustering is used to find a single cluster which will better fit in some way than the existing clustering when a number of various (input) clustering have been established for a given dataset. In recent years, a number of ways to solving ensemble clustering problems have been developed. The majority of these ensemble techniques, on the other hand, are intended for partitional clustering algorithms. Only a few studies have been done on ensemble hierarchical clustering algorithms. The work [5] by Li Zheng, Tao Li, and Chris Ding introduces a hierarchical ensemble clustering framework that allows partitional and hierarchical clustering findings to be organically merged. We acknowledge the importance of ultra metric distance in hierarchical clustering and present a new method for learning ultra metric distance from a collection of distance matrices and generating final hierarchical clustering with improved cluster separation. Experimental evidence backs up the utility of our suggested approaches.

3. Problem Definition

It's possible that the DADC cluster's data points were assigned to the wrong clusters. Because clusters are combined via IDS, CDS, CCD, and CFD. However, the data points in those resulting clusters could differ. The numerous data points within each cluster that is appropriate for merging must therefore be precisely identified. The outliers must be identified and removed in order to stop those clusters from merging. These clusters must meet certain requirements in order to join forces with other clusters.

4. Methodology

We recommend a micro-clustering with outlier detection/ elimination strategy for the Domain-Adaptive Density

Clustering method [11]. Microclustering is used to maintain compact data item information in data streams. Micro-clusters use information from several data points in a particular area to calculate local density. To determine probable cluster centres, domain densities and delta distances are used as decision factors. Micro-clusters are formed from the initial clusters, and in order to stop them from joining other clusters based on similarity matching, Outliers are discovered and eliminated from the micro-clusters that have been left alone. Fig 1 shows the framework of Microclustering with Outlier Detection for DADC.

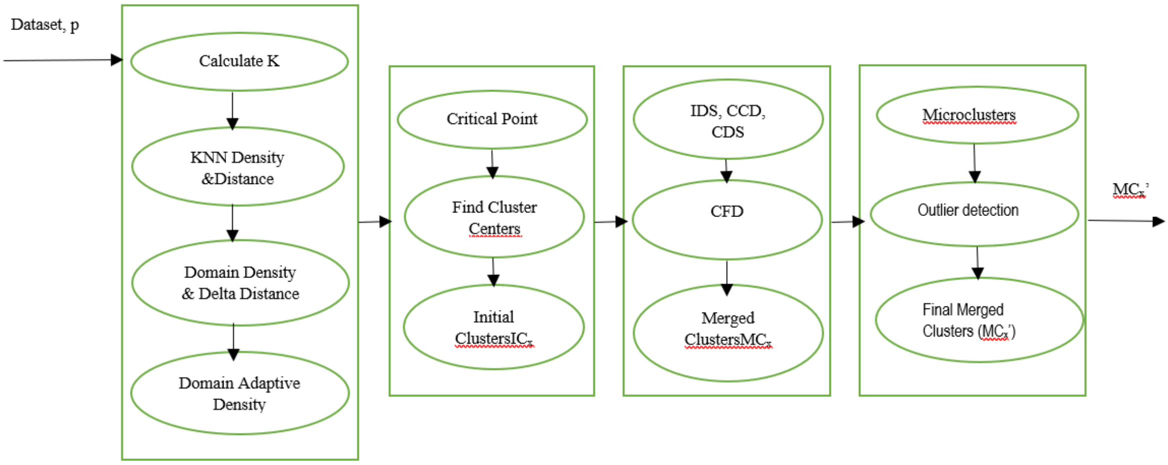


Figure 1: Microclustering with Outlier Detection for DADC

Algorithm 1: Domain Adaptive Density Measurement

```
Input:
X: The dataset for Microclustering
p: Percentage to calculate the K value

Output:
The domain adaptive densities and delta distance of the data points of X

1)  Calculate the Euclidean Distance for X
2)  For each datapoint xi in X do,
    2.1 calculate K-Nearest Neighbours of xi
    2.2 Calculate KNN distance
    2.3 Calculate KNN density
    2.4 Calculate Domain Density
    2.5 Find the datapoint with maximum domain density
3)  For each datapoint xi in X check
    3.1 whether (∂i< ∂max)
    3.2 Then for each neighbourxj of xi, do
    3.2.1 if (∂j > ∂i and δj<δi)
    3.2.2 set Domain Adaptive density of xi as ∂i*max(δij)
    3.3 set Domain Adaptive density of xi as ∂i*(δij)
4)  Domain Adaptive density and delta distance of all xi
```

This section describes a domain-adaptive density calculation approach for adaptively recognising the domain-density peaks in the different density regions of the dataset. Domain distance and domain density computing methods are provided [6] [7] based on the KNN approach. When applied to large datasets, which in real applications usually include a range of distribution densities, these techniques are especially beneficial and useful.

