Weighted Density for The Win: Accurate Subspace Density Clustering

APPENDIX

Due to space limitations, many experimental details could not be shown in detail in the main text. The following is what the appendix contains: Appendix A section details the main notations used in the text to help the reader better understand the theoretical foundations of the algorithm. Appendix B section provides detailed statistical information on the 12 datasets used in the experiments. Appendix C section details the parameter settings of the comparative clustering algorithm, which are crucial for the performance of the algorithm. Appendix D section demonstrates the clustering performance of the WDSC algorithm on different datasets, including comparisons with seven other algorithms. Appendix E section verifies the statistical significance of the WDSC algorithm in comparison to the other algorithms through the Wilcoxon signed rank test. Finally, Appendix F section evaluates the effectiveness of the WDSC core components through ablation studies. Therefore, this appendix aims to provide these detailed information additionally to provide readers with a more complete perspective to evaluate the effectiveness and applicability of our WDSC.

A. Details of the Equations

Table I presents the major symbols and notations of the equations used in the paper.

B. Details of the Datasets

The statistics of the 12 experimental datasets are shown in Table II. [1]–[4] “Sonar”, “Wine”, “Iris”, “Heart”, “Pima” and Leaf datasets are collected from the UCI machine learning repository. Datasets “Pathbased”, “Aggregation”, “Jain” and “Spiral” are from University of Eastern Finland, “Moon” and “CMC” datasets are collected from GitHub.

C. Parameters Settings of Algorithms

Some parameters need to be pre-specified for these compared clustering algorithms. For WKM approach, the parameter β for updating the attribute weights is set at 2 [5]. The number of clusters k is set according to the labels of the datasets. DBSCAN has two important parameters, the maximum radius Eps and the minimum points MinPts. d_c is the cutoff distance of DPC and its variants, which is set to be a distance at which 2%–3% of points are included on average [6]. K for SNN-DPC and DenMune represents the number of nearest neighbors. Similarly, this applies to the WDSC algorithm. T_r of DPC-CE controls the sensitivity of the connectivity estimate, while P_r adjusts the distance penalty to balance connectivity with spatial distance. δ_t of VDPC is used to select representatives. The parameter k in ICKDP controls the selection of local centers to optimize clustering performance, and K refers to the number of K-nearest neighbors in the dataset. The parameters configurations of regarding compared algorithms and proposed methods are listed in Table III.

TABLE I
FREQUENTLY USED NOTATIONS.

Symbol	Meaning
k	The number of neighbors considered
m	The number of clusters
W	The weights matrix
$D = \{d_{ij}\}$	The distances of the pairs of data points in X
$SNN(i) = (x_1, \dots)$	The set of shared nearest neighbors of point i
$S_w(ij)$	The weighted SNN similarity between points i and j
$P = (P_1, \dots, P_n)$	The weighted local density
$\Delta = (\Delta_1, \dots, \Delta_n)$	The weighted distance from a larger density point
$\Gamma = (\Gamma_1, \dots, \Gamma_n)$	The weighted decision value the element-wise product of ρ and Δ
$X = (x_1, \dots, x_n)$	The dataset with x_i as its i-th data point
$N(i) = (x_1, \dots, x_k)$	The set of k-nearest neighbors of point i
$Z(i) = (x_1, \dots, x_k)$	The set of k points with the highest similarity to point i

TABLE II
THE FEATURES OF COMMON-USED SYNTHETIC AND UCI DATASETS, N DENOTES THE NUMBER OF SAMPLES, D DENOTES THE NUMBER OF DIMENSIONS, AND NC DENOTES THE NUMBER OF NATURAL CLUSTERS BASED ON GROUND-TRUTHS.

ID	Type	Datasets	<N, D, NC>	ID	Type	Datasets	<N, D, NC>
1	Real	Sonar	<208, 60, 2>	7	Synthetic	Pathbased	<300, 2, 3>
2	Real	Wine	<178, 13, 3>	8	Synthetic	Aggregation	<788, 2, 7>
3	Real	Iris	<150, 4, 3>	9	Synthetic	Jain	<373, 2, 2>
4	Real	Heart	<303, 13, 2>	10	Synthetic	Spiral	<312, 2, 3>
5	Real	Pima	<768, 8, 2>	11	Synthetic	Moon	<210, 2, 2>
6	Real	Leaf	<340, 15, 30>	12	Synthetic	CMC	<1002, 2, 3>

TABLE III
PARAMETER CONFIGURATIONS OF ALGORITHMS.

