A Novel Fuzzy Adaptive Density-Based Clustering Algorithm

Mohamed-Ali Belloum, Valentin Fouillard, Jean-Philippe Poli

Université Paris-Saclay
CEA List
F-91120 Palaiseau, France
{*firstname.lastname*}@cea.fr

***Abstract*—Density-based clustering methods are of paramount importance in data science. They present the advantage of automatically determining the number of clusters, relying on a definition of density. Among these methods, DBSCAN also detects outliers. However, it suffers from hyperparameters that may be difficult to set, in particular by end-users. Different fuzzy approaches to DBSCAN have been proposed to fuzzify the frontiers between clusters and also the hyperparameters. In this paper, we present FA-DBC (Fuzzy Adaptive Density-Based Clustering), a fuzzy density-based clustering algorithm with self-adaptive thresholding, which significantly reduces manual parameter tuning. We compare the performance of this algorithm on different datasets.**

***Index Terms*—Fuzzy clustering, density-based clustering, unsupervised learning, adaptive threshold.**

I. Introduction

Clustering plays an important role in data science and in particular in the induction of fuzzy rule-based systems. Many rule induction approaches rely on data fuzzy clustering to define the linguistic variables, for instance before inducing fuzzy decision trees [1], or to directly extract the rules [2], [3]. Classical clustering methods are commonly categorized into partition-based (e.g., k-means [4]), hierarchical [5], density-based (e.g., DBSCAN [6]), and fuzzy approaches (e.g., Fuzzy C-Means [7]). Each family presents specific advantages and limitations. Partition-based algorithms are computationally efficient but require the number of clusters to be predefined and perform well mainly for convex-shaped clusters. Hierarchical methods produce dendrograms that expose nested structures but tend to be sensitive to noise and computationally demanding. Density-based methods are effective in discovering clusters of arbitrary shapes and detecting outliers; however, they often struggle with datasets containing clusters of varying densities and depend strongly on the tuning of sensitive parameters [8].

Despite substantial progress in clustering research, several fundamental challenges persist. First, most clustering algorithms require user-specified parameters that critically influence results. DBSCAN, for instance, requires the neighborhood radius (ε) and minimum points (MinPts) parameters, whose optimal values are dataset-dependent and difficult to determine a priori. Then, real-world datasets often contain clusters with significantly different densities. Traditional density-based methods using global density thresholds fail to correctly identify all clusters in such scenarios [9], [10]. Moreover, classical algorithms assign each point to exactly one cluster (or noise), which inadequately represents boundary points that may reasonably belong to multiple clusters [11], [12]. Finally, many clustering algorithms lack a principled approach for classifying new data points after initial clustering, necessitating complete re-clustering [13].

In this paper, we propose a novel algorithm called Fuzzy Adaptive Density-Based Clustering (FA-DBC) that combines density-based spatial clustering with fuzzy set theory to enable soft cluster assignments while maintaining the ability to detect arbitrarily shaped clusters. We develop an adaptive density threshold calculation method enabling effective handling of clusters with varying densities within a single dataset. Our approach incorporates fuzzy membership functions to provide flexible density estimation and to enable smooth transitions between clusters and noise regions. Furthermore, we extend the fuzzy assignment principle to perform multi-criteria allocation of unassigned points and to provide a principled prediction mechanism for efficiently classifying new data points.

The paper is structured as follows. Section II presents the related work, limited to fuzzy density-based clustering methods. The FA-DBC algorithm is presented in Section III. Section IV presents the results of our experiments, which are discussed in Section V before drawing some conclusions and perspectives.

II. Related Work

We focus on density-based clustering methods since our method belongs to this category, whose seminal crisp algorithm is DBSCAN [6]. DBSCAN excels at discovering arbitrary-shaped clusters and handling noise but suffers from several limitations such as its sensitivity to hyperparameters and its difficulty with varying-density clusters [8]. Many clustering approaches are derived from DBSCAN: for instance, HDBSCAN [14] uses hierarchical density estimates with improved handling of varying densities, and more recently deep

This research was funded by the French National Research Agency under the "France 2030 Initiative" and the "DIADEM Program", grant number 22-PEXD-0016 ("ARTEMIS").

density-based approaches have also been proposed [15]. To tackle the problems related to hyperparameters, self-adaptive DBSCAN [16] uses k-nearest neighbor statistics to estimate local ϵ values, while Adaptive DBSCAN [17] employs learning-based models for regional parameter adjustment. Zhang et al. [18] used reinforcement learning to dynamically tune clustering parameters during execution.

Several fuzzy extensions of DBSCAN have been proposed with different approaches to integrate fuzziness. *$\mathcal{F}$-DBSCAN* [19] is dedicated to uncertain data and uses fuzzy distance measures defined as probabilities of density-reachability. However, it maintains binary density counting and was designed specifically for uncertain object positions rather than general fuzzy density estimation. *FN-DBSCAN* [20] employs fuzzy neighborhood relations with membership functions on normalized distances. It uses fuzzy cardinality for core point identification, which can create fuzzy clusters with heterogeneous density characteristics and allows overlapping clusters. Points can belong to multiple clusters with different membership degrees. *AF-DBSCAN* [21] extends FN-DBSCAN with automatic parameter estimation based on k-nearest neighbor plots, addressing the parameter sensitivity issue. Soft-DBSCAN [22] introduces a fuzzy variant of DBSCAN where the MinPts constraint is relaxed, allowing for smoother membership degrees and partial cluster assignment. We include it in the benchmark as a reference for fuzzy density-based clustering approaches.

