Fuzzy Density-Based Clustering in Dense Datasets: A Modified DBSCAN Algorithm

Erind Bedalli[1,*], Rexhep Rada[1], Luan Sinanaj[2]

[1]*Department of Informatics, University of Elbasan 'Aleksander Xhuvani', Elbasan, Albania*
[2]*Department of Information Technology, 'Aleksander Moisiu' University, Durres, Albania*

Abstract

In the context of unsupervised learning, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a well-established clustering algorithm that groups together data points which belong to dense regions, and denotes as noise the points located in low density regions. This algorithm is very convenient in detecting clusters of various shapes, including non-convex shapes which are challenging for many other cluster algorithms. However, in dense datasets, the assignment of data points into clusters may become abrupt. This paper introduces a modified version of the DBSCAN algorithm incorporating fuzzy membership degrees for points that are close to meeting the criterion of being part of a cluster. The core and border points are still assigned with a complete membership degree as in the classical DBSCAN, while some of the noise points will receive a fuzzy degree of membership based on the proportion of core, border, and noise points in their local neighborhood. The proposed approach is evaluated using several synthetic datasets to demonstrate its ability to provide a smoother cluster assignment in high-density scenarios.

1. Introduction

Clustering is an important form of unsupervised learning which aims at arranging the data points into clusters (subsets) such that instances within the same cluster are significantly more similar to each other compared to instances belonging to the other clusters. This is essentially a data-driven procedure as it is oriented merely by the distance or similarity measures that the points have with respect to each other, without any information about the intrinsic structures of the dataset being provided. Its contribution into a wide range of important problems such as customer profiling in marketing, image segmentation in computer vision, genomic data analysis in bioinformatics, clinical trial analysis in medicine etc. makes clustering a valuable and versatile technique [1].

Clustering plays a vital role in both exploration and summarization of data. Its flexibility makes it applicable on both small and large datasets, and as in the nowadays world data continues to grow in volume and complexity, clustering constitutes an essential method for pattern discovery and intrinsic structures explorations. Furthermore, clustering is frequently a key step in exploratory data analysis, with its results often serving as an intermediate output for further machine learning processes [2, 3].

Based on the way the clusters are conceived and on the way the cluster generation process is conducted, several important forms of clustering may be distinguished:

- Prototype-based clustering (centroid-based clustering) where each cluster is represented by a central point (centroid) and instances are assigned to the closest cluster. Some well-known prototype-based clustering algorithms are K-Means, K-Medoids, K-Means++ etc [4].

- Hierarchical clustering where a tree-like structure (dendrogram) is built by progressively merging smaller clusters into larger ones (agglomerative) or breaking larger clusters into smaller ones (divisive) [5].
- Density-based clustering where clusters are conceived as dense regions of instances separated by sparse regions considered as noise. Some well-known density-based clustering algorithms are DBSCAN, OPTICS, Mean-Shift clustering etc [6].
- Model-based clustering where the data are conceived as mixtures of underlying probability distributions (typically Gaussians) and the assignment of the points into clusters is done based on statistical likelihoods. Some well-known algorithms include Expectation-Maximization, Bayesian Gaussian Mixture Models etc [7].
- Fuzzy Clustering where instances are allowed to belong to multiple clusters simultaneously with partial degrees of membership [8, 9].

The core idea of this work is to blend the partial membership approach of fuzzy clustering into a density-based clustering algorithm (DBSCAN) aiming to capture clusters of various shapes and sizes and avoiding abrupt assignments. Although the DBSCAN is a robust and intuitive algorithm, it is sensitive to the choice of its hyper-parameters, therefore a fine-tuning procedure of these parameters is crucial to the quality of the generated clusters. Nevertheless, even with fine-tuning, the risk of an abrupt assignment for boundary points is still present. The partial membership approach introduces a gradual assignment policy in the border region, ensuring that the points that are close to meeting the criterion, will not be categorized as noise, but instead are assigned a partial membership. The gradual assignment will be a policy considering the quantitative presence of border and noise points in the neighborhood, as well as the homogeneity of these points.