The average distance between each data point xi and its k nearest neighbours is referred to as the KNN-distance [11] for a dataset X. The KNN-Distance is calculated as below;

$$KDisti = \frac{1}{K}\ dij$$

The reciprocal of the KNN-distance [11] is referred to as the KNN-density of the data point xi in dataset X. A data point's lower KNN density suggests indicating a more sparse area is where this data point is situated. The KNN-Density calculated as below;

$$KDeni = \frac{1}{KDisti}$$

Each data point's domain density in dataset X is calculated as the sum of that data point's KNN-density and the weighted KNN-density of its K-nearest neighbours. The Domain Density ∂i is calculated as;

$$\partial i = KDeni + \sum_{j\epsilon N\ (xi)} (KDenj \times wj)$$

Where wj =1/dij weighted KNN-density between each neighbour, xj and xi. In comparison to the KNN-density, the domain density can more precisely depict the density distribution of nearby data points.

Each data point's (xi) Delta distance δi is calculated as a clustering decision parameter based on the domain density. By figuring out the shortest path between xi and any other places with higher densities, the delta distance of xi is calculated. In this scenario, the maximum value of the Delta distance is only present at the places with the highest global density. Compared to the remaining points in a moderately dense region, the domain density peak in a scattered zone produces a smaller Delta distance value. Given that a dataset contains various regions with various densities, data points in a dense region have greater domain densities than those in a scattered zone. We improve the definition of domain-adaptive density [11] by integrating the values of the domain density and Delta distance in order to adaptively identify the density peaks of each region.. The domain-adaptive density ∂i of each data point in dataset is calculated as;

$$\partial i = \partial i \times \delta i = \left\{ \begin{pmatrix} \partial i \ \times \max(dij) \\ \partial i \ \times \min(dij) \end{pmatrix} \right.$$

We multiply the domain density ∂ and Delta distance δ for each location to easily identify the density peaks of a region.

A clustering decision graph [11] is created to show the candidate cluster centres based on the domain density ∂ and delta distance δ values. The horizontal axis in the clustering decision graph stands for ∂ and the vertical axis for δ. Points having high values of ∂ and δ are considered as cluster centers, while points with a low ∂ and a high δ are considered as outliers.

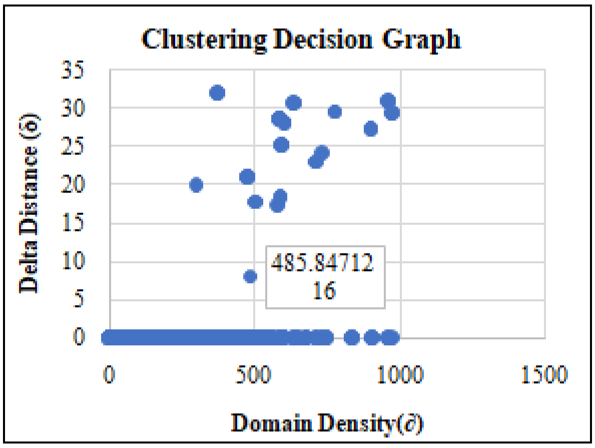
Algorithm 2: Cluster center identification

```
INPUT:
X: The dataset
∂: The domain densities of all xi in X
∂: The delta distances of all xi in X

OUTPUT:
Initial Clusters of X
1) Find Cp(x)=∂max/2
2) Find Cp(y)=δmax/4
3) For each xi in X do
    3.1 if ∂i> Cp(x) and δi> Cp(y)
    Append the datapoint xi to set of cluster centers
    1.2  else append the xi to the set of remaining datapoints
4) For each xi in the set remaining datapoints
    4.1 if δij< max value and ∂i<  ∂j
    4.2 set neighbour of xi as xj
    4.3 If CC(xj)!= set of cluster centers
    4.4 Find the nearest neighbour of xj (repeat step 4.1 to
    4.3)
5) Return Initial Clusters formed
```

We present a cluster centre identification approach to extract initial cluster centres by automatically identifying the parameter thresholds of the clustering decision graph in order to address the issue of cluster fragmentation. Candidate cluster centres for a dataset can be found from the related clustering decision graph using the domain-adaptive densities and Delta distances. Each of the remaining data points is then assigned to the cluster with the closest and highest density neighbours once the cluster centres have been determined.

