Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 14, No. 3, 2026, 453-463.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

DPC-FSNN: Adaptive Fuzzy Shared Nearest Neighbor Density Peak Clustering Algorithm

Afsaneh Shamsaddini-Farsangi and Mostafa Ghazizadeh-Ahsaee*

Department of Computer Engineering, Shahid Bahonar University of Kerman, Kerman, Iran.

Article Info

Article History:
Received 17 November 2025
Revised 05 July 2026
Accepted 30 July 2026

DOI:10.22044/jadm.2026.17159.2851

Keywords:
Clustering, Multi-density, Nearest Neighbor, Fuzzy Neighborhood.

**Corresponding author: mghazizadeh@uk.ac.irr (M. Ghazizadeh-Ahsaee).*

Abstract

Clustering datasets with varying densities is challenging because classical Density Peak Clustering (DPC) may fail to identify valid centers in sparse regions and may incorrectly assign sparse samples to dense clusters. To address this issue, this paper proposes a Density Peak Clustering algorithm based on Fuzzy Shared Nearest Neighbors (DPC-FSNN). The proposed method combines k-nearest neighbors and fuzzy neighborhood information to construct a fuzzy nearest neighbor kernel for adaptive local density estimation. It also incorporates shared nearest neighbor information into the local distance computation and sample assignment process to improve center detection and assignment reliability. Experiments on datasets with varying densities (e.g., Jain, Compound, and LineBlobs), complex-shaped datasets, and real-world datasets are evaluated using AMI, ARI, and FMI. The results show that DPC-FSNN achieves competitive and often superior clustering performance compared with IDPC-FA, FKNN-DPC, DPC-FWSN, DPCSA, FNDPC, and the classical DPC algorithm. A computational complexity analysis is also provided, showing that the proposed method has an overall complexity of $O(n^2)$ in the standard implementation.

1. Introduction

Clustering is an important unsupervised learning task in data mining that aims to group similar samples and separate dissimilar ones [1,2]. Among existing approaches, density-based methods are particularly attractive because they can identify clusters with arbitrary shapes.

The Density Peak Clustering (DPC) algorithm proposed by Rodriguez and Laio [3] has received considerable attention due to its simple principle and effective performance. DPC identifies cluster centers based on local density and relative distance, then assigns the remaining samples by following the nearest point with higher density.

Despite its advantages, the original DPC often struggles on datasets with heterogeneous density distributions. Its density estimation tends to favor dense regions, which may overlook valid centers in sparse clusters. In addition, its assignment strategy may produce unreliable labels for boundary or ambiguous samples. These limitations reduce its effectiveness on multi-density and structurally complex datasets.

To address these issues, several DPC variants have been developed by improving density estimation or assignment strategies. However, many existing methods still have limited adaptability to heterogeneous neighborhood structures and remain sensitive to density imbalances between dense and sparse regions.

Motivated by these challenges, this paper proposes a new clustering method called DPC-FSNN for datasets with heterogeneous density distributions. The proposed method introduces a fuzzy nearest neighbor kernel to obtain a more adaptive local density estimation and incorporates shared nearest neighbor (SNN) information into the assignment stage to improve the treatment of boundary and ambiguous samples. By combining adaptive density modeling with neighborhood-consistent

assignment, DPC-FSNN offers a more balanced handling of dense and sparse regions.

1. The main contributions of this paper are summarized as follows: A fuzzy nearest neighbor kernel is introduced to redefine local density for heterogeneous-density data.
2. An SNN-based assignment strategy is developed to improve the reliability of sample allocation, especially for boundary samples.
3. The proposed DPC-FSNN framework improves the robustness of density peak clustering on datasets with varying densities and complex structures.

Experimental results in Section 4show that DPC-FSNN achieves competitive and often superior clustering performance compared with several .representative DPC-based methods

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed method. Section 4 reports the experimental results and analysis. Section 5 concludes the paper.

2. Related Work

The density peak clustering (DPC) algorithm, introduced by Rodriguez and Laio [3], identifies cluster centers based on two conditions: 1) a center must have higher local density than its neighbors, and 2) centers must be well-separated. DPC computes local density (ρ) and relative distance (δ) for each sample:

$$\rho_i = \sum_{i \neq j} X(d_{ij} - d_c) \quad (1)$$

$$(x) = \begin{cases} 1 & x < 0 \\ 0 & x \geq 0 \end{cases} \quad (2)$$

For large datasets, the Gaussian kernel can be used:

$$\rho_i = \sum_{i \neq j} \exp(-\frac{d_{ij}}{d_c})^2 \quad (3)$$

The relative distance δ_i to the nearest higher-density sample is:

$$\delta_i = \min_{j:\rho_j > \rho_i}(d_{ij}) \quad (4)$$

$$\delta_i = \max_{j}(d_{ij}) \quad (5)$$

Cluster centers are selected using $\gamma_i = \rho_i \times \delta_i$, and remaining points are assigned to the nearest high-density cluster [3,7]. While DPC is efficient and non-iterative, it struggles with parameter sensitivity, label propagation errors, and datasets with varying densities.

Unsupervised learning approaches have also been investigated to improve the representation and analysis of unlabeled data. For example, Kharghanian and Mohammadpoory [19] developed K-means-CRBM for efficient unsupervised feature learning. In the context of density-based clustering, however, the challenges of reliable local density estimation and neighborhood-based sample assignment remain important.

Existing DPC-based methods mainly improve the algorithm from three perspectives: local density estimation, similarity/neighborhood modeling, and sample assignment.

Several studies have enhanced local density estimation to better handle multi-density datasets. Zhou et al. [8], Amagata et al. [9], Li et al. [10], and Wang et al. [11,12] proposed adaptive density estimation, probabilistic modeling, or hybrid clustering frameworks to improve DPC performance. While these methods increase adaptability to varying densities, they still rely heavily on accurate density estimation and reliable sample assignment.

Another research direction improves DPC through fuzzy modeling, weighted similarity, and neighborhood-aware strategies. Representative approaches include fuzzy density estimation [13], weighted distance and similarity measures [6,16], adaptive density refinement [15], hybrid DPC frameworks [17], kernel-based separation [18], trilateral clustering theory [7], self-adaptive fuzzy clustering [5], and methods for imbalanced data [14]. These approaches improve clustering robustness, but most focus on either density estimation or assignment reliability rather than optimizing both simultaneously.

