DBHD: Density-based clustering for highly varying density

https://ieeexplore.ieee.org/document/10027741

Problem statement

Traditional clustering algorithms have strong underlying assumptions about the homogeneity of a cluster's shape, density, or size. As a result, they may not perform well when a dataset violates one or more of these assumptions. This poses a challenge for clustering datasets with varying sizes, densities, and shapes. The goal is to find an algorithm that can effectively cluster such datasets and discover hidden patterns that other algorithms may miss.

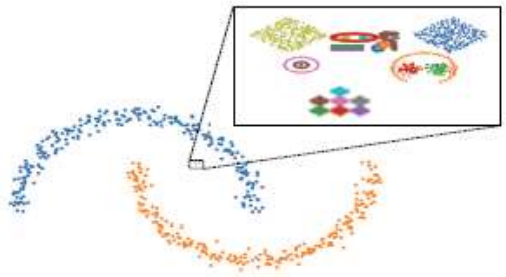

Fig. 1: The diagram shows a toy dataset in which clusters of various shapes with extremely different densities and sizes exist. In this scenario, DBHD achieves an AMI score of 0.990.

Motivation

1. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (KDD-96) (pp. 226-231).

2. Chaudhuri, K., & Dasgupta, S. (2010). Rates of convergence for the cluster tree. *The Annals of Statistics*, 38(5), 3167-3194.

I cite two relevant papers that justify the need for DBHD algorithm. This first paper cited introduced that DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm that works well for datasets with clusters of similar density, but it can have limitations when it comes to clustering datasets with varying sizes and densities. Because if the dataset has small clusters surrounded by larger ones, DBSCAN may not be able to accurately identify the smaller clusters as separate clusters, as they may be considered as noise or part of the larger cluster. The second paper cited discusses the limitations of traditional clustering algorithms like the hierarchical clustering algorithm called Cluster Tree, which is effective but computationally expensive.

Key ideas and techniques

1. Local density estimation

DBHD estimates the local density of each data point by considering the distances between the data point and its k nearest neighbors. This approach takes into account the local density information, which is important in identifying clusters of varying densities.

2. Two new conditions for data point assignment

DBHD introduces two new conditions for data point assignment that make it possible to identify clusters of varying densities, sizes, shapes, and scaling.

3. Adaptively computed density information

DBHD uses adaptively computed density information to detect clusters with varying sizes and densities.

4. Robustness to noise

DBHD is robust to noise because it considers the local density information of each point and requires that a data point should have a local density consistent with the local densities of the other points in the cluster to be assigned to the cluster. This robustness to noise allows DBHD to effectively cluster datasets with high levels of noise.

Contributions

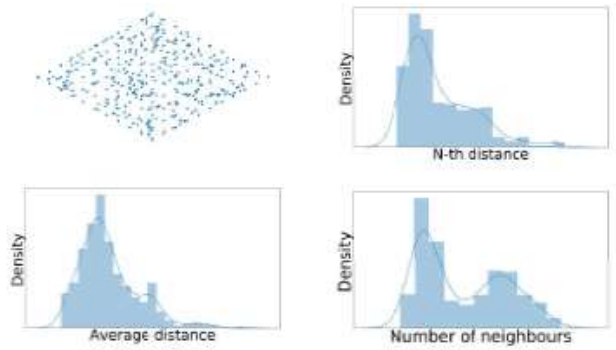

Fig. 2: The figures illustrate how key properties of a data point can vary within a cluster.

Effort of DBHD:

The algorithm has three input parameters, such as n, which defines the minimum size of a cluster; ρ, which sets the threshold for deviation within a cluster; and β, which defines the maximum deviation of average distances within a cluster. The algorithm works by iteratively extracting clusters from the dataset, starting with the whole dataset and dividing it into smaller clusters until each cluster contains no more than n data points.

Robust clustering results

This new approach provides robust and stable clustering results for a wide range of parameters compared to classical density-based approaches like DBSCAN or OPTICS.

Improved performance over state-of-the-art algorithms

The experimental results demonstrate that DBHD outperforms state-of-the-art-clustering algorithms while simultaneously being more robust against outliers.

Experimental evaluation

1. Technical perspective

The experimental study in the paper evaluates the performance of DBHD algorithm against several state-of-the-art clustering algorithms, including DBSCAN, OPTICS, HDBSCAN, k-means, Spectral ACL, Spectral DPC, Mean-

shift, SNN-DPC, LSDBC, and DBADV. The major metrics used to evaluate the proposed approach include accuracy and robustness to noise.