The paper continues in the second section with a literature review of the most relevant research works related to fuzzy extensions applied in the field of density-based clustering algorithms. The third section follows with a theoretical overview of the classical DBSCAN algorithm, highlighting its main workflow, applicability, and limitations. The proposed fuzzy modifications on DBSCAN, are introduced in the fourth section, describing the evaluation process of fuzzy membership values and how the classical algorithm is modified via the incorporation of these values. The fifth section covers a series of experimental studies conducted on various synthetic datasets comprising intrinsic structures of non-convex shapes and an increased ratio of boundary values. These experimental studies compare the quality of the generated clusters by the classical and modified DBSCAN algorithms, based on the generalized silhouette score performance measure. The paper concludes with a discussion of the relevance of the findings, the challenges and limitations inherent in their applicability, as well as potential directions for future work.

2. Related work

The idea of a fuzzy approach to density-based clustering algorithms is not new to the machine learning community; several modifications on DBSCAN and other density-based algorithms have been presented in previous works. In this section, the main approaches to fuzzy modifications of density-based clustering proposed in various studies will be described, and the differences in our approach will be highlighted.

H.P. Kriegel et al. have proposed the F-DBSCAN algorithm which is capable of operating on vague data such as sensor databases or biometric information systems. The central idea was the integration of a fuzzy distance function into the density-based algorithm [10].

E. Nasibov et al. have proposed initially the Fuzzy Joint Point methods and have revised and optimized this methodology in several of their subsequent works. In addition, the same authors have proposed the FN-DBSCAN algorithm, a fuzzy neighborhood where points are allowed partial

membership into clusters based on the distance from the nearest points in the clusters. In all their approaches the key idea is the evaluation of the partial memberships based on the comparison of the distances of the neighborhood points to the overall distribution inside a cluster and they have rendered these techniques more robust alleviating the sensitivity to the choice of hyper-parameters [11-12].

A. Smiti and Z. Eloudi have also presented the idea of fuzzy neighborhood where partial memberships are also evaluated based on the distances, but instead of the classical Euclidean distance function, they have employed the Mahalanobis distance function which is more adaptable to various distributions [13].

S. Jebari et al. extend these ideas further proposing the AF-DBSCAN (Automatic Fuzzy DBSCAN) algorithm which strives to automatically determine the hyper-parameters in the FN-DBSCAN algorithm based on the k-neighbors plots [14].

G. Bordogna and D. Ienco have proposed the idea of utilizing the minimum number of points hyper-parameter to evaluate the partial memberships in the fuzzy neighborhood, but without discriminating between border and noise points [15].

In addition, there exist more specialized approaches such as the TSF-DBSCAN (Temporal Streaming Fuzzy DBSCAN) by A. Bechini et al., which is applied for the fuzzy clustering of streaming data [16]. The approach presented in this paper is in the same direction as in the work by G. Bordogna and D. Ienco, thus utilizing the minimum number of points hyper-parameter for the evaluation of the partial memberships, but adding as significant novelties the discrimination between border and noise points during the evaluation of partial memberships and also incorporating a penalty component for neighbors belonging to different clusters.

3. The classical DBSCAN algorithm

The classical DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm is a well-established clustering technique that conceives clusters as contiguous dense regions consisting of points packed closely to each-other separated by low-density regions consisting of noise, outliers or ideally just void. The algorithm has gained significant attraction in the machine learning community due to its capabilities of capturing clusters of arbitrary shapes and sizes, while being robust towards noise and outliers [17]. DBSCAN operates based on two hyper-parameters: ε (epsilon) which represents the radius of the neighborhood centered at the given point and minPts which represents the minimum number of neighbors within the ε radius required for a point to be considered a core point (i.e. the internal part of a dense region). At the first phase of the algorithm, each of the points of the dataset is categorized into one of these categories [18]:

- Core point, when there are at least minPts points within a circle of radius ε centered at the given point
- Border point, when a point does not reach the minPts threshold, but it has at least one core point in its ε neighborhood
- Noise, when the point does not qualify for being a core point or a border point.