Overall, several important challenges remain. First, many DPC variants still favor dense regions during density estimation, causing cluster centers in sparse regions to be overlooked. Second, boundary and ambiguous samples are often assigned unreliably when neighborhood structures are weak or density distributions are highly uneven.

Third, the joint integration of adaptive fuzzy density estimation and structurally informed neighborhood assignment has not been sufficiently explored within a unified DPC framework. These limitations motivate the proposed DPC-FSNN method, which combines a fuzzy nearest neighbor kernel for adaptive density estimation with shared nearest neighbor information for more reliable sample assignment in heterogeneous-density datasets.

3. Proposed Method

The proposed Density Peak Clustering based on Fuzzy Shared Nearest Neighbors (DPC-FSNN) is designed for clustering datasets with heterogeneous density distributions, where conventional methods often perform poorly. It integrates fuzzy kernel-based density estimation with shared nearest neighbor (SNN) structural analysis to achieve more adaptive local density estimation and more robust clustering in the presence of uneven densities and structural noise. The details of the proposed method are presented in the following subsections.

3.1 Local Density of Fuzzy Kernel Nearest Neighbors

In the standard Density Peak Clustering (DPC) algorithm, local density is estimated using a cutoff distance parameter, d_c. Since the optimal value of $d_c$varies across datasets, an inappropriate choice may bias density estimation, favor dense regions, and overlook cluster centers in sparse areas, reducing clustering performance on heterogeneous density datasets.

To overcome this limitation, we propose a Fuzzy Kernel Nearest Neighbor (FKNN)-based density estimation strategy. By combining the k-nearest neighbors (KNN) mechanism with fuzzy membership modeling, FKNN provides a more adaptive estimation of local density. Instead of relying on the global cutoff distance d_c, the proposed method uses the neighborhood size k, making density estimation less sensitive to dataset characteristics. Furthermore, integrating neighborhood information with fuzzy membership values balances the contribution of samples from both dense and sparse regions.

The neighborhood size k controls the locality of density estimation. A smaller value of k emphasizes local structures and is more sensitive to subtle density variations, whereas a larger value produces smoother density estimates by incorporating a broader neighborhood. In our experiments, k was selected through a grid-search procedure over the range $[1, 20]$, and the value yielding the best clustering performance was adopted for each dataset. This strategy enables the proposed method to adapt to datasets with different density characteristics while avoiding the need for manually specifying a global cutoff distance.

Consequently, DPC-FSNN improves the reliability of cluster center detection and enhances clustering robustness for datasets with heterogeneous density distributions.

Definition 1 (Fuzzy Kernel Nearest Neighbor)

The Fuzzy Kernel Nearest Neighbor (FKNN) function is constructed by combining the k-nearest neighbors concept with fuzzy neighborhood modeling. The FKNN membership of sample x_i with respect to sample x_j is defined as:

$$\mu(i,j)=\begin{cases}\dfrac{e^{-d_{ij}^2}}{(d_{ij}+1)}^2 & if\ j\in kNN(i)\\ \dfrac{e^{-(\varphi d_{ij})^2}}{(d_{ij}+1)^2} & others\end{cases} \tag{6}$$

where d_{ij} denotes the Euclidean distance between samples x_i and x_j. The fuzzy membership is computed differently according to the neighborhood relationship between the two samples. If x_j belongs to the k-nearest neighbors of x_i, the membership depends only on their pairwise distance, thereby preserving local neighborhood information. Otherwise, the scattering coefficient φ is incorporated to attenuate the contribution of distant samples according to the overall data dispersion. Consequently, nearby samples receive larger membership values, while distant samples contribute progressively less to the local density estimation.

Here, $kNN(i)$ denotes the set of the k-nearest neighbors of x_i. The fuzzy membership $\mu(i,j)$ varies depending on whether x_j belongs to the nearest-neighbor set of x_i. For nearest neighbors, the membership is primarily determined by the pairwise distance d_{ij}, which reflects local similarity. For non-nearest neighbors, the scattering coefficient φ is additionally incorporated to reflect the global dispersion of the dataset. The scattering coefficient φ is computed based on the average distance ω as follows:

$$\omega=\frac{1}{N.(N-1)}\sum_{i=1}^{N}\sum_{j=1}^{N}d_{ij} \tag{7}$$

$$\varphi=\sqrt{\frac{1}{N-1}\sum_{i\neq j}^{N}(d_{ij}-\omega)^2} \tag{8}$$

The scattering coefficient φ measures the global variability of pairwise distances in the dataset. Larger values of φ indicate a higher degree of distance dispersion and consequently produce stronger attenuation for non-neighbor samples in Eq. (6). Conversely, when the dataset exhibits relatively uniform distances, φ becomes smaller, allowing non-neighbor samples to retain a moderate influence on density estimation. Since φ is computed directly from the pairwise distance matrix using Eqs. (7) and (8), no additional parameter tuning is required.

Using these memberships, the local density ρ_i of sample x_i is defined as

$$\rho_i = \frac{1}{k}\sum_{j=1}^{k}\mu(i,j) + \frac{1}{N-k}\sum_{v=k+1}^{N}\mu(i,v) \quad (9)$$

The first term represents the contribution of the k-nearest neighbors, which usually share similar local density characteristics with x_i. The second term represents the contribution of the remaining samples.

The weighting scheme in Eq. (9) allows the local density to simultaneously exploit local neighborhood information and global structural information. The first term emphasizes nearby samples through larger fuzzy memberships, whereas the second term introduces a weak contribution from the remaining samples that is automatically regulated by the scattering coefficient. As a result, the proposed density estimation becomes less biased toward dense regions and provides a more balanced description of heterogeneous-density datasets.

Consequently, the FKNN strategy improves the possibility of identifying cluster centers from both dense and sparse clusters.

Definition 2 (Shared Nearest Neighbors):

Let x_i and x_j be two samples in the dataset, and let $\Gamma(i)$ and $\Gamma(j)$ denote their respective sets of k-nearest neighbors. The shared nearest neighbors (SNN) of x_i and x_j are defined as:

$$SNN(i,j) = \Gamma(i) \cap \Gamma(j) \quad (10)$$

This measure reflects the degree of neighborhood overlap between the two samples. A larger overlap indicates a higher structural similarity and a greater likelihood that the two samples belong to the same cluster. Because SNN captures local structural consistency, it is effective in handling clusters with different densities and irregular shapes.