The CFSFDP algorithm's [1] clustering outcomes rely on a rigid restriction that only one local density maximum is expected to exist in each candidate cluster. The CFSFDP algorithm has the drawback of considering data points with high local density and delta distance values as cluster centres. However, in actual use, these parameter levels are frequently set by hand. Based on the clustering decision graph, we suggest a self-identification technique to automatically extract the cluster centres. A crucial point of the clustering decision graph is established to automatically compute the parameter threshold values for domain density and Delta distance. A splitting point by which the candidate cluster centres and remaining points can be divided is known as the crucial point Cp(x, y) of a clustering decision graph.



Cluster centers are the points with high domain density. The Critical point is calculated as Cp(x)=∂max/2 & Cp(y)=∂max/4. Therefore, the value of critical point Cp(x, y) of the clustering decision graph is defined as:

$$Cp(x, y) = \left(\frac{\partial max}{2}, \frac{\delta max}{4}\right)$$

Where ∂max and δmax are the maximum values of ∂ &δ .

Each of the remaining data points is assigned to the cluster that contains the closest and highest domain-density neighbours after cluster centres are found. Repeat this process until all data points have been assigned to the appropriate clusters.

Algorithm 3: Cluster Self Ensemble

INPUT:
ICx: Initial Clusters of X
th: Threshold value of cluster fusion degree for Cluster merging

OUTPUT:

MCx: Merged Clusters of X

```
1)  Calculate Inter Cluster Density Similarity $S_{a, b}$
2)  Calculate Cluster Crossing points
3)  Calculate Cluster Crossover Degree Ca, b
4)  Calculate Cluster Density Similarity Da, b
5)  Calculate Cluster Fusion Degree Fa, b
6)  If Fa, b>th, then
        1.1  Merge Clusters Ca, Cb
7)  Return MCx
```

As we all know, the essential concept of clustering analysis is that each datapoints who are members of the same cluster are distinct from those who are members of various clusters while yet sharing a high degree of similarity. In order to determine whether clusters are improperly fragmented into several subclusters, we offer an inter-cluster similarity measurement and a cluster fusion degree model for merging the fractured cluster. To create a single cluster, cluster fusion and clusters with a higher density of similarity are joined.

Inter-cluster Density Similarity (IDS): The degree of similarity between two clusters' intercluster densities is known as intercluster density similarity. The average value of the domain densities of all the data points in a cluster is the cluster's average density.

Let $S_{a,\ b}$ [11]represent the similarity in inter-cluster density between clusters ca and cb. The more comparable the density of the two clusters is, the bigger the value of $S_{a,\ b}$. The $S_{a,\ b}$ is calculated as;

$$S_{a,b} = \frac{2\sqrt{\overline{KDen_{c_a}} \times \overline{KDen_{c_b}}}}{\overline{KDen_{c_a}} + \overline{KDen_{c_b}}}$$

The distance between each pair of clusters is taken into account. The distance between two clusters was calculated using a variety of ways. To calculate the separation between clusters, we provide a novel approach. Every two clusters have a crossing point, and the crossover degree of the clusters is determined by finding these crossing points.

The crossover degree c(i, a→b) [11]of a crossing point xi in ca between clusters ca and cb is defined as:

$$C_{i,a\rightarrow b} = \frac{2\sqrt{|N_{i,a}| \times |N_{(i,b)}|}}{|N_{(i,a)}| + |N_{(i,b)}|}$$

Cluster crossover degree Ca, b [11]of two clusters ca and cb is calculated by the sum of the crossover degrees of all crossing points between ca and cb. The formula of CCD is defined as:

$$C_{a,b} = \sum_{x_i \in CP_{a\rightarrow b}} c_{(i,a\rightarrow b)} + \sum_{x_j \in CP_{b\rightarrow a}} c_{(j,b\rightarrow a)}$$