Later, during the second phase of the algorithm, the clusters are constructed one by one starting with a random unassigned core point and progressively assigning to the current cluster all core points that are density-connected to the initial core point. The density-connection means that there exist a path of core points connecting the initial core point with some other core point, such that the length of each edge in this path does not exceed the value ε (as illustrated in Figure 1). This process continues "greedily" until no density-connected core point has remained. Afterwards the algorithm proceeds constructing another cluster starting with another random unassigned core point [19].
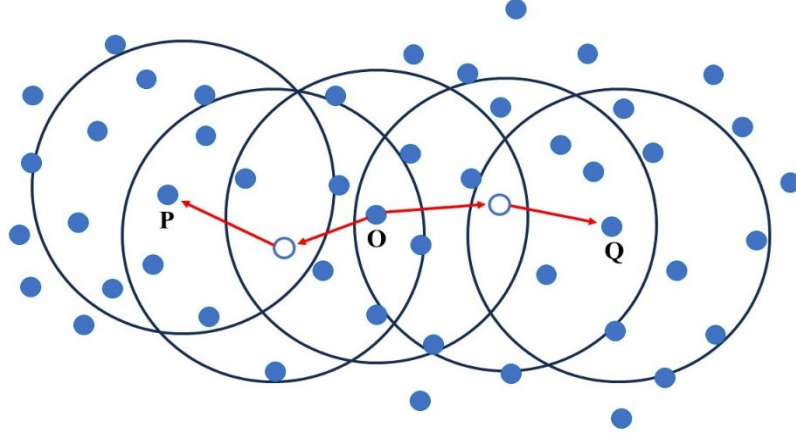
Figure 1: An illustration of density-connection

After all the core points are assigned into clusters, the border points are also assigned into clusters, with each border point assigned into the cluster of its closest core neighbor. Finally, the remaining points are marked as noise and are not assigned into any of the clusters. The entire DBSCAN algorithm can be summarized by the following pseudocode [20]:

1. Categorize all the points as core, border or noise based on the number of points located in their respective ε neighborhoods.
2. While there are unassigned core points, repeat:
   2.1 Randomly select an unassigned core point (denoting it x).
   2.2 Start a new cluster containing initially only x.
   2.3 Expand the current cluster adding all the other core points which are density-connected to x.
3. Assign each border point into the cluster of its closest core neighbor.
4. Leave all the noise points unassigned into any of the clusters.

4. A fuzzy modification of the DBSCAN algorithm

Despite of the desirable properties characterizing the classical DBSCAN such as the ability to capture clusters of arbitrary shapes and sizes, robustness towards noise and the automatic determination of the number of clusters, there are drawbacks such as an abrupt assignment of points into clusters. So, while two points are very close to each other, one may be assigned into one of the clusters, while the other may remain noise.

In order to improve this impediment, a fuzzy modification on the original DBSCAN is proposed. The idea is to assign partial memberships to the border points and to assign partially some of the noise points, which are close to being a border point. These partial memberships are evaluated based on the number of core, border and noise points located in the ε neighborhood of the respective point. Moreover, a penalty term is introduced for the cases when in the ε neighborhood there are points with different assignments.

More concretely, the membership value to be assigned to a border point in the case that all its core and border neighbors belong to the same cluster will be evaluated as:

$$\mu = \frac{w_c N_c + w_b N_b + w_n N_n}{(w_c + w_b + w_n)(N_c + N_b + N_n)} \tag{1}$$

Here N_c denotes number of core points in the ε neighborhood of the given border point, N_b denotes number of border points in the ε neighborhood of the given border point and N_n denotes the

number of noise points in the ε neighborhood of the given border point. On the other hand, w_c, w_b, w_n are respective weights to control the importance of the core, border and noise points. These three weights are expected as hyper-parameters by the modified fuzzy algorithm, and in absence of input they have the default values $w_c = 1.0$, $w_b = 0.55$, $w_n = 0.1$. General hyper-parameter tuning algorithms, such as grid search, are applicable in this context.

If the ε neighborhood contains core points or border points from several different clusters (symbolically denoted as 1,2, ..., k), then the calculation of the membership values will be carried out as follows for every $i \in \{1, 2, \ldots, k\}$:

$$\mu_i = \frac{w_{c_i} N_{c_i} + w_{b_i} N_{b_i} + w_n N_n}{(w_{c_i} + w_{b_i} + w_n)(N_{c_i} + N_{b_i} + N_n)} \tag{2}$$

Here N_{c_i} denotes number of core points belonging to i-th cluster in the ε neighborhood of the given border point, N_{b_i} denotes number of border points belonging to i-th cluster in the ε neighborhood of the given border point and N_n denotes the number of noise points in the ε neighborhood of the given border point. Naturally the presence of assignments into more than one clusters among the points in the ε neighborhood leads to penalization, i.e. smaller membership values as the core points or border points cannot have a joint contribution. On the other hand, the special phenomenon occurring in these circumstances is the partial membership into more than one cluster, an epitome of fuzzy clustering.