Definition 3. (Local Distance):

For a sample x_i, the local distance is computed with respect to samples having higher local density. Let $KNN(i)$ and $KNN(j)$ denote the sets of k-nearest neighbors of $x_i$and x_j, respectively. The choice of k plays an important role in defining the neighborhood structure. A larger k generally produces smoother and more stable neighborhoods and improves noise tolerance, whereas a smaller k is better suited for capturing compact local structures but may be more sensitive to outliers. Therefore, the selection of k directly affects density estimation, local distance computation, and neighborhood consistency in multi-density datasets. In this study, the neighborhood size k was determined using the same parameter selection strategy described in Section 4.1. Specifically, k was searched within the interval $[1, 20]$, and the value yielding the best clustering performance was selected for each dataset. This adaptive selection allows the local distance measure to remain effective across datasets with different density distributions and structural characteristics. The local distance δ_i is defined as:

$$\delta_i = \min_{j:\rho_j>\rho_i}\left[d_{ij}\Big(\sum_{p\in kNN(i)} d_{ip} + \sum_{q\in kNN(j)} d_{iq}\Big)\right] \quad (11)$$

If x_i has the highest local density, its local distance is defined as:

$$\delta_i = \max_i(d_{ij}) \quad (12)$$

This formulation incorporates both relative distance and neighborhood structure. When samples within a cluster are more scattered and have larger neighborhood distances, the proposed local distance can compensate for sparsity and reduce the risk of underestimating the importance of such samples.

Definition 4. Fuzzy Hybrid Score (FHS)

Similar to the original DPC algorithm, the values of ρ and δ are used to identify cluster centers. However, instead of directly using their product, the proposed method introduces a Fuzzy Hybrid Score (FHS) that adaptively balances density and distance through fuzzy weighting. The score is defined as:

$$FHS_i = [\alpha_i(\frac{\rho_i}{\max(\rho)}) \times (1-\alpha_i)(\frac{\delta_i}{\max(\delta)})]^2 \quad (13)$$

The adaptive fuzzy weight α_i is calculated as:

$$\alpha_i = 1 - \frac{\rho_i - \rho_{\min}}{\rho_{\max} - \rho_{\min}} \quad (14)$$

In this formulation, samples in dense regions receive lower fuzziness, allowing density to play a stronger role in the score. In contrast, samples in sparse regions receive higher fuzziness, which increases the influence of the distance term. This mechanism improves the robustness of cluster center detection in datasets with varying density distributions. After calculating the FHS values, all samples are sorted in descending order, and the top Ksamples are selected as cluster centers.

For the assignment stage, the remaining samples are allocated by using shared nearest neighbor information between unassigned samples and already assigned samples. This strategy alleviates the limitation of one-pass propagation in the original DPC algorithm. Let x_i be an unassigned sample. It is assigned to a cluster only if the following condition is satisfied:

$$|SNN(i,j)| \geq \frac{k}{2} \quad (15)$$

The parameter $k$is important in this stage because each sample has exactly $k$nearest neighbors. Therefore, the threshold $\frac{k}{2}$ acts as a majority consistency rule: an unassigned sample is linked to an assigned sample only when at least half of their nearest neighbors overlap. This condition ensures that cluster assignment is performed only under sufficiently strong local structural similarity, thereby reducing incorrect assignment of noisy or weakly related samples. After completing these steps, the clustering process is finalized for datasets with varying density distributions. The complete DPC-FSNN procedure is summarized in Algorithm 1.

Algorithm 1: DPC-FSNN algorithm.

Input : Dataset $U = \{x_1, x_2, \cdots, x_n\}$, neighborhood size k.
Output : Clustering result $\{C_1, C_2, \ldots, C_m\}$

1. Compute the pairwise distance matrix D = d_{ij} for all samples in *U*.
2. Construct the k-nearest neighbor set $kNN(i)$ for each sample x_i.
3. For *i* = 1: *n* do
4. Compute the fuzzy memberships $\mu(i,j)$ according to Eq. (6).
5. Compute the local density ρ_i according to Eq. (9).
6. End For
7. Compute the shared nearest neighbor matrix $SNN(i,j)$ for all sample pairs according to Eq. (10).
8. For *i* = 1: *n* do
9. Compute the local distance δ_i according to Eqs. (11) and (12).
10. Compute the Fuzzy Hybrid Score FHS_i according to Eq. (13).
11. End For
12. Sort all samples in descending order of FHS_i.
13. Select the top m samples as cluster centers $c_1, c_2, \ldots, c_m$.
14. Assign each cluster center c_t to cluster C_t, for $t = 1\ to\ m$.
15. Initialize a queue Q and insert all cluster centers into Q.
16. While Q is not empty do
17. Remove the first sample x_j from Q.
18. Find the unassigned samples that are neighbors of x_j.
19. For each unassigned neighbor x_i do
20. If $|SNN(i,j)| \geq k/2$ then
21. Assign xi to the same cluster as x_j.
22. Insert xi into the tail of Q.
23. End If
24. End For
25. End While
26. Return $C = c_1, c_2, \ldots, c_m\}$.

3.2. Computational Complexity Analysis

Let $n$denote the number of samples. In DPC-FSNN, the main computational cost comes from pairwise distance calculation, fuzzy neighborhood construction, local density estimation, and shared nearest neighbor-based assignment. Computing the pairwise distance matrix requires $O(n^2)$. Based on this matrix, the construction of the k-nearest neighbor sets for all samples requires $O(n^2)$ in the standard implementation. The computation of fuzzy memberships and local densities also requires $O(n^2)$. Similarly, the calculation of the relative distance δ_i for all samples has worst-case complexity $O(n^2)$. The ranking of the FHS values requires $O(n\log n)$, and the final assignment step based on shared nearest neighbors requires at most $O(n^2)$.

Therefore, the overall computational complexity of DPC-FSNN is dominated by the pairwise distance computation and neighborhood analysis, resulting in an overall complexity of $O(n^2)$in the standard implementation. Although the proposed method has a computational complexity comparable to many DPC-based algorithms, it provides improved robustness for datasets with heterogeneous density distributions.