Fuzzy Core DBSCAN, or *FCore* [23], fuzzifies the number of samples in a neighborhood to be considered as a core using a soft constraint while keeping distance crisp. It generates non-overlapping fuzzy clusters with fuzzy cores where membership degrees reflect variable local densities. This approach is complementary to FN-DBSCAN, which fuzzifies distance but not this hyperparameter.

The authors of [24] proposed two fuzzy extensions of DBSCAN. The first, *FBorder*, relaxes the neighborhood radius (ϵ) to allow the generation of clusters with overlapping borders. The second, *FDBSCAN*, combines FBorder and FCore to generate clusters with both fuzzy cores and fuzzy overlapping borders. In the next section, we describe our novel approach.

III. FA-DBC Algorithm

A. Algorithm Outline

As any density-based clustering, the idea of our method is that a cluster is a group of high point density, separated from the other clusters by sparse regions. It thus relies on the notions of density and neighborhood.

Our approach, FA-DBC introduces three key novelties compared to existing fuzzy DBSCAN variants:

- it integrates fuzzy membership functions directly into the density computation, rather than applying them post-hoc or limiting fuzziness to distance measures (as in *$\mathcal{F}$-DBSCAN* and FN-DBSCAN);
- it incorporates almost fully automatic parameter estimation, reducing the need for manual parameter tuning and eliminating the need for user-specified soft constraints (unlike Fuzzy Core DBSCAN and FN-DBSCAN);
- it employs an adaptive, prototype-based multi-criteria assignment rule for core, border, and noise points, enabling a principled treatment of heterogeneous density regions.

Let $X = \{x_1, x_2, \ldots, x_n\}$ denote a dataset of n points in d-dimensional Euclidean space, where $x_i \in \mathbb{R}^d$ for $i = 1, \ldots, n$. We denote the Euclidean distance between points x_i and x_j as $d(x_i, x_j)$. FA-DBC employs the following hyperparameters: $k \in \mathbb{N}$, the number of neighbors used for neighborhood-radius estimation; $\mu(\cdot)$, the fuzzy membership function used in fuzzy density computation; and $\phi(\cdot)$, the prototype-based affinity function used in the multi-criteria fuzzy reassignment phase.

All other quantities, including the neighborhood radius $\epsilon \in \mathbb{R}^+$, the adaptive density threshold τ, and the final acceptance threshold $\eta_{\min}$, are automatically estimated from the data. The pseudocode, with the reassignment as a mandatory step, is presented in Algorithm 1.

Algorithm 1 Fuzzy Adaptive Density-Based Clustering with mandatory multi-criteria fuzzy reassignment

Require: Dataset $X = \{x_1, \ldots, x_n\}$, parameters k, μ, ϕ
Ensure: Cluster assignments $C_1, \ldots, C_M$, prototypes $\{p_k\}$, noise set N
Estimate ϵ *(see sec. III-C1)*
Compute fuzzy density for each point *(see sec. III-B)*
Compute adaptive threshold τ *(see sec. III-C2)*
// Cluster formation
Mark all points $x_i \in X$ as unvisited
$M \leftarrow 0$ (cluster count)
for each unvisited point $x_i \in X$ **do**
  if $\rho(x_i) \geq \tau$ **then**
    $M \leftarrow M + 1$
    Initialize new cluster $C_M \leftarrow \{x_i\}$
    Initialize queue Q with ϵ-neighbors(x_i)
    while $Q \neq \emptyset$ **do**
      $x_j \leftarrow$ dequeue from Q
      if x_j unvisited **then**
        Add x_j to C_M and mark as visited
        if $\rho(x_j) \geq \tau$ **then**
          Enqueue ϵ-neighbors(x_j) to Q
        end if
      end if
    end while
  end if
end for
$N \leftarrow X \setminus \bigcup_{k=1}^{M} C_k$ (noise points)
Compute the prototype of each cluster *(see sec. III-E)*
// Multi-criteria fuzzy reassignment of initially noise points (see sec. III-F)
for each point $x_i \in N$ **do**
  Compute global affinity to each cluster using $\phi(d(x_i, p_k))$
  Compute local neighborhood support
  Assign x_i to clusters with membership degrees proportional to combined affinity
end for
Update cluster sets $C_1, \ldots, C_M$ accordingly
Remove reassigned points from N
return Clusters $\{C_1, \ldots, C_M\}$, prototypes $\{p_k\}$, remaining noise N

B. Fuzzy Density Assessment

Unlike traditional density-based methods that use binary counting within neighborhoods, FA-DBC computes a *fuzzy density* where each neighbor contributes to a certain extent with a membership degree based on its distance.