Additionally, the calculation of partial membership values for noise points would follow a similar logic but with the major distinction that there will be no core points. So, the calculation of the membership value of a noise point into the i-th cluster will be performed as:

$$\mu_i = \begin{cases} \dfrac{(w_{b_i} N_{b_i} + w_n N_n)}{(w_{b_i} + w_n)(N_{b_i} + N_n)}, \text{if } N_{b_i} > 0 \\ 0, \text{if } N_{b_i} = 0 \end{cases} \tag{3}$$

Based on the aforementioned modifications, now the pseudocode of the fuzzy modified DBSCAN algorithm will be:

1. Categorize all the points as core, border or noise based on the number of points located in their respective ε neighborhoods.
2. While there are unassigned core points, repeat:
   2.1 Randomly select an unassigned core point (denoting it x).
   2.2 Start a new cluster containing initially only x.
   2.3 Expand the current cluster adding all the other core points which are density-connected to x.
3. Assign each border point a partial membership according to equation 1 (if the neighbors are from the same cluster) or according to equation 2 (if the neighbors are from several different classes).
4. Assign each noise point a partial membership into clusters according to equation 3.
5. Mark all the points whose overall memberships are 0 (from equation 3) as noise

5. Experimental results

In order to assess the quality of the results generated by the fuzzy modified DBSCAN algorithm, a series of experimental studies were conducted on several synthetic datasets. These datasets are characterized by non-convex shapes and some 'disputable' points in the boundaries. The structure of these datasets is intentionally devised to be challenging in order to highlight the differences between the classical DBSCAN and the fuzzy modified DBSCAN. In Figure 2 below are shown the visualizations of these two algorithms on the first dataset consisting of 3 crescents (non-convex shapes), where the results of classical DBSCAN are shown in the left and the results of the fuzzy modified algorithm are shown on the right. The axes represent the natural coordinates (i.e. the two attributes that the points in the crescents dataset have). Each cluster is depicted with a separated color (red, green or blue), while noise points are depicted in black. Furthermore, in the fuzzy modified version, there are points with partial memberships which are depicted in lighter shades of the original color of the cluster they belong.
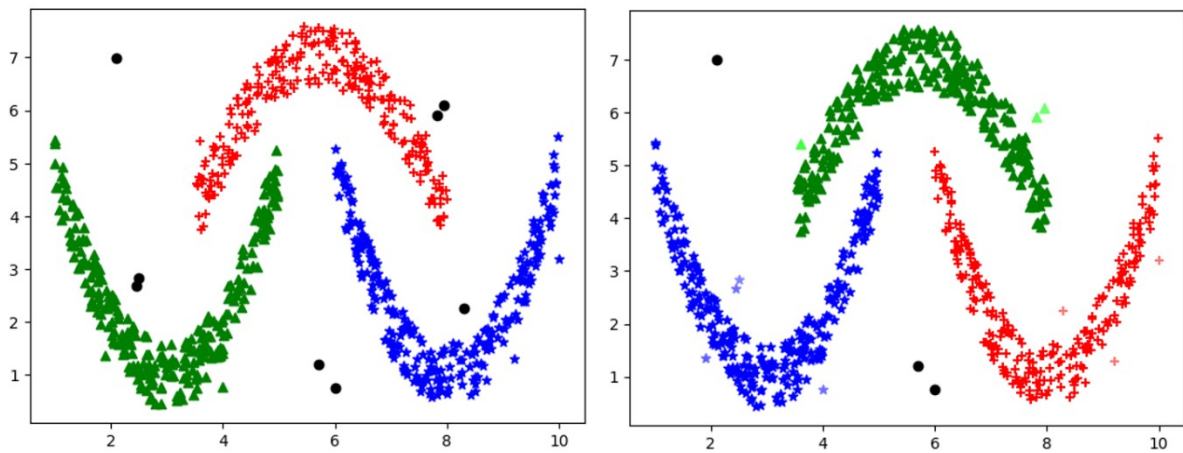
Figure 2 - Visualization of classical DBSCAN (left) and fuzzy modified DBSCAN (right)

As it can be easily noticed in the above visualizations, both algorithms properly capture the overall structures of the three clusters of this dataset as the core points are the same in both cases, while the differences lie in the assignment of border points versus noise points. In the left image can be noticed the abrupt assignment by the classical DBSCAN algorithm where some of the 'disputable' points have become full members of the respective clusters, while others are disqualified as noise (depicted in black color). In the right image can be noticed that the fuzzy modified DBSCAN algorithm assigns partial memberships to the 'disputable' points, depicting them as lighter shades of the cluster colors. Naturally, points far from the clusters will remain noise even in the fuzzy version of the algorithm (again depicted in black in the right image).