To complement the theoretical complexity analysis, an empirical runtime evaluation has been conducted on benchmark datasets with different sample sizes and structural characteristics. The runtime comparison, presented in Section 4.5 (Table 7), demonstrates that the practical computational cost of DPC-FSNN is competitive with existing DPC-based methods while achieving improved clustering performance. These experimental results are consistent with the theoretical $O(n^2)$ complexity derived in this section.

4. Experimental Results and Analysis

In this section, we describe the evaluation metrics used to assess clustering performance. We compare the proposed DPC-FSNN algorithm with several existing methods using various benchmark datasets. Table 1 summarizes the characteristics of the datasets employed. Following this, we outline the experimental setup, discuss the comparison metrics, and analyze the clustering accuracy for each method.

Table 1. Features of datasets with different densities.

Dataset	Number of Data Points	Number of Features	Number of Clusters
Jain	373	2	2
Compound	399	2	6
LineBlobs	266	2	3

4.1 Experimental Settings

To evaluate the proposed method, experiments were conducted on three varying-density synthetic datasets, six complex-shape

synthetic datasets, and six real-world datasets. DPC-FSNN was compared with six state-of-the-art clustering algorithms: IDPC-FA [20], FKNNDPC [21], DPC-FWSN [16], DPCSA [22], FNDPC [23], and the original DPC [3].

All algorithms were implemented under identical experimental conditions using the same initialization procedures and evaluation metrics.

DPCSA and IDPC-FA require no parameter tuning. For DPC-FWSN and FKNN-DPC, the parameter k was varied from 1 to 50 with a step size of 1, whereas for DPC-FSNN, k was searched from 1 to 20. For FNDPC, the parameter β ranged from 0.01 to 1 with a step size of 0.01, while for DPC, the cutoff distance varied from 0.1% to 5% with a step size of 0.1%.

To ensure a fair comparison, all tunable parameters were selected using the same exhaustive search strategy within the ranges recommended in the corresponding original publications. For each algorithm, the parameter combination yielding the best clustering performance on each dataset was adopted for evaluation. Algorithms such as DPCSA and IDPC-FA were executed using their default configurations because they do not require parameter tuning. Therefore, no algorithm was given a preferential tuning strategy, and each competing method was evaluated under its most favorable parameter setting according to its original design.

It should be noted that the search ranges differ across algorithms because they are intrinsic to the corresponding methods rather than being arbitrarily selected. Specifically, DPC-FSNN employs a neighborhood size within the interval$[1, 20]$, whereas DPC-FWSN and FKNN-DPC follow the broader search ranges suggested in their original studies. This protocol ensures that every algorithm is evaluated within its recommended operating conditions while maintaining a consistent and unbiased parameter selection procedure.

Clustering performance was evaluated using the Adjusted Mutual Information (AMI) [24], Adjusted Rand Index (ARI) [25], and Fowlkes–Mallows Index (FMI) [26], where higher values indicate better clustering performance.

As summarized in Tables 1, 3, and 5, the benchmark datasets cover a broad range of sample sizes, feature dimensions, and cluster structures. Specifically, the number of samples ranges from 150 (Iris) to 5,000 (S1), while the feature dimensionality varies from 2 to 90. Evaluating DPC-FSNN on datasets with such diverse characteristics enables a comprehensive assessment of its robustness with respect to dataset scale, dimensionality, and density variation. The reported AMI, ARI, and FMI values therefore reflect the performance of the proposed method under substantially different data complexities.

4.2 Results on Datasets with Varying Densities

Table 2 summarizes the performance of the seven clustering algorithms on datasets with varying densities. Overall, DPC-FSNN achieves the best or among the best results, demonstrating its effectiveness in handling heterogeneous density distributions.

Table 2. Performance of seven clustering algorithms on datasets with different densities.

Algorithm	FMI	ARI	AMI	FMI	ARI	AMI	FMI	ARI	AMI
	Compound			Jain			LineBlobs		
DPC-FSNN	**0.9302**	**0.8871**	**0.8951**	1	1	1	1	1	1
DPC-FWSN	0.9115	0.878	0.8664	1	1	1	1	1	1
IDPC-FA	0.8815	0.8327	0.7922	1	1	1	0.8842	0.8237	0.8375
FKNN-DPC	0.8877	0.8418	0.8381	0.9359	0.8224	0.7092	1	1	1
DPCSA	0.8707	0.8284	0.8392	0.5924	0.0442	0.2167	1	1	1
FNDPC	0.644	0.5337	0.7239	0.9051	0.7257	0.5961	0.7218	0.5769	0.6386
DPC	0.6876	0.591	0.7754	0.8819	0.7146	0.6183	0.8166	0.721	0.7799

On the Compound dataset, the original DPC achieves relatively low performance (AMI = 0.7754, ARI = 0.5910, FMI = 0.6876). Although IDPC-FA, DPCSA, and FNDPC improve these results, DPC-FSNN attains the highest scores (AMI = 0.8951, ARI = 0.8871, FMI = 0.9302), indicating that the proposed fuzzy density estimation better preserves sparse cluster structures while reducing bias toward dense regions during cluster center selection.

On the LineBlobs dataset, DPC-FSNN also achieves perfect AMI, ARI, and FMI scores. Its performance is comparable to DPC-FWSN and FKNN-DPC and is superior to the original DPC and FNDPC.

Overall, the results demonstrate that DPC-FSNN provides accurate and stable clustering across datasets with varying densities.

The combination of fuzzy nearest neighbor-based density estimation and SNN-based assignment enhances the detection of sparse cluster centers, preserves local structural consistency, and reduces the misclassification of boundary samples. These findings are further supported by Figures 1 and 2.