Definition 1 (ϵ-neighborhood [6])**.** *The ϵ-neighborhood of point $x_i \in \mathcal{X} = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ is:*

$$N_\epsilon(x_i) = \{x_j \in \mathcal{X} : d(x_i, x_j) \leq \epsilon\} \tag{1}$$

where $d(\cdot,\cdot)$ is a distance metric (typically Euclidean).

Definition 2 (Fuzzy Density)**.** *The fuzzy density of a point x_i is defined as:*

$$\rho(x_i) = \frac{1}{|N_\epsilon(x_i)|} \sum_{x_j \in N_\epsilon(x_i),\, j \neq i} \mu\left(\frac{d(x_i, x_j)}{\epsilon}\right) \tag{2}$$

where $\mu : [0,\infty) \to [0,1]$ is a decreasing fuzzy membership function satisfying $\mu(0) = 1$.

The fuzzy density represents the fuzzified number of neighbors, where closer points have higher membership degrees. Unlike binary counting in DBSCAN (with crisp membership), each neighbor x_j contributes to a certain extent regarding $\mu(d(x_i, x_j)/\epsilon) \in [0,1]$. This provides a smooth, gradual transition in density estimates rather than discontinuous jumps, addressing a key limitation of *$\mathcal{F}$-DBSCAN* [19] that uses fuzzy distances but maintains binary density counting.

C. Parameter Estimation

1) Bandwidth selection via k-Distance: The k-distance for a point x_i is defined as the distance between x_i and its k-th nearest neighbor, i.e., $d_k(x_i) = d(x_i, x_i^{(k)})$, where $x_i^{(k)}$ denotes the k-th nearest neighbor of x_i [6]. It reflects the local density around each point: points in dense regions have small k-distances, whereas points in sparse regions or near cluster boundaries exhibit larger ones. Sorting all k-distances in ascending order typically yields a curve with a marked rise. The transition point, often referred to as the "knee" or "elbow", is used as a data-driven estimate of the neighborhood radius ε.

Proposition 1 (Heuristic knee-point interpretation)**.** *The knee point in the sorted k-distance plot approximates the transition from intra-cluster to inter-cluster distances. For well-separated clusters with density ratio $\lambda = \rho_{\min}/\rho_{\max}$, the knee estimator satisfies:*

$$\epsilon_{knee} = O\left(\lambda^{-1/d}\right) \tag{3}$$

This proposition should be understood as a heuristic justification under simplified separation assumptions, rather than as a full finite-sample guarantee.

In the implementation, the neighborhood radius ϵ is estimated from the sorted k-distance curve using a knee-detection procedure (`KneeLocator`) configured for an increasing convex profile [25]. When k is not provided, we use the default rule $k = \max(5, \lceil \log n \rceil)$. If no knee is detected, a fallback based on the $80th$ percentile of the sorted k-distances is used; if knee detection fails numerically, a robust alternative based on the median and interquartile range is applied.

This automatic parameter estimation addresses a key limitation of FN-DBSCAN [20] and is similar in spirit to AF-DBSCAN [21], but applied to our fuzzy density framework.

2) Weighted fuzzy density threshold: Unlike DBSCAN's fixed minimum number of points and Fuzzy Core DBSCAN's user-specified soft constraint [23], we compute a fully automatic adaptive threshold based on the fuzzy density distribution:

$$\tau = \frac{\sum_{i:\rho(x_i)>0} \rho(x_i)^2}{\sum_{i:\rho(x_i)>0} \rho(x_i)}. \tag{4}$$

This weighted threshold gives more influence to high-density points than to boundary or sparse points. As a result, τ is biased toward the density level of core regions rather than transition areas, which makes it suitable for core-point detection in heterogeneous-density settings.

This weighted average emphasizes high-density regions, automatically adapting to the fuzzy density distribution. The quadratic weighting ensures that the threshold is pulled toward the density of core cluster regions rather than boundary areas.

D. Core Point Detection and Cluster Expansion

Definition 3 (Core Point [6])**.** *A point x_i is a core point if $\rho(x_i) \geq \tau$, with $\rho(x_i)$ its fuzzy density.*

Definition 4 (Density-Reachable [6])**.** *Point x_j is density-reachable from core point x_i if there exists a chain of core points $x_i = x^{(1)}, x^{(2)}, \ldots, x^{(m)} = x_j$ such that $x^{(t+1)} \in N_\epsilon(x^{(t)})$ for all t.*

Cluster expansion proceeds by depth-first propagation through core points only. Border points (non-core points in neighborhoods of core points) receive cluster labels but do not propagate, ensuring cluster coherence [6]. Once the cluster structure has been determined through density-based core expansion, each resulting cluster is summarized by a representative element selected according to its fuzzy density characteristics.

E. Prototype Selection

For each discovered cluster C_k, the prototype is the point with maximum fuzzy density:

$$p_k = \arg\max_{x_i \in C_k} \rho(x_i). \tag{5}$$

This choice is motivated by the mode-seeking principle in fuzzy clustering [26] and ensures prototypes represent cluster centers in fuzzy density space. The prototype serves as the representative element with highest fuzzy neighborhood support.

F. Second Chance Multi-Criteria Fuzzy Assignment

Points initially marked as noise after the core expansion phase — i.e., neither core nor border points — are reassigned using a multi-criteria fuzzy strategy that combines prototype-based global affinity and local neighborhood support. This step is particularly important in datasets with heterogeneous densities or ambiguous boundary regions, and it also provides a principled mechanism for extending the clustering to new data points.