The impact factors investigated in the study include the size of input data and the complexity of queries. The authors conducted experiments on both synthetic and real-world datasets with varying sizes and densities to evaluate the performance of DBHD algorithm. They also tested the algorithm's robustness to noise by adding different levels of noise to the datasets.

The experimental results show that DBHD outperforms all other clustering algorithms in terms of accuracy and robustness to noise. The authors also demonstrate that DBHD is more computationally efficient than some of the other algorithms tested.

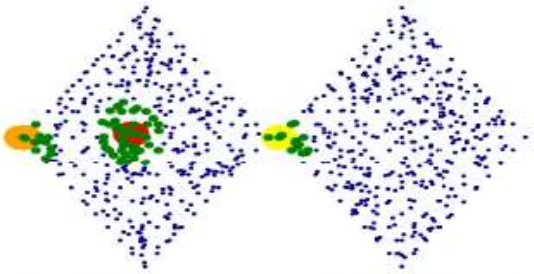

Fig. 3: The diagram shows some exemplary data points, a red central point, an orange border point and a yellow connecting point together with their neighborhoods (highlighted in green). We show a red central point, an orange border point, and a yellow connecting point. The neighborhood of each point is thereby highlighted in green.

One representative figure from the paper is Figure 2, which shows a comparison of clustering results for different algorithms on a synthetic dataset with varying densities. Another representative figure is Figure 3, which shows a comparison of clustering results for different algorithms on a real-world dataset.

2. Criteria for the line of research

Expressive Power:

Two new conditions are introduced to distinguish between different types of data points, which enables more effective filtering of clusters compared to state-of-the-art algorithms. Therefore, the algorithm has a high expressive power.

Complexity:

The proposed DBHD algorithm has a complexity of O(n^2), which is higher compared to some other clustering algorithms. However, the algorithm is optimized for performance, and the authors provide a detailed analysis of the computational complexity of the algorithm, which can be useful for understanding its performance.

Accuracy:

There are extensive experimental results, which confirm that the DBHD algorithm outperforms other state-of-the-art algorithms in terms of clustering accuracy for both synthetic and real-world datasets. The experiment also uses two commonly used metrics, the adjusted mutual information (AMI) and the adjusted rand index (ARI), to evaluate the clustering outcomes of all methods quantitatively.

Algorithm 1: DBHD

input : dataset: $\mathcal{D}$, minimum cluster size: n, cluster member ratio: ρ, density deviation threshold: β
output: clusterSet, outlierSet
// Initialization:
$\mathcal{D}^* := \mathcal{D}$ // The set with all unprocessed points
clusterSet := {}
// Main Loop:
while $|\mathcal{D}^*| > n$ **do**
  $\epsilon := \min_{x \in \mathcal{D}^*} \max_{y \in Q(x)} \text{dist}(x, y)$
  $\sigma := \min_{x \in \mathcal{D}^*} \text{avg}_Q(x)$
  $C := \{\sigma\}$ // The object set for the new cluster
  $P := \mathcal{H}(\mathcal{D}^*, \sigma, \epsilon)$**repeat**
    $M := C \cup P$ **for** $x \in P$ **do**
      $N = \{\}$ // The set of new cluster points
      if $\beta \times \text{avg}_Q(x) \leq \text{avg}_Q(\sigma)$ $\wedge\ \rho \times |\mathcal{H}(\mathcal{D}^*, x, \epsilon) \setminus M| \leq |\mathcal{H}(\mathcal{D}^*, x, \epsilon) \cap M|$ **then**
        $N := N \cup \{x\}$
      end
    end
    $C := C \cup N$
    $P := \left(\bigcup_{x \in N} \mathcal{H}(x, \epsilon)\right) \setminus C$
  until C *does not change anymore*;
  clusterSet := clusterSet $\cup \{C\}$
  $\mathcal{D}^* := \mathcal{D}^* \setminus C$
end
outlierSet := $\mathcal{D}^*$
return clusterSet, outlierSet

Scalability:

This algorithm is capable of handling large datasets with large numbers of clusters, which makes it scalable. However, the complexity of the algorithm is O(n^2), which can be a limitation for very large datasets.

3. Conclusion

The Density-Based Hierarchical Density (DBHD) clustering algorithm was proposed to handle large datasets with varying cluster properties. It can adaptively determine the density of clusters and is robust against outliers. Two new conditions were introduced to distinguish between different types of data points, making cluster filtering more effective. Extensive experiments were conducted with synthetic and real datasets, and the results showed that DBHD outperformed other state-of-the-art algorithms. Therefore, DBHD is an attractive candidate for future data mining tasks.