 Besides visual comparison, the classical DBSCAN algorithm and the fuzzy modified DBSCAN algorithms are compared using the silhouette score performance measure once that they are applied on the same dataset. In the following table are summarized all the synthetic datasets where the two algorithms were applied:

Table 1 - Summary of used datasets

Dataset	Number of attributes	Number of instances	Number of clusters	Description
Synth-1	2	814	3	3 crescents
Synth-2	2	2270	5	5 semicircles
Synth-3	2	2560	4	4 concentric circles
Synth-4	2	1725	2	2 spirals
Synth-5	3	3250	7	7 gaussian mixtures
Synth-6	3	3630	6	6 interlocking tori

For the comparison of the performance of these algorithms was used the fuzzy (generalized) silhouette score which is an extension of the classical silhouette score. Similarly to the classical silhouette score, the generalized silhouette score also aims to measure the performance of clustering by measuring how well each data point fits with the points of the same cluster compared to other clusters. For each point are taken into consideration the average distance from the other points belonging to the same cluster and the lowest among the average distance to points of some other cluster. The main difference is the adaption of the partial memberships in the calculations. More concretely, the calculation of the generalized silhouette score is applied as [21, 22]:

$$s_f(i) = \frac{b_f(i) - a_f(i)}{max(\ a_f(i),\ b_f(i))} \tag{3}$$

where the evaluation of $a_f(i)$ and $b_f(i)$ is generalized as:

$$a_f(i) = \sum_j \mu_{ic}\mu_{jc} d(i,j) \tag{4}$$

$$b_f(i) = \min_{k \neq c} \sum_j \mu_{ik}\mu_{jk} d(i,j) \tag{5}$$

To make the comparison fairer, in both cases the noise points are included in the evaluation, and their silhouette score a noise point takes the default value 0. The following table summarizes the silhouette scores of both classical DBSCAN and fuzzy modified DBSCAN for each algorithm.

Table 2 – Generalized silhouette scores of the algorithms

Dataset	Classical DBSCAN	Fuzzy modified DBSCAN
Synth-1	0.81	0.85
Synth-2	0.76	0.81
Synth-3	0.79	0.83

Synth-4	0.80	0.85
Synth-5	0.72	0.75
Synth-6	0.71	0.73

In the overall, it can be noticed that the generalized score is better for the fuzzy modified DBSCAN, compared to the classical DBSCAN. This is mainly due to the penalization that presence of noise points gives to the Classical DBSCAN, while the fuzzy version assigns partial memberships to some of the noise points.

6. Conclusions

This paper presented a fuzzy extension of the conventional DBSCAN algorithm aiming to improve the cluster assignment in dense data sets through the provision of partial memberships to boundary points and some noise points. Compared to the abrupt decision boundaries of conventional DBSCAN, the new method provides smoother assignments, especially for points placed near the boundaries of the clusters. By incorporating the weighted number of core, border, and noise points within the ε neighborhood, the algorithm presents smoother and insightful clustering results. Experimental analysis on a collection of synthetic datasets with complex structures demonstrated that both versions of DBSCAN recognize core points identically, but the fuzzy version handles boundary points better. Application of the generalized silhouette score with incorporation of noise points default scores, highlighted the superiority of the fuzzy approach in producing better clusters. These findings suggest that fuzzy DBSCAN is particularly well-suited for datasets with high density and indistinct cluster boundaries.

The mainline of this work is towards demonstration of the relevance of fuzzy modified algorithm in several datasets, but the classical challenge of the clustering problem is the detection of the circumstances where a clustering algorithm operates effectively. In order to make the given approach more robust, it should be carefully adapted with a preceding hyper-parameter tuning procedure and assessed by several performance measures.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] Ezugwu, Absalom E., Abiodun M. Ikotun, Olaide O. Oyelade, Laith Abualigah, Jeffery O. Agushaka, Christopher I. Eke, and Andronicus A. Akinyelu. "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects." *Engineering Applications of Artificial Intelligence* 110 (2022): 104743.

[2] Oyewole, Gbeminiyi John, and George Alex Thopil. "Data clustering: application and trends." *Artificial intelligence review* 56, no. 7 (2023): 6439-6475.

[3] Xu, Dongkuan, and Yingjie Tian. "A comprehensive survey of clustering algorithms." *Annals of data science* 2, no. 2 (2015): 165-193.