Let $\{C_k\}_{k=1}^{K}$ denote the set of discovered clusters and p_k their associated prototypes, as introduced in Section III-E.

Criterion 1: Prototype-Based Global Affinity. The first criterion evaluates the degree of global compatibility between an unassigned point and each cluster prototype. This quantity does not model local proximity but rather the attraction of a point toward the global representative of a cluster. We define a prototype affinity function $\phi : \mathbb{R}^+ \to [0, 1]$, satisfying the following properties: (i) $\phi(0) = 1$, indicating maximal affinity at the prototype, (ii) ϕ is monotonically non-increasing with respect to distance and (iii) $\lim_{t\to+\infty} \phi(t) = 0$. Without loss of generality, we adopt in this paper a Gaussian affinity function: $\phi(d) = \exp\left(-\frac{d^2}{2\sigma^2}\right)$, which is widely used in prototype-based fuzzy clustering models and radial basis formulations [27]. In the experiments reported in this paper, we set $\sigma = \epsilon/2$, which keeps the prototype-based affinity on a scale consistent with the neighborhood radius ϵ.

For each unassigned point x_i, the normalized global affinity to cluster C_k is defined as

$$u_{ik}^{(\text{proto})} = \frac{\phi(d(x_i, p_k))}{\sum_{j=1}^{K} \phi(d(x_i, p_j))}. \tag{6}$$

Criterion 2: Proximity-Based Local Support. While global prototype affinity captures coarse cluster structure, it may be unreliable for points located in sparse regions or near cluster boundaries. To incorporate local geometric information, we introduce a second criterion based on neighborhood support. For this purpose, we consider a local proximity set $\mathcal{P}(x_i; \kappa_{\text{prox}})$, defined as a finite set of nearest already assigned points to x_i gathered during the reassignment phase. This local support is distinct from the parameter k used in the estimation of ϵ. In the experiments, we use a small fixed value of κ_{prox} so that the reassignment step remains local without introducing an additional data-dependent hyperparameter. Let $C_k = \{x_j \in X : \text{label}(x_j) = k\}$.

For an unassigned point x_i, the local proximity support for cluster C_k is defined as

$$u_{ik}^{(\text{prox})} = \max_{x_j \in \mathcal{P}(x_i;\kappa_{\text{prox}}) \cap C_k} \mu\left(\frac{d(x_i, x_j)}{\epsilon}\right), \tag{7}$$

where μ is the fuzzy membership function used in the density estimation step (Section III-B). If no assigned neighbor from C_k exists, we set $u_{ik}^{(\text{prox})} = 0$.

Adaptive Fusion of Global and Local Criteria. The final membership degree is obtained by adaptively combining the global affinity and the local support:

$$u_{ik}^{(\text{final})} = \alpha_i \, u_{ik}^{(\text{proto})} + (1 - \alpha_i) \, u_{ik}^{(\text{prox})}, \tag{8}$$

where $\alpha_i = \max_j u_{ij}^{(\text{proto})}$. The coefficient α_i acts as a confidence indicator for the global prototype-based affinity. Points close to a prototype yield $\alpha_i \approx 1$, resulting in a dominance of global structure, whereas points far from all prototypes yield $\alpha_i \approx 0$, in which case local neighborhood support drives the assignment.

1) Assignment Rule and Noise Rejection: A point x_i is assigned to cluster

$$k^* = \arg\max_k u_{ik}^{(\text{final})} \textbf{ if } u_{ik^*}^{(\text{final})} \geq \eta_{\min}. \tag{9}$$

Equation (9) assigns an initially unlabeled point only when its strongest final membership is sufficiently large. This avoids forcing weakly supported points into a cluster during the reassignment phase. Points that fail to satisfy (9) are labeled as noise or outliers.

The minimum acceptance threshold $\eta_{\min}$ is computed automatically using a fuzzy-density-weighted statistic:

$$\eta_{\min} = \text{clip}\left(\frac{\sum_{i=1}^{n} \rho(x_i) \, \max_k u_{ik}}{\sum_{i=1}^{n} \rho(x_i)}, \, 0.05, \, 0.5\right), \tag{10}$$

where clipping prevents degenerate decisions while preserving the fuzzy interpretation. Equation (10) defines $\eta_{\min}$ as a density-weighted acceptance level. High-density points contribute more to this statistic, so the resulting threshold reflects the confidence typically observed in well-supported regions while remaining bounded away from degenerate values through clipping.