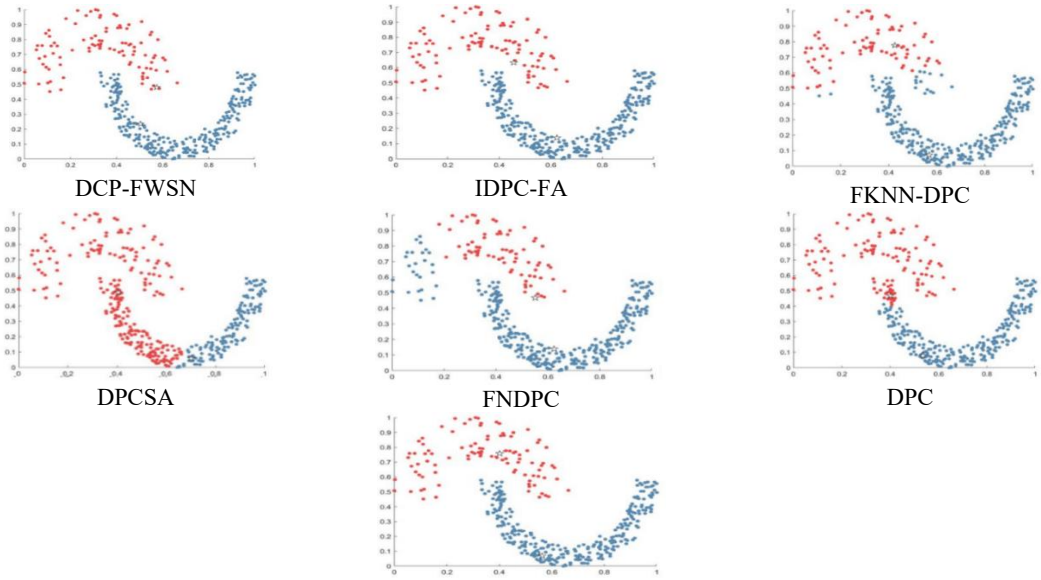

Figure 1. Clustering Performances on the Jain Dataset.

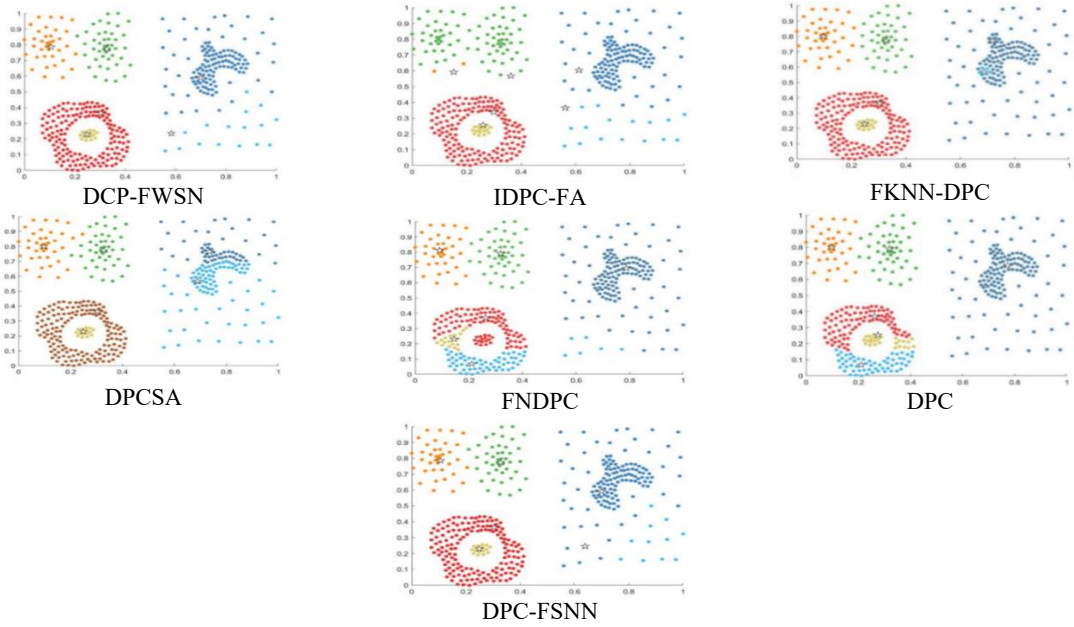

Figure 2. Clustering Performances on the Compound Dataset.

4.3. Experimental Results for Complex Morphology Datasets

Six challenging synthetic datasets (Flame, Aggregation, S1, Spiral, R15, and D31) were used to evaluate clustering performance on irregular shapes, multiple clusters, and complex spatial distributions.

Table 3 reports the AMI, ARI, and FMI results of the seven clustering algorithms.

Overall, DPC-FSNN achieves highly competitive performance across most datasets, demonstrating its effectiveness on complex morphological structures.

Table 3. Clustering Performances of Seven Algorithms on Complex Morphological Datasets.

Algorithm	FMI	ARI	AMI	FMI	ARI	AMI	FMI	ARI	AMI
	Aggregation			Flame			D31		
DPC-FSNN	0.9809	0.9915	0.9906	1	1	1	**0.9676**	**0.9689**	**0.9802**
DPC-FWSN	**0.9977**	**0.9971**	**0.995**	1	1	1	0.9496	0.9479	0.9637
IDPC-FA	1	1	1	1	1	1	0.9421	0.9402	0.9575
FKNN-DPC	0.996	0.9949	0.9905	0.9845	0.9666	0.9267	0.9482	0.9465	0.9621
DPCSA	0.9673	0.9581	0.9537	1	1	1	0.9374	0.9353	0.9552
FNDPC	0.9932	0.9913	0.9864	1	1	1	0.9385	0.9364	0.9555
DPC	0.9966	0.9956	0.9922	1	1	1	0.9385	0.9365	0.9554
	Spiral			S1			R15		
DPC-FSNN	1	1	1	**0.9904**	**0.9897**	**0.9895**	**0.9933**	**0.9928**	**0.9938**
DPC-FWSN	1	1	1	0.9904	0.9897	0.9895	0.9933	0.9928	0.9938
IDPC-FA	1	1	1	0.9904	0.9897	0.9895	0.9933	0.9928	0.9938
FKNN-DPC	1	1	1	0.99	0.9893	0.9892	0.9933	0.9928	0.9938
DPCSA	1	1	1	0.9888	0.988	0.9879	0.9866	0.9857	0.9885
FNDPC	1	1	1	0.9896	0.9889	0.9887	0.9933	0.9928	0.9938
DPC	1	1	1	0.9904	0.9897	0.9895	0.9932	0.9928	0.9938

Table 4. Characteristics of datasets with different densities.