Type	Algorithms	Parameters setting	Reference
Conventional	WKM	input k, β=2	[5]
	DBSCAN	Eps = 0.033 - 1.9, Minpts = 4 - 61	[7]
Representative	SNNDPC	$d_c = 2\% - 3\%, 5 \leq K \leq 30$	[8]
	DenMune	$1 \leq K \leq 50$	[9]
Sota	DPC-CE	$d_c = 2\%, T_r = 0.25, P_r = 0.3$	[10]
	VDPC	$d_c = 0.2\% - 50\%, \delta_t = 0.04 - 40$	[11]
	ICKDP	$d_c = 2\%, K/4 \leq k < K/2$	[12]
Ours	WDSC	$d_c = 2\% - 3\%, 5 \leq K \leq 30$	-

D. Clustering Performence of WDSC

All the results are averaged by 10 runs of the experiments. The bold numbers indicate the best performance for a given dataset and metric. WDSC consistently outperforms other algorithms on most datasets, as can be seen in Table IV where WDSC’s column has the most bold entries, highlighting its validity and reliability. **Observations: (1)** Although WDSC cannot reach 1 in the pathbased and aggregation datasets, it is

TABLE IV
ARI AND FMI PERFORMANCE OF WDSC AND SEVEN COUNTERPARTS.

ARI								
Data	**W-KM**	**DBSCAN**	**SNNDPC**	**DenMune**	**DPC-CE**	**VDPC**	**ICKDP**	**WDSC**
Pathbased	0.3113	0.4735	0.9294	0.9239	0.4623	**1**	0.4311	0.9699
Aggregation	0.5577	0.7997	0.9731	0.9941	0.9978	**1**	0.9978	0.9920
Jain	0.4445	0.4215	0.5612	**1**	**1**	**1**	0.6965	**1**
Spiral	-0.0048	0.4079	**1**	0.6887	**1**	**1**	0.0414	**1**
Moon	0.7748	0.0403	**1**	**1**	0.1426	0.3807	0.3574	**1**
CMC	0.0749	0.8434	0.8198	**1**	0.2140	0.1476	0.4490	**1**
Sonar	-0.0047	0.0012	0.0533	0.0001	0.0253	0.0052	0.0443	**0.0581**
Wine	0.8620	0.2524	0.8992	0.8819	0.3715	0.2860	0.3715	**0.9637**
Iris	0.6943	0.6789	0.9038	0.7455	0.6634	0.7074	0.7445	**0.9222**
Heart	0.1851	0.0626	0.1508	0.1410	0.1723	0.1723	0.1507	**0.4382**
Pima	0.0206	0.0199	0.0131	0.0107	0.0119	**0.1707**	0.0787	0.0996
Leaf	0.2576	0.0238	0.2510	0.0592	0.0114	0.0100	0.0430	**0.2620**
Total Rank	**69**	**77**	**41**	**48**	**57**	**45**	**58**	**18**
Avg Rank	**5.75**	**6.42**	**3.42**	**4.00**	**4.75**	**3.75**	**4.83**	**1.50**
Rank	**7**	**8**	**2**	**4**	**5**	**3**	**6**	**1**
FMI								
Data	**W-KM**	**DBSCAN**	**SNNDPC**	**DenMune**	**DPC-CE**	**VDPC**	**ICKDP**	**WDSC**
Pathbased	0.5457	0.5980	0.9529	0.9489	0.6652	**1**	0.6508	0.9799
Aggregation	0.6508	0.8842	0.9789	0.9953	0.9983	**1**	0.9983	0.9937
Jain	0.8200	0.6210	0.8123	**1**	**1**	**1**	0.8739	**1**
Spiral	0.3275	0.5650	**1**	0.8297	**1**	**1**	0.4184	**1**
Moon	0.9084	0.4114	**1**	**1**	0.6470	0.7116	0.7030	**1**
CMC	0.5048	0.8143	0.8896	**1**	0.6235	0.4759	0.6862	**1**
Sonar	0.5144	0.0012	0.5702	**0.7070**	0.5132	0.6653	0.5217	0.6087
Wine	0.9233	0.2524	0.9330	0.9215	0.5834	0.7009	0.5834	**0.9759**
Iris	0.9233	0.6789	0.9355	0.8321	0.7824	0.8084	0.8306	**0.9479**
Heart	0.5543	0.0626	0.6177	0.4556	0.6299	0.6299	0.6251	**0.7231**
Pima	0.5275	0.0199	**0.7106**	0.7080	0.7093	0.5550	0.7050	0.6274
Leaf	0.3167	0.0238	0.2930	0.0858	0.0811	0.1133	0.0946	**0.3452**
Total Rank	**68**	**88**	**38**	**41**	**53**	**42**	**60**	**23**
Avg Rank	**5.67**	**7.33**	**3.17**	**3.42**	**4.42**	**3.50**	**5.00**	**1.92**
Rank	**7**	**8**	**2**	**3**	**5**	**4**	**6**	**1**

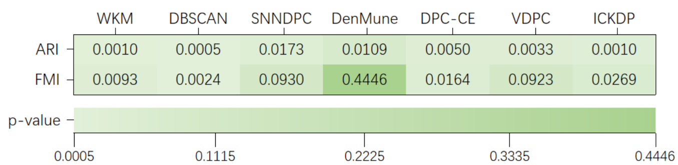

Fig. 1. p-values of Wilcoxon signed rank test in comparing out method against the other methods on ARI and FMI metrics.

the first and second closest algorithm to 1. **(2)** Real datasets like “Wine”, “Heart”, “Leaf” present a different challenge as they often contain noise and may have clusters without clear boundaries. The WDSC algorithm demonstrates strong performance on these datasets as well, suggesting that it is capable of dealing with the complexities of real-world data. **Conclusion:** WDSC’s superior performance across diverse datasets can be attributed to its innovative hybrid approach that combines subspace and density-based clustering with adaptive weighting and iterative optimization.