IV. Experimental Validation and Benchmarking

We report external and fuzzy clustering metrics computed against ground-truth labels. Internal validation indices (such as Silhouette, Davies–Bouldin, or Calinski–Harabasz) are not reported in this work. For fuzzy algorithms (FDCA, FCM, FN-DBSCAN, fuzzy DBSCAN variants), crisp metrics (ARI, AMI) are computed by assigning each object to the cluster with maximum membership (hardening by $\arg\max_k u_{ik}$). Conversely, for crisp algorithms (DBSCAN, HDBSCAN), fuzzy metrics (FARI, fuzzy F-measure, FPI) are computed from a one-hot membership matrix where each object has membership 1 to its predicted cluster and 0 to the others. We use the following external metrics:

- **Adjusted Rand Index** (ARI) **[28]**: measures pairwise agreement between predicted and true labels, corrected for chance. Values close to 1 indicate nearly perfect agreement, values around 0 correspond to random partitions, and negative values indicate worse-than-random agreement.
- **Adjusted Mutual Information** (AMI) **[29]**: measures the mutual information shared by predicted and true clusters, also adjusted for chance. Its interpretation is similar to ARI, but AMI is more robust to highly unbalanced cluster sizes.

We use the following fuzzy metrics to evaluate the fuzzy partition:

- **Fuzzy F-measure [24]**: combines fuzzy precision (how well clusters correspond to ground-truth classes) and fuzzy recall (how well each class is covered by the clusters). Larger values indicate better agreement between the fuzzy clustering and the labels.
- **Fuzzy Partition Index** (FPI) **[30]**: normalized version of the partition coefficient that measures how crisp the fuzzy partition is. Values close to 1 indicate sharp memberships (crisp-like clusters), whereas values closer to 0 indicate

TABLE I: Performance comparison on synthetic benchmark datasets (ARI / AMI / FARI; best in bold)

Dataset	Metric	Algorithm				
		DBSCAN	HDBSCAN	FN-DBSCAN	FCM	FA-DBC
Aggregation	ARI	0.81	0.87	0.68	0.74	**0.93**
	AMI	0.89	0.93	0.82	0.84	**0.94**
	FARI	0.82	0.84	0.70	0.80	**0.93**
Compound	ARI	0.74	**0.81**	0.53	0.52	0.80
	AMI	0.80	**0.83**	0.70	0.72	**0.83**
	FARI	0.77	0.74	0.55	0.61	**0.79**
Flame	ARI	**0.98**	**0.98**	0.50	0.49	0.93
	AMI	**0.96**	0.95	0.45	0.44	0.88
	FARI	**0.94**	0.69	0.50	0.68	0.93
Jain	ARI	0.94	**0.96**	0.54	0.30	0.82
	AMI	0.86	**0.87**	0.50	0.35	0.70
	FARI	**0.93**	0.86	0.55	0.31	0.81
Moons	ARI	**1.00**	**1.00**	0.50	0.26	**1.00**
	AMI	**1.00**	**1.00**	0.40	0.20	**1.00**
	FARI	**1.00**	**1.00**	0.50	0.31	0.73
R15	ARI	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
	AMI	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
	FARI	0.78	0.80	**0.99**	0.94	**0.99**

highly diffuse memberships. FPI is therefore interpreted as a measure of boundary sharpness rather than clustering correctness, and must be read jointly with external fuzzy metrics such as fuzzy F-measure or FARI.

- **Fuzzy Adjusted Rand Index** (FARI) **[31]**: soft version of ARI that accounts for partial memberships while preserving adjustment for chance. It replaces hard co-occurrence counts by fuzzy co-association degrees derived from the fuzzy partition and a one-hot encoding of the true labels.

For the fuzzy DBSCAN variants reported in [24] (FCore, FBorder, FDBScan, SOFT-DBSCAN), we follow the original evaluation protocol and report fuzzy-specific indices (fuzzy F-measure, FPI). ARI/AMI/FARI are not reported for these methods because they are not available in [24], and our goal is to enable a direct comparison under the same experimental conditions. We evaluate FA-DBC on two categories of datasets:

- **Real-world UCI datasets**: Breast Cancer Wisconsin (569 samples, 30 features), Diabetes (768, 8), Ecoli (336, 7), Glass (214, 9), Iris (150, 4), Parkinsons (195, 22), and Vehicle (846, 18).
- **Synthetic benchmark datasets**: standard shape-based datasets (Aggregation, Compound, Flame, Jain, Moons, R15) and large-scale varying-density datasets (Cluto-t4-8k and Cluto-t8-8k with 8000 samples each).

All features are standardized using z-score normalization prior to clustering. This preprocessing is important for FA-DBC and the distance-based baselines because neighborhood construction and density estimation are sensitive to feature scaling.

We compare FA-DBC against five baseline algorithms:

- **DBSCAN**: parameters $(\epsilon, \text{min\_samples})$ are optimized via grid search over $\epsilon \in [0.1, 2.0]$ (20 values) with min_samples $= 5$, maximizing ARI on ground-truth labels. This provides an optimistic upper bound for DBSCAN performance.