Dataset	Number of Data	Number of Features	Number of Clusters
Flame	240	2	2
Aggregation	788	2	7
S1	5000	2	5
Spiral	312	2	3
R15	600	2	5
D31	3100	2	31

For the S1 dataset, it obtains AMI = 0.9895, ARI = 0.9897, and FMI = 0.9904, performing comparably to DPC-FWSN, IDPC-FA, and the original DPC. On the R15 dataset, DPC-FSNN also ranks among the best methods (AMI = 0.9938, ARI = 0.9928, FMI = 0.9933), outperforming DPCSA and the original DPC.

For the Flame dataset, DPC-FSNN achieves perfect scores, while on the Aggregation dataset it remains highly competitive with all evaluation metrics above 0.98, although IDPC-FA achieves perfect performance. All methods obtain perfect results on the Spiral dataset. Overall, these results demonstrate that DPC-FSNN consistently maintains accurate cluster center detection and reliable sample assignment across datasets with irregular shapes and complex cluster structures.

4.4. Results on Real-World Datasets

To further evaluate the proposed method, experiments were conducted on six real-world datasets from the UCI repository with different sample sizes, feature dimensions, cluster numbers, and structural complexities.

Table 5 summarizes the dataset characteristics, while Table 6 reports the performance of the seven clustering algorithms. Overall, DPC-FSNN achieves competitive and often superior results across these datasets.

On the Iris dataset, DPC-FSNN, DPC-FWSN, FKNN-DPC, DPCSA, and FNDPC achieve similarly strong performance, whereas IDPC-FA and the original DPC perform slightly worse.

Table 5. Characteristics of real-world datasets.

Dataset	Number of Data	Number of Features	Number of Clusters
Iris	150	4	3
Wine	178	13	3
Seeds	210	7	3
Ecoli	336	8	8
Ionosphere	351	34	2
Libras	360	90	15

Table 6. Clustering Performances of Seven Algorithms on Real-World Datasets.

Algorithm	FMI	ARI	AMI	FMI	ARI	AMI	FMI	ARI	AMI
	Wine			Iris			Libras		
DPC-FSNN	0.9298	0.9199	0.8809	0.9355	0.9038	0.8831	0.4429	0.3956	0.5762
DPC-FWSN	0.9319	0.8975	0.8716	0.9355	0.9038	0.8831	0.3969	0.3325	0.5408
IDPC-FA	0.8478	0.7713	0.7675	0.9233	0.8857	0.8623	0.4247	0.3816	0.5733
FKNN-DPC	0.9229	0.8839	0.8481	0.9355	0.9038	0.8831	0.4044	0.3459	0.5554
DPCSA	0.8283	0.7414	0.748	0.9355	0.9038	0.8831	0.3791	0.3095	0.5388
FNDPC	0.8686	0.8025	0.7898	0.9355	0.9038	0.8831	0.3869	0.329	0.5494
DPC	0.7835	0.6724	0.7065	0.8032	0.7037	0.7247	0.3717	0.3193	0.5358
	Ecoli			Seeds			Ionosphere		
DPC-FSNN	0.8309	0.7656	0.6725	0.8912	0.7856	0.8802	0.8023	0.5012	0.3706
DPC-FWSN	0.7585	0.7796	0.8532	0.6637	0.7331	0.8059	0.7842	0.4746	0.3664
IDPC-FA	0.7299	0.767	0.8444	0.6638	0.7561	0.8284	0.6432	0.2183	0.1355
FKNN-DPC	0.7118	0.7303	0.8029	0.5878	0.5894	0.7027	0.7716	0.479	0.3485
DPCSA	0.6609	0.6873	0.7918	0.4406	0.4593	0.6467	0.639	0.2135	0.1335
FNDPC	0.7136	0.7545	0.8361	0.4833	0.5618	0.7178	0.6513	0.2483	0.163
DPC	0.7298	0.767	0.8444	0.4978	0.4465	0.5775	0.6476	0.2231	0.1376

For the Ecoli dataset, DPC-FSNN achieves the best performance (AMI = 0.6725, ARI = 0.7656, FMI = 0.8309). It also achieves the best results on the Libras (AMI = 0.5762, ARI = 0.3956, FMI = 0.4429) and Seeds (AMI = 0.8802, ARI = 0.7856, FMI = 0.8912) datasets.

On the Wine dataset, DPC-FSNN outperforms DPC-FWSN, FKNN-DPC, DPCSA, FNDPC, and the original DPC, achieving AMI = 0.8809, ARI = 0.9199, and FMI = 0.9298. For the Ionosphere dataset, it obtains the highest ARI and FMI values while remaining competitive in AMI, demonstrating its effectiveness on higher-dimensional data.

Overall, the results confirm that DPC-FSNN provides accurate and stable clustering across diverse real-world datasets. Its adaptive fuzzy density estimation and shared nearest neighbor-based assignment improve both cluster center detection and sample assignment, enabling robust performance on datasets with varying sizes, densities, and cluster structures.

The experimental results indicate that DPC-FSNN maintains competitive performance across datasets with substantially different sample sizes and feature dimensions.

For example, the proposed method achieves strong results not only on low-dimensional synthetic datasets but also on higher-dimensional real-world datasets such as Libras (90 features) and Ionosphere (34 features), demonstrating that its effectiveness is not restricted to a particular data scale or dimensionality.

4.5. Runtime Analysis

To complement the clustering-quality evaluation, we further compared the runtime of DPC-FSNN with the baseline algorithms on several benchmark datasets.

Table 7 reports the execution time in seconds for datasets with different sizes and structural characteristics. Overall, DPC-FSNN shows competitive computational efficiency and achieves lower runtime than several representative methods, including FKNN-DPC, DPCSA, and FNDPC, on most datasets. Although the classical DPC is faster on some small datasets because of its simpler formulation, the proposed method provides a favorable balance between clustering effectiveness and computational cost.

These empirical observations are consistent with the theoretical complexity analysis in Section 3.2, which indicates an overall complexity of $O(n^2)$ for the standard implementation of DPC-FSNN.

Although classical DPC is faster on some relatively small datasets because of its simpler formulation, the proposed method provides a favorable balance between clustering effectiveness and computational cost.

Overall, the runtime comparison indicates that the additional computations introduced by the FKNN-based density estimation and SNN-based assignment do not impose a significant computational burden in practice.

Despite having the same theoretical time complexity as many existing DPC-based algorithms, DPC-FSNN achieves competitive execution times while consistently providing

superior or comparable clustering performance across the benchmark datasets. These empirical observations further validate the practical applicability of the proposed method.