Cluster density stability [11] is the inverse of cluster density variance, which is determined by the difference between the average domain density of the cluster and the domain density of each individual point. If the domain density differences between each point in a cluster are smaller, then the larger the CDS of the cluster is. The CDS of a cluster ca is defined as:

$$d_a = \log(\sqrt{\sum_{i \in c_a} (KDen_i - \overline{KDen_{c_a}})^2})$$

According to the values of IDS, CCD, and CDS, the Cluster Fusion Degree (CFD) [11]of two clusters is the degree of correlation between the clusters in terms of position and density distribution. The following requirements should be met by two clusters to achieve a high degree of fusion: (1) high IDS, (2) high CCD, and (3) the fused cluster's CDS should be close to the average value of the two originating clusters' CDSs. A high fusion degree is present between two crossing and nearby clusters that have a high IDS.

$$F_{a,b} = S(S_{a,b}, C_{a,b}, D_{a,b})$$

$$= \frac{\sqrt{3}}{4}(S_{a,b} \times C_{a,b} + C_{a,b} \times D_{a,b} + D_{a,b} \times S_{a,b})$$

If the value of $F_{a,\ b}$go beyond a given threshold, then clusters ca and cb are combined to a single cluster.

Algorithm 4: Microclustering and Outlier detection

```
INPUT:
MCx: Merged Clusters of X
mc:No:of data points in each micro cluster

OUTPUT:
Mix: Micro Clusters of MCx
MCx’: Merged clusters after the outlier has been appended
to appropriate cluster or after removal of the outlier

1) Get the datapoints with same cluster center (cci).
2) If No:of datapoints with same cci > mc
    1.1  Do Microclustering
3) Calculate the total Domain Density of each merged
   Clusters (MCx)
4) Calculate the total Domain Density of each microclusters
5) While (Size of microcluster<mc )
    1.2  Consider these microclusters to be the outliers
    1.3  Calculate the total domain density(tdd) while
         adding these outliers to each of the Merged Clusters
         (MCx)
    1.4  Append the outlier to the MCxi, where the tdd is
         more when compared to other MCx
    1.5  If the tdd< the domain density of the MCx
         1.5.1  Remove the microcluster from the list
6) Return the Final Merged Cluster MCx’
```

Microclustering, a stream clustering approach, records only the most basic information about the data items in data streams. The microcluster feature, a temporal extension of the cluster feature, effectively compresses data. The term "microclustering" is used to describe clustering techniques that result in small clusters or techniques that cause the size of the clusters to rise nonlinearly as the number of samples rises. Outlier detection is the technique of removing anomalies produced in a database. Outlier values stand out from the norm and are extraordinary. Other findings point to some measurement volatility, which may be the result of experimental, measurement, or novelty errors. Anomalies and outliers can be found in the database. A training machine may have serious problems with this. You can apply statistical methods or discover new algorithms. They don't depict how the underlying system functions generally since they are typically the result of measuring errors or unexpected system conditions. Data input errors, experimental errors, data processing errors, sampling errors, purposeful ones, and natural causes are the most frequent sources of outliers in a set of data. It is true that it is best practise to eliminate outliers before continuing with the study. It might be challenging to find outliers in data. In the modern world, it is essential to identify aberrant behaviour in order to unearth crucial facts, observations, and accurate data projections. Detecting outliers is an important data mining task that seeks to find objects that differ from the typical data’s predicted pattern.

5. Result and Discussion

The experiments make use of a collection of synthetic datasets. As stated in Table 1, these benchmark datasets were downloaded from publicly available online benchmarks such the clustering benchmark datasets [8] and UCI Machine Learning Repository [9].

Table 1: Clustering Accuracy of each dataset

S. No	Dataset	No. of Data points	Dimensions
1	Aggregation	788	2
2	Compound	399	2
3	Heart Shapes	213	2

Clustering Results on Heart Shaped Dataset

Heart shaped dataset is a synthetic dataset, which is composed of 215 datapoints. Domain adaptive density and delta distance for each of the datapoints are calculated. From the experiment, it is obvious that there is only one cluster center formed. From the clustering decision graph it is clear that, there were only one decion point with highest values of both domain density and delta distance. Also while doing the microclustering, we got many microclusters with size threshold value. Out of these microcluster, we need to find out the outlier which needs to removed from the clusters. 100% is the clustering accuracy formed after microclustering with outlier detection.