TABLE II: Performance comparison on large-scale synthetic datasets with varying densities (ARI / AMI / FARI; best in bold)

Dataset	Metric	Algorithm				
		DBSCAN	HDBSCAN	FN-DBSCAN	FCM	FA-DBC
Cluto-t4-8k	ARI	0.34	**0.67**	0.47	0.47	0.63
	AMI	0.51	**0.84**	0.60	0.60	0.79
	FARI	0.44	**0.66**	0.47	0.47	**0.66**
Cluto-t8-8k	ARI	0.30	0.79	0.47	0.47	**0.87**
	AMI	0.54	**0.88**	0.64	0.64	0.87
	FARI	0.42	0.74	0.48	0.48	**0.85**

- **HDBSCAN**: parameters (min_cluster_size, min_samples) are optimized via grid search over min_cluster_size $\in \{5, 10, 15, 20, 30\}$ and min_samples $\in \{5, 10, 15\}$, maximizing ARI on non-noise points (labels $\neq -1$). This yields a fair comparison with a modern hierarchical density-based method.
- **FN-DBSCAN**: a fuzzy DBSCAN variant with fuzzy neighborhoods and density thresholds. Parameters are set as $\epsilon_{\min} = 0.9\,\epsilon_{\text{best}}$, $\epsilon_{\max} = 1.1\,\epsilon_{\text{best}}$ (where ϵ_{best} is DBSCAN's optimal value), min_pts $= 5$, and max_pts $= 7$, providing a fuzzy density baseline with comparable parameter complexity. This choice ensures a comparable scale with DBSCAN and avoids introducing additional tuning complexity.
- **Fuzzy C-Means (FCM)**: the number of clusters is set to the true number of classes. We use the standard choice $m = 2$, which is the most common setting in the literature and provides a conventional reference baseline. A dedicated sensitivity study with respect to m is left for future work.
- **FA-DBC**: almost fully automatic parameter selection. The only user-defined hyperparameter is k, used for neighborhood-radius estimation. However, we used the heuristic defined Section III-C to estimate it. We use a Gaussian membership function for fuzzy density estimation and set $\sigma = \epsilon/2$ in the prototype-affinity term; all remaining quantities are inferred automatically from the data.

For datasets with ground-truth labels, we report **ARI**, **AMI**, and **FARI**. For fuzzy partitioning quality, we compute **FPI** and **Fuzzy F-Measure**. Higher FPI and Fuzzy F-Measure indicate better-defined fuzzy clusters. All experiments are implemented in Python 3.10 using scikit-learn 1.3 for standard clustering baselines, hdbscan 0.8.33 for HDBSCAN, and custom implementations for FN-DBSCAN and FA-DBC. All data and code are publicly available at: 10.5281/zenodo.18257171.

To contextualize FA-DBC within the fuzzy DBSCAN family, we include published results from [24] for four variants: **FCore** (fuzzy core points), **FBorder** (fuzzy border regions), **FDBScan** (combined fuzzy core/border), and **SOFT-DBSCAN** (soft MinPts constraint). Note that we do not include *$\mathcal{F}$-DBSCAN* in our benchmark because it represents an earlier variant whose principles are largely encompassed by FN-DBSCAN, while FDBScan provides a more direct state-of-the-art reference. Similarly, AF-DBSCAN extends FN-DBSCAN with automatic parameter tuning, but since our comparisons already include FN-DBSCAN at its optimal

TABLE III: Fuzzy F-Measure on real-world UCI datasets (FCore/FBorder/FDBScan/SOFT-DBSCAN from [24])

Dataset	DBSCAN	HDBSCAN	FCore	FBorder	FDBScan	SOFT-DBSCAN	FN-DBSCAN	FCM	FA-DBC
Breast Cancer	0.21	0.27	0.76	0.79	0.77	0.42	**0.91**	0.74	0.77
Diabetes	0.71	0.68	0.73	0.75	0.71	0.56	0.71	0.57	**0.77**
Ecoli	0.59	0.47	0.60	**0.61**	**0.61**	0.32	0.59	0.39	0.60
Glass	0.54	0.54	0.55	**0.74**	0.48	0.25	0.44	0.34	0.61
Iris	0.78	0.78	0.78	0.78	0.78	0.39	**0.84**	0.75	0.78
Parkinsons	0.40	0.20	0.79	**0.84**	0.78	0.46	0.59	0.60	0.83
Vehicle	0.40	0.10	0.41	0.41	0.41	0.30	0.38	0.33	**0.41**
Mean	0.52	0.43	0.66	**0.70**	0.65	0.39	0.64	0.53	0.68

TABLE IV: Fuzzy Partition Index (FPI) on real-world UCI datasets (FCore/FBorder/FDBScan/SOFT-DBSCAN from [24])

Dataset	DBSCAN	HDBSCAN	FCore	FBorder	FDBScan	SOFT-DBSCAN	FN-DBSCAN	FCM	FA-DBC
Breast Cancer	0.13	0.00	0.90	0.96	0.94	0.37	**1.00**	0.79	**1.00**
Diabetes	0.93	0.76	0.92	0.96	0.94	0.00	**1.00**	0.65	**1.00**
Ecoli	0.97	0.08	**1.00**	**1.00**	**1.00**	0.07	**1.00**	0.21	0.99
Glass	0.78	0.68	0.78	0.72	0.84	0.28	**1.00**	0.39	**1.00**
Iris	**1.00**	0.95	**1.00**	**1.00**	**1.00**	0.36	**1.00**	0.67	0.97
Parkinsons	0.24	0.15	0.94	0.72	0.94	0.12	**1.00**	0.53	0.98
Vehicle	0.97	0.00	**1.00**	**1.00**	**1.00**	0.04	**1.00**	0.59	**1.00**
Mean	0.72	0.37	0.93	0.91	0.95	0.18	**1.00**	0.55	0.99

settings, we omit a separate benchmark to avoid redundancy. These methods were evaluated on the same UCI datasets under comparable conditions, providing a reference for state-of-the-art fuzzy density-based clustering.