Table 7. Runtime comparison (seconds) of DPC-FSNN and baseline clustering methods on benchmark datasets.

Dataset	n	DPC-FWSN	DPC	IDPC-FA	FKNN-DPC	DPCSA	FNDPC	DPC-FSNN
aggregation	788	0.1627	0.0551	0.3424	0.753	0.8904	0.3438	0.0419
jain	373	0.0263	0.0148	0.0729	0.3302	0.5615	0.4115	0.0157
spiral	312	0.0525	0.0108	0.0538	0.2013	0.3301	0.6222	0.0464
flame	240	0.025	0.0073	0.0236	0.2056	0.2557	0.7562	0.0926
R15	600	0.0578	0.0543	0.2123	0.2359	0.6435	0.7336	0.4766
wine	178	0.0412	0.0049	0.0263	0.066	0.203	0.6507	0.0126
iris	150	0.0369	0.0041	0.0139	0.0703	0.2312	0.2965	0.0612

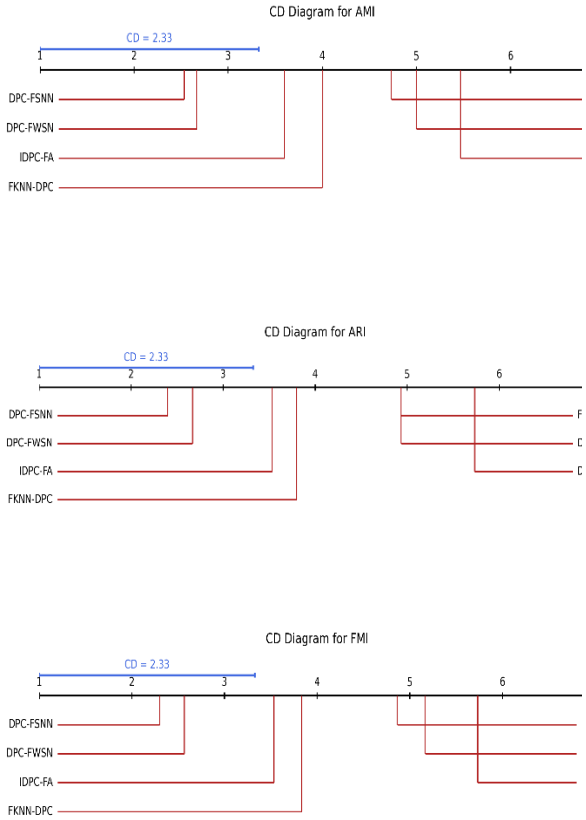

Figure 3. Comparison of the seven clustering algorithms using the Friedman-Nemenyi test.

4.6. Statistical Significance Analysis

To further evaluate the reliability of the comparative results, we performed the Friedman test followed by the Nemenyi post-hoc analysis using the obtained AMI, ARI, and FMI scores on the benchmark datasets. The corresponding critical difference (CD) diagrams are presented in Figure 3. As shown in Figure 3, DPC-FSNN consistently achieves the best or among the best average ranks across the three evaluation metrics. This finding indicates that the strong performance of the proposed method is not limited to a few isolated cases but is maintained across benchmark datasets with varying densities and complex structures. The statistical comparison, therefore, provides further evidence for the robustness and generalizability of DPC-FSNN.

5. Conclusion

This paper presented DPC-FSNN, a density-based clustering method for datasets with varying density distributions. By combining a fuzzy nearest neighbor kernel for adaptive local density estimation with shared nearest neighbor information for sample assignment, the proposed method improves the identification of cluster centers in sparse regions and reduces misclassification of low-density samples into dense clusters. Experimental results on datasets with varying densities, complex morphological datasets, and real-world datasets demonstrate that DPC-FSNN achieves competitive and often superior performance compared with representative baseline methods. These findings confirm the effectiveness and robustness of the proposed method for clustering data with heterogeneous densities and complex structures. Although the proposed method shows strong performance, its standard implementation still has quadratic computational complexity. Future work will focus on improving its scalability and extending it to larger and higher-dimensional datasets.

References

[1] M. Ghazizadeh-Ahsaee, H. Sadoghi-Yazdi, and M. Naghibzadeh, "Curve fitting space for classification, "

Neural Computing and Applications, vol. 20, no. 2, pp. 273–285, 2011.

[2] K. J. Cios, W. Pedrycz, R. W. Swiniarski, and L. A. Kurgan, *Data Mining: A Knowledge Discovery Approach*. New York, NY, USA: Springer, 2007.

[3] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks, " *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[4] A. Lotfi, P. Moradi, and H. Beigy, "Density peaks clustering based on density backbone and fuzzy neighborhood, " *Pattern Recognition*, vol. 107, Art. no. 107449, 2020.

[5] K. Qiao, J. Chen, and S. Duan, "Self-adaptive two-stage density clustering method with fuzzy connectivity, " *Applied Soft Computing*, vol. 154, Art. no. 111355, 2024.

[6] J. Xie, X. Liu, and M. Wang, "SFKNN-DPC: Standard deviation weighted distance-based density peak clustering algorithm, " *Information Sciences*, vol. *653*, Art. no. 119788, 2024.

[7] H. Ju, Y. Lu, W. Ding, J. Cao, and X. Yang, "Three-way evidence theory-based density peak clustering with the principle of justifiable granularity, " *Applied Soft Computing*, vol. 152, Art. no. 111217, 2024.

[8] X. Yuan, H. Yu, J. Liang, and B. Xu, "A novel density peaks clustering algorithm based on K nearest neighbors with adaptive merging strategy, " *International Journal of Machine Learning and Cybernetics*, vol. 12, pp. 2825–2841, 2021.

[9] D. Amagata and T. Hara, "Efficient density-peaks clustering algorithms on static and dynamic data in Euclidean space, " *ACM Transactions on Knowledge Discovery from Data*, vol. 18, no. 1, Art. no. 2, pp. 1–27, 2023.

[10] M. Li, X. Bi, L. Wang, and X. Han, "A method of two-stage clustering learning based on improved DBSCAN and density peak algorithm, " *Computer Communications*, vol. 167, pp. 75–84, 2021.