[4] Rada, Rexhep, Erind Bedalli, Sokol Shurdhi, and Betim Çiço. "A comparative analysis on prototype-based clustering methods." In *2023 12th Mediterranean Conference on Embedded Computing (MECO)*, pp. 1-5. IEEE, 2023.

[5] Ran, Xingcheng, Yue Xi, Yonggang Lu, Xiangwen Wang, and Zhenyu Lu. "Comprehensive survey on hierarchical clustering algorithms and the recent developments." *Artificial Intelligence Review* 56, no. 8 (2023): 8219-8264.

[6] Bhattacharjee, Panthadeep, and Pinaki Mitra. "A survey of density-based clustering algorithms." *Frontiers of Computer Science* 15 (2021): 1-27.

[7] McNicholas, Paul D. "Model-based clustering." *Journal of Classification* 33 (2016): 331-373.

[8] Ruspini, Enrique H., James C. Bezdek, and James M. Keller. "Fuzzy clustering: A historical perspective." *IEEE Computational Intelligence Magazine* 14, no. 1 (2019): 45-55.

[9] Bagherinia, Ali, Behrooz Minaei-Bidgoli, Mehdi Hosseinzadeh, and Hamid Parvin. "Reliability-based fuzzy clustering ensemble." *Fuzzy Sets and Systems* 413 (2021): 1-28.

[10] Kriegel, Hans-Peter, and Martin Pfeifle. "Density-based clustering of uncertain data." In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 672-677. 2005.

[11] Ulutagay, G., and E. Nasibov. "FN-DBSCAN: A novel density-based clustering method with fuzzy neighborhood relations." In *8th International Conference on Application of Fuzzy Systems and Soft Computing (ICAFS-2008)*, pp. 101-110. 2008.

[12] Nasibov, Efendi, Can Atilgan, Murat Ersen Berberler, and Resmiye Nasiboglu. "Fuzzy joint points based clustering algorithms for large data sets." *Fuzzy sets and Systems* 270 (2015): 111-126.

[13] Smiti, Abir, and Zied Eloudi. "Soft DBSCAN: Improving DBSCAN clustering method using fuzzy set theory." In *2013 6th International Conference on Human System Interactions (HSI)*, pp. 380-385. IEEE, 2013.

[14] Jebari, Sihem, Abir Smiti, and Aymen Louati. "AF-DBSCAN: An unsupervised Automatic Fuzzy Clustering method based on DBSCAN approach." In *2019 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, pp. 000001-000006. IEEE, 2019.

[15] Bordogna, Gloria, and Dino Ienco. "Fuzzy core DBSCAN clustering algorithm." In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 100-109. Cham: Springer International Publishing, 2014.

[16] Bechini, Alessio, Francesco Marcelloni, and Alessandro Renda. "TSF-DBSCAN: A novel fuzzy density-based approach for clustering unbounded data streams." *IEEE Transactions on Fuzzy Systems* 30, no. 3 (2020): 623-637.

[17] Gan, Junhao, and Yufei Tao. "DBSCAN revisited: Mis-claim, un-fixability, and approximation." In *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, pp. 519-530. 2015.

[18] Khan, Kamran, Saif Ur Rehman, Kamran Aziz, Simon Fong, and Sababady Sarasvady. "DBSCAN: Past, present and future." In *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*, pp. 232-238. IEEE, 2014.

[19] Bedalli, Erind, Enea Mançellari, and Esteriana Haskasa. "Exploring user feedback data via a hybrid fuzzy clustering model combining variations of FCM and density-based clustering." In *Advances in Intelligent Networking and Collaborative Systems: The 10th International Conference on Intelligent Networking and Collaborative Systems (INCoS-2018)*, pp. 71-81. Springer International Publishing, 2019.

[20] Bedalli, Erind, Enea Mançellari, and Rexhep Rada. "A semi-supervised fuzzy clustering approach via modifications of the DBSCAN algorithm." In *International Conference on Theory and*

Application of Soft Computing, Computing with Words and Perceptions, pp. 229-236. Cham: Springer International Publishing, 2019.

[21] Shahapure, Ketan Rajshekhar, and Charles Nicholas. "Cluster quality analysis using silhouette score." In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pp. 747-748. IEEE, 2020.

[22] Vardakas, Georgios, Ioannis Papakostas, and Aristidis Likas. "Deep clustering using the soft silhouette score: Towards compact and well-separated clusters." *arXiv preprint arXiv:2402.00608* (2024).