V. Discussion

Our experiments highlight three aspects: (i) fuzzy clustering quality on real-world UCI datasets, (ii) behavior on synthetic datasets with complex shapes and varying densities, and (iii) the impact of fuzziness (FPI) and noise handling.

On UCI datasets (Table III), FBorder attains the highest mean fuzzy F-measure (0.70), slightly above FA-DBC (0.68), using several tuned fuzzy parameters. FA-DBC relies on a single heuristic parameter and automatically infers fuzzy density and membership assignments, while clearly outperforming DBSCAN (0.52), HDBSCAN (0.43), and FN-DBSCAN (0.64). For partition quality (Table IV), FN-DBSCAN and FA-DBC reach near-perfect mean FPI (1.00 and 0.99), indicating well-separated fuzzy clusters, but FA-DBC achieves a higher fuzzy F-measure (0.68 vs. 0.64). HDBSCAN shows a much lower FPI (0.37) due to excessive noise classification.

On synthetic datasets with non-convex shapes and overlapping clusters (Aggregation, Compound, Flame, Jain, Moons, R15; Table I), FA-DBC is highly competitive with DBSCAN, HDBSCAN, and FCM in ARI, AMI, and FARI. It achieves the best scores on Aggregation and remains competitive on Compound, Flame, Jain, and Moons, although DBSCAN or HDBSCAN obtain the highest ARI/AMI on several of these datasets. These results suggest that FA-DBC does not dominate all baselines uniformly, but provides a favorable trade-off between clustering quality, fuzzy memberships, and reduced parameter tuning. On large-scale varying-density datasets (Cluto-t4-8k, Cluto-t8-8k; Table II), FA-DBC improves over DBSCAN and FN-DBSCAN and attains ARI and FARI comparable to or better than HDBSCAN.

FA-DBC is particularly attractive when one seeks a fuzzy density-based method with limited manual tuning, smooth treatment of boundary points, and robustness to heterogeneous-density structures. In contrast, on simple shape-based datasets with very favorable density separation, classical crisp density-based methods may still achieve slightly better hard-partition scores.

Noise handling differentiates density-based algorithms (DBSCAN, HDBSCAN, FN-DBSCAN, FA-DBC) from FCM, which assigns every point to a cluster. FN-DBSCAN and FA-DBC both produce high-quality fuzzy partitions (high FPI), but FA-DBC's multi-criteria assignment yields better clustering accuracy by adapting memberships to local density variations. HDBSCAN often over-classifies noise, degrading fuzzy F-measure and FPI, whereas FA-DBC's adaptive fuzzy density threshold gives low but non-zero memberships to borderline points, preventing excessive noise labeling and producing smoother fuzzy clusters.

The proposed approach also has limitations. Although most quantities are inferred automatically, the method still depends on the heuristic choice of k, and its behavior can be influenced by feature scaling. In addition, FA-DBC is defined procedurally rather than through the optimization of a single global objective, which makes its theoretical analysis less direct than that of objective-driven fuzzy clustering methods.

From a computational viewpoint, the most demanding steps of the proposed method are neighborhood construction, fuzzy density computation, and cluster expansion. The overall worst-case complexity is quadratic, i.e., $\mathcal{O}(n^2)$, due to pairwise distance computations involved in neighborhood queries and expansion. In practice, this cost can be significantly reduced by using spatial indexing structures such as KD-trees or Ball-trees, which accelerate nearest-neighbor searches. Therefore, the practical runtime is typically closer to sub-quadratic behavior depending on the data distribution and dimensionality.

A detailed empirical runtime analysis is left for future work.

In summary, compared to existing fuzzy density-based methods (FN-DBSCAN, Fuzzy Core DBSCAN, FDBSCAN, AF-DBSCAN, and the approach by Ienco and Bordogna [24]), FA-DBC combines reduced manual parameter tuning, adaptive membership assignment, and non-overlapping fuzzy clusters within a density-based fuzzy framework.

VI. Conclusion

We introduced FA-DBC, a fuzzy density-based clustering algorithm that combines adaptive density estimation, multi-criteria membership assignment, and almost fully automatic parameter selection. On real-world UCI datasets, FA-DBC achieves a higher mean fuzzy F-measure than DBSCAN, HDBSCAN, FN-DBSCAN, and FCM, and remains competitive with the best fuzzy DBSCAN variants while using only a single heuristic hyperparameter k for neighborhood-radius estimation. On synthetic datasets with complex shapes and varying densities, FA-DBC remains competitive with strong density-based baselines in ARI, AMI, and FARI, while additionally providing fuzzy memberships and reduced manual parameter tuning. Although these results on standard UCI and synthetic benchmarks ensure comparability with prior work, further experiments on larger and more heterogeneous data are needed to fully assess the limits of FA-DBC. Future work will investigate incremental and streaming extensions of FA-DBC, as well as applications to high-dimensional domains such as images and spatio-temporal data. FA-DBC provides a practical compromise between interpretability, flexibility, and reduced parameter tuning in density-based clustering.