[11] Y. Wang, D. Wang, Y. Zhou, X. Zhang, and C. Quek, "VDPC: Variational density peak clustering algorithm, " *Information Sciences*, vol. 621, pp. 627–651, 2022.

[12] Y. Wang, W. Pang, and J. Zhou, "An improved density peak clustering algorithm guided by pseudo labels," *Knowledge-Based Systems*, vol. 252, Art. no. 109374, 2022.

[13] J. Zhao, G. Wang, J.-S. Pan, T. Fan, and I. Lee, "Density peaks clustering algorithm based on fuzzy and weighted shared neighbor for uneven density datasets, " *Pattern Recognition*, vol. 139, Art. no. 109406, 2023.

[14] F. Ge and X. Liu, "Density peaks clustering algorithm based on a divergence distance and tissue-like P system, " *Applied Sciences*, vol. 13, no. 4, Art. no. 2293, 2023.

[15] P. Wang, T. Wu, and Y. Yao, "A three-way adaptive density peak clustering (3W-ADPC) method, " *Applied Intelligence*, vol. 53, no. 20, pp. 23966–23982, 2023.

[16] S. Ding, W. Du, C. Li, X. Xu, L. Wang, and L. Ding, "Density peaks clustering algorithm based on improved similarity and allocation strategy, " *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 4, pp. 1527–1542, 2023.

[17] L. Guo, W. Qin, Z. Cai, and X. Su, "Hybrid clustering algorithm based on improved density peak clustering, " *Applied Sciences*, vol. 14, no. 2, Art. no. 715, 2024.

[18] S. Zhang and K. Li, "A novel density peaks clustering algorithm with isolation kernel and K-induction, " *Applied Sciences*, vol. 13, no. 1, Art. no. 322, 2023.

[19] R. Kharghanian and Z. Mohammadpoory, "K-means-CRBM: An efficient unsupervised tool for feature learning, " *Journal of Advanced Data Mining and Applications*, vol. 14, no. 1, pp. 61–69, Jan. 2026

[20] J. Zhao, J. Tang, A. Shi, T. Fan, and L. Xu, "Improved density peaks clustering based on firefly algorithm," *International Journal of Bio-Inspired Computation*, vol. 15, no. 1, pp. 24–42, 2020.

[21] J. Xie, H. Gao, W. Xie, X. Liu, and P. W. Grant, "Robust clustering by detecting density peaks and assigning points based on fuzzy weighted K-nearest neighbors, " *Information Sciences*, vol. 354, pp. 19–40, 2016.

[22] D. Yu, G. Liu, M. Guo, X. Liu, and S. Yao, "Density peaks clustering based on weighted local density sequence and nearest neighbor assignment, " *IEEE Access*, vol. *7*, pp. 34301–34317, 2019.

[23] M. Du, S. Ding, and Y. Xue, "A robust density peaks clustering algorithm using fuzzy neighborhood, " *International Journal of Machine Learning and Cybernetics*, vol. 9, no. 7, pp. 1131–1140, 2018.

[24] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance, " *Journal of Machine Learning Research*, vol. 11, pp. 2837–2854, 2010.

[25] L. Bai, X. Cheng, J. Liang, H. Shen, and Y. Guo, "Fast density clustering strategies based on the k-means algorithm, " *Pattern Recognition*, vol. 71, pp. 375–386, 2017.

[26] R. Liu, H. Wang, and X. Yu, "Shared-nearest-neighbor-based clustering by fast search and find of density peaks, " *Information Sciences*, vol. 450, pp. 200–226, 2018.

الگوریتم خوشه‌بندی قله چگالی مبتنی بر همسایگان مشترک نزدیک فازی تطبیقی (DPC-FSNN)

افسانه شمس‌الدینی-فرسنگی و مصطفی قاضی‌زاده-احسائی*

گروه مهندسی کامپیوتر، دانشگاه شهید باهنر کرمان، کرمان، ایران.

ارسال ۲۰۲۵/۱۱/۱۷؛ بازنگری ۲۰۲۶/۰۷/۰۵؛ پذیرش ۲۰۲۶/۰۷/۳۰

چکیده:

خوشه‌بندی مجموعه‌داده‌های دارای چگالی‌های متفاوت همواره یکی از چالش‌های مهم در یادگیری بدون ناظر محسوب می‌شود، زیرا الگوریتم کلاسیک خوشه‌بندی قله چگالی (DPC) ممکن است در شناسایی صحیح مراکز خوشه در نواحی کم‌تراکم با مشکل مواجه شود و نمونه‌های کم‌تراکم را به‌اشتباه به خوشه‌های متراکم تخصیص دهد. در این مقاله، الگوریتمی جدید با عنوان DPC-FSNN مبتنی بر همسایگان مشترک نزدیک فازی پیشنهاد شده است. در روش پیشنهادی، اطلاعات همسایگان نزدیک و همسایگی فازی برای ساخت یک هسته همسایگان نزدیک فازی به‌منظور برآورد تطبیقی چگالی محلی ترکیب شده است. همچنین، اطلاعات همسایگان مشترک نزدیک در محاسبه فاصله محلی و فرایند تخصیص نمونه‌ها به کار گرفته شده تا دقت شناسایی مراکز خوشه و قابلیت اطمینان تخصیص نمونه‌ها افزایش یابد. عملکرد روش پیشنهادی بر روی مجموعه‌داده‌های با چگالی‌های متفاوت، مجموعه‌داده‌های با ساختارهای پیچیده و مجموعه‌داده‌های واقعی با استفاده از معیارهای AMI، ARI و FMI ارزیابی شد. نتایج تجربی نشان می‌دهد که روش DPC-FSNN در مقایسه با الگوریتم‌های IDPC-FA ، FKNN-DPC، DPC-FWSN، DPCSA، FNDPC و الگوریتم کلاسیک DPC ، عملکردی رقابتی و در بسیاری از موارد برتر ارائه می‌دهد. همچنین تحلیل پیچیدگی زمانی نشان می‌دهد که پیچیدگی محاسباتی روش پیشنهادی در پیاده‌سازی استاندارد برابر با $O(n^2)$ است.

کلمات کلیدی: خوشه‌بندی، قله چگالی، همسایگان مشترک نزدیک، همسایگی فازی، داده‌های با چگالی متفاوت، برآورد چگالی محلی.