Clustering Results on Aggregation Dataset

Aggregation dataset is a synthetic dataset, which is composed of 787 datapoints. For each points in the dataset, domain adaptive density and delta distance are calculated. From the experiment, there were 33 cluster formed initially. After the implementation of DADC algorithm, we got 11 clusters. After performing the microclustering with outlier detection, the no of clusters has been reduced to 10. Also the outlier points of each initial cluster have been appended with the cluster to which it needs to belong. After doing the microclustering with outlier detection, received a clustering accuracy of 98%.

Clustering Results on Compound Dataset

Compound dataset is a synthetic dataset, which is composed of 398 datapoints. For each points in the dataset, domain adaptive density and delta distance is calculated. From the experiment, there were 24 cluster centers formed initially. After the implementation of DADC algorithm, we got 9 clusters. After performing the microclustering with outlier detection, the number of clusters has been reduced to 8. Also the outlier points of each initial cluster have been appended with the cluster to which it needs to belong. After doing the microclustering with outlier detection, received a clustering accuracy of 96%.

6. Performance Evaluation

Clustering Accuracy (CA) [11]is used to evaluate the performance of the clustering algorithms. It measures the ratio of the perfectly clustered datapoints to the predefined class labels. CA is calculated as,

$$CA = \sum_{i=0}^{K-1} \frac{\max\left(\frac{Ci}{Li}\right)}{|X|}$$

Where Ci refers to the ith datapoint in the ith cluster, K is number of Class labels & L be the set of predefined class labels.

Table 2: Dataset Information.

S. No	Dataset	No. of Data points	Clusters	Accuracy
1	Aggregation	787	10	98%
2	Compound	398	8	96%
3	Heart Shapes	215	1	100%

7. Conclusion

Our goal was to devise a mechanism for swiftly and precisely creating clusters. Outlier detection or the process of extracting anomalous datapoints is a time consuming task. To address the problem of finding outliers in DADC, we presented a clustering based method called Microclustering with Outlier Detection for DADC in this study. The proposed approach employs the microcluster technology to cluster nearby related data elements. In the case of memory use, not every object inside the microclusters was kept in memory, and any outlier data points were removed from memory to save on memory use. The tests are carried out using fictitious datasets, and our method will reveal to be more effective in terms of clustering accuracy.

References

[1] WA. Rodriguez and A. Laio, “Clustering by fast search and find of density peaks, ” Science, vol. 344, no. 6191, pp. 1492–1496, 2014.

[2] G. Wang and Q. Song, “Automatic clustering via outward statistical testing on density metrics,” IEEE Trans. Knowl. Data Eng., vol. 28, no. 8, pp. 1971–1985, 2016.

[3] Fahim, ”A Clustering Algorithm for Varied Density Clusters based on Similarity of Local Density of Objects, ” 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020, pp. 26-31.

[4] LINA LIU and DONGHUA YU, ”Density Peaks Clustering Algorithm Based on Weighted k-Nearest Neighbors and Geodesic Distance ”, September 24, 2020.

[5] L. Zheng, T. Li, and C. Ding, “A framework for hierarchical ensemble clustering, ” ACM Trans. Knowl. Discov. Data, vol. 9, no. 2, pp. 9:1–23, 2014.

[6] S. Yang, M. A. Cheema, X. Lin, Y. Zhang, andW. Zhang, “Reverse k nearest neighbors queries and spatial reverse top-k queries, ” The VLDB Journal, vol. 26, no. 2, pp. 151–176, 2017.

[7] D. Jiang, G. Chen, B. C. Ooi, K.-L. Tan, and S. Wu, “Epic: An extensible and scalable system for processing big data, ” in VLDB’14, 2014, pp. 541–552.

[8] P. F. et al, “Clustering datasets, ” 2017. [Online]. Available: http://cs.uef.fi/sipu/datasets/

[9] U. of California, “Uci machine learning repository, ” Website, 2017, http://archive.ics.uci.edu/ml/datasets.

[10] Mohamed JawardBah, Hongzhi Wang, Li-Hui Zhao, Ji Zhang, and Jie Xiao, ”EMMCLODS:An Effective Microcluster and Minimal Pruning CLustering-Based Techniquefor Detecting Outliers in Data Streams”, September 2021.

[11] JianguoChen, and Philip S. Yu, "A Domain Adaptive Density Clustering Algorithm for Data with Varying Density Distribution", vol 33, IEEE Transactions on Knowledge and Data Engineering, 2021