References

[1] A. G. Sébert and J. Poli, "Material classification from imprecise chemical composition : Probabilistic vs possibilistic approach," in *2018 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2018, Rio de Janeiro, Brazil, July 8-13, 2018*, pp. 1–8, IEEE, 2018.

[2] J. Viaña, S. Ralescu, A. Ralescu, K. Cohen, and V. Kreinovich, "Explainable fuzzy cluster-based regression algorithm with gradient descent learning," *Complex Engineering Systems*, vol. 2, no. 2, 2022.

[3] O. Rousselle, J. Poli, and N. B. Abdallah, "Towards an interpretable fuzzy approach to experimental design," in *Information Processing and Management of Uncertainty in Knowledge-Based Systems - 20th International Conference*, vol. 1174 of *Lecture Notes in Networks and Systems*, pp. 219–232, Springer, 2024.

[4] J. B. McQueen, "Some methods of classification and analysis of multivariate observations," in *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.*, pp. 281–297, 1967.

[5] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[6] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "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, p. 226–231, AAAI Press, 1996.

[7] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media, 2013.

[8] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "Dbscan revisited, revisited: why and how you should (still) use dbscan," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 3, pp. 1–21, 2017.

[9] Y. Chen, Z. Cai, and Y. Li, "Density-ratio based clustering for discovering clusters with varying densities," *Pattern Recognition*, vol. 94, pp. 95–109, 2019.

[10] S. Li and H. Zhao, "A domain adaptive density clustering algorithm for data with varying density distribution (dadc)," *arXiv preprint arXiv:1911.10293*, 2020.

[11] M. Kearns, Y. Mansour, and A. Y. Ng, "An information-theoretic analysis of hard and soft assignment methods for clustering," in *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 282–293, 1997.

[12] N. Bore and N. Gupta, "A comparative study between fuzzy clustering algorithm and hard clustering algorithm," *arXiv preprint arXiv:1404.6059*, 2014.

[13] V. Cohen-Addad, B. Guedj, V. Kanade, and G. Rom, "Online k-means clustering," in *International Conference on Artificial Intelligence and Statistics*, pp. 1126–1134, PMLR, 2021.

[14] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia conference on knowledge discovery and data mining*, pp. 160–172, Springer, 2013.

[15] A. Beer, P. Weber, L. Miklautz, C. Leiber, W. Durani, C. Böhm, and C. Plant, "Shade: Deep density-based clustering," in *2024 IEEE International Conference on Data Mining (ICDM)*, pp. 675–680, IEEE, 2024.

[16] A. Ram, S. Jalal, A. S. Jalal, and M. Kumar, "A density based algorithm for discovering density varied clusters in large spatial databases," *International Journal of Computer Applications*, vol. 3, no. 6, pp. 1–4, 2010.

[17] T. Boonchoo, X. Ao, Y. Liu, W. Zhao, F. Zhuang, and Q. He, "Grid-based dbscan: Indexing and inference," *Pattern Recognition*, vol. 90, pp. 271–284, 2019.

[18] R. Zhang, H. Peng, Y. Dou, J. Wu, Q. Sun, Y. Li, J. Zhang, and P. S. Yu, "Automating dbscan via deep reinforcement learning," in *Proceedings of the 31st acm international conference on information & knowledge management*, pp. 2620–2630, 2022.

[19] H.-P. Kriegel and M. 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.

[20] E. N. Nasibov and G. Ulutagay, "Robustness of density-based clustering methods with various neighborhood relations," *Fuzzy Sets and Systems*, vol. 160, no. 24, pp. 3601–3615, 2009.

[21] S. Jebari, A. Smiti, and A. Louati, "Af-dbscan: An unsupervised automatic fuzzy clustering method based on dbscan approach," in *2019 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, IEEE, 2019.

[22] A. Smiti and Z. 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.

[23] G. Bordogna and D. Ienco, "Fuzzy core dbscan clustering algorithm," in *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 100–109, Springer, 2014.

[24] D. Ienco and G. Bordogna, "Fuzzy extensions of the dbscan clustering algorithm," *Soft Computing*, vol. 22, no. 5, pp. 1719–1730, 2018.

[25] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan, "Finding a" kneedle" in a haystack: Detecting knee points in system behavior," in *2011 31st international conference on distributed computing systems workshops*, pp. 166–171, IEEE, 2011.

[26] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[27] C. Borgelt, *Prototype-based classification and clustering*. PhD thesis, Magdeburg, Univ., Habil.-Schr., 2006, 2005.

[28] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

[29] N. X. Vinh and J. Epps, "Bailey, j2738784: Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. vol. 11," *J Mach Learn Res*, vol. 11, pp. 2837–2854, 2010.

[30] M. Roubens, "Fuzzy clustering algorithms and their cluster validity," *European journal of operational research*, vol. 10, no. 3, pp. 294–301, 1982.

[31] E. Bedalli, S. Hajrulla, R. Rada, and R. Kosova, "Fuzzy clustering approaches based on numerical optimizations of modified objective functions," *Algorithms*, vol. 18, no. 6, p. 327, 2025.