4. Evaluation based on my own criteria

Strong points:

- DBHD is a flexible algorithm that can be adapted to different applications and data types. Additionally, it can be extended to incorporate domain-specific knowledge or constraints, making it a powerful tool for

data analysis and exploration.

- The experimental study demonstrates that DBHD outperforms state-of-the-art clustering algorithms in terms of accuracy and robustness to noise.
- While DBHD has a higher computational complexity than some other clustering algorithms, it is still scalable to handle large datasets. This is because it uses an adaptive approach to density estimation and clustering, which allows it to handle datasets with varying sizes and densities more efficiently.

Weak points:

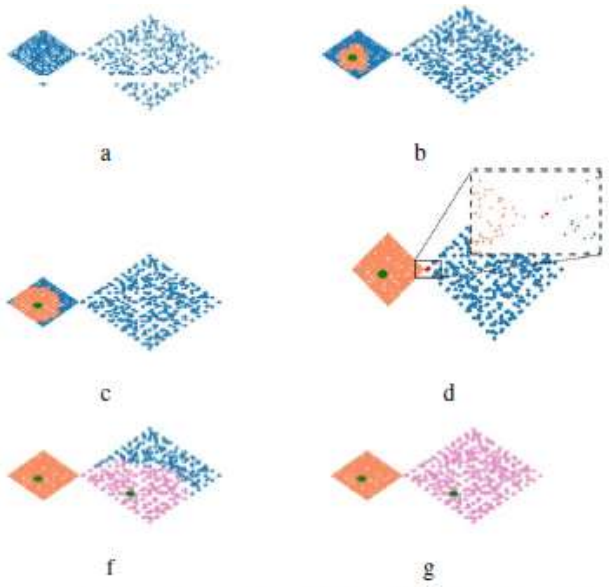

Fig. 4: We illustrate the steps of DBHD on a variation of the twoDiamond dataset [16].

- Some readers may find it difficult to understand the local density estimation approach used in the algorithm.
- The authors do not provide a thorough analysis of the computational complexity of DBHD algorithm. It would be helpful to know how the algorithm scales with increasing dataset sizes.
- While the experimental study is comprehensive, it would be interesting to see how DBHD performs on datasets with different characteristics, such as high-dimensional data or datasets with imbalanced class distributions.

Extended Discussion

If I were the authors, here is what I would plan to do to address the weaknesses or improve/extend any part of the solution:

- To address the weakness of not providing a thorough analysis of the computational complexity of DBHD algorithm, I would conduct experiments on datasets with varying sizes and dimensions to evaluate how the algorithm scales. This would help to provide a better understanding of the algorithm's performance on large-scale datasets.
- To address the weakness of not providing detailed explanations of the key ideas and techniques used in DBHD algorithm, I would include more examples and illustrations in the paper to help readers understand how local density estimation works and how it is used in DBHD.
- To improve/extend the solution, I would explore ways to make DBHD more efficient without sacrificing accuracy. This could involve developing new algorithms for local density estimation or exploring different clustering techniques that can handle datasets with varying sizes and densities.

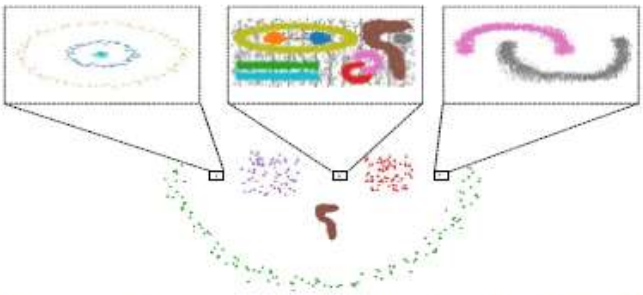

Fig. 5: The synthetic dataset 'ArtificialLocalN' with local noise in a dense region.

If I were to continue working on this paper, here is the next topic I would like to pursue to enrich this research:

One area that could be explored further is how DBHD performs on datasets with imbalanced class distributions. It would be interesting to see if the algorithm can effectively cluster datasets where one or more classes are significantly underrepresented.

Other new research topics or applications that this proposed solution could impact include:

An example application where DBHD could be useful is in identifying anomalies in network traffic data. By clustering network traffic data based on their local densities, it may be possible to identify patterns that are indicative of malicious activity or other anomalies. Another research topic that could build upon this work is developing new algorithms for density-based clustering that can handle high-dimensional data. Many real-world datasets have hundreds or thousands of features, which can make traditional clustering algorithms ineffective. Developing new algorithms that can handle high-dimensional data while maintaining accuracy and efficiency could have significant practical applications.