E. Significance Experiment Results

In the significance experiments, we compared the scores of the WSDC algorithm with other clustering algorithms on different datasets and evaluation metrics (ARI and FMI) using the Wilcoxon Signed Rank Test [13]. This was done to verify whether there is a statistically significant difference or advantage of our proposed algorithm over other clustering algorithms. The experimental results are shown in Fig. 1. In this experiment, we set the significance level at 0.05 to balance error control and result sensitivity. The larger the p-values, the darker the color.

F. Effectiveness of the Core Components of WDSC

We assessed the effectiveness of the core components of WDSC, namely the weight computation method and the learning mechanism. The effectiveness of the weighted approach was verified by comparing the representative density peak clustering algorithm SNNDPC (Baseline) with its weighted version, SNNDPC+W (Baseline+W), hereinafter referred to as WDC. The learning mechanism’s effectiveness was validated by comparing WDC with the complete version of WDSC. Results shown in Table V demonstrate that WDC outperforms SNNDPC on most datasets, proving the efficacy of the weight computation method, while WDSC significantly surpasses WDC, confirming the success of the learning mechanism.

TABLE V
ABLATION STUDY OF WDSC ON EIGHT DATASETS. THE BEST RESULTS IN TERM OF ARI AND FMI VALIDITY METRIC IS MARKED IN **BOLDFACE**.

Data	**Baseline**		**WDC**		**WDSC**	
	ARI	**FMI**	**ARI**	**FMI**	**ARI**	**FMI**
Pathbased	0.9294	0.9529	**0.9699**	**0.9799**	**0.9699**	**0.9799**
Aggregation	0.9731	0.9789	**0.9920**	**0.9937**	**0.9920**	**0.9937**
Jain	0.5612	0.8123	0.6017	0.8310	**1.0000**	**1.0000**
Sonar	0.0533	0.5702	**0.1026**	0.5518	0.0581	**0.6087**
Wine	0.8992	0.9330	0.9325	0.9552	**0.9637**	**0.9759**
Heart	0.1508	0.6177	0.4037	0.7028	**0.4382**	**0.7231**
Pima	0.0131	**0.7106**	0.0826	0.6047	**0.0996**	0.6274
Leaf	0.2510	0.2930	0.2575	0.3245	**0.2620**	**0.3452**

REFERENCES

[1] Aristides Gionis, Heikki Mannila, and Panayiotis Tsaparas, “Clustering aggregation,” *Acm Transactions on Knowledge Discovery from Data (tkdd)*, vol. 1, no. 1, pp. 4–es, 2007.

[2] Hong Chang and Dit-Yan Yeung, “Robust path-based spectral clustering,” *Pattern Recognition*, vol. 41, no. 1, pp. 191–203, 2008.

[3] Anil K. Jain and Martin HC Law, “Data clustering: A user’s dilemma,” in *PReMI*. Springer, 2005, pp. 1–10.

[4] Markelle Kelly, Rachel Longjohn, and Kolby Nottingham, “The uci machine learning repository,” `https://archive. ics. uci. edu`, 2023.

[5] Joshua Zhexue Huang, Michael K. Ng, Hongqiang Rong, and Zichen Li, “Automated variable weighting in k-means type clustering,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 657–668, 2005.

[6] Alex Rodriguez and Alessandro Laio, “Clustering by fast search and find of density peaks,” *science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[7] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al., “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *KDD*, 1996, vol. 96, pp. 226–231.

[8] Rui Liu, Hong Wang, and Xiaomei Yu, “Shared-nearest-neighbor-based clustering by fast search and find of density peaks,” *Information Sciences*, vol. 450, pp. 200–226, 2018.

[9] Mohamed Abbas, Adel El-Zoghabi, and Amin Shoukry, “Denmune: Density peak based clustering using mutual nearest neighbors,” *Pattern Recognition*, vol. 109, pp. 107589, 2021.

[10] Wenjie Guo, Wenhai Wang, Shunping Zhao, Yunlong Niu, Zeyin Zhang, and Xinggao Liu, “Density peak clustering with connectivity estimation,” *Knowledge-Based Systems*, vol. 243, pp. 108501, 2022.

[11] Yizhang Wang, Di Wang, You Zhou, Xiaofeng Zhang, and Chai Quek, “Vdpc: Variational density peak clustering algorithm,” *Information Sciences*, vol. 621, pp. 627–651, 2023.

[12] Wenjie Guo, Wei Chen, and Xinggao Liu, “Density peak clustering by local centers and improved connectivity kernel,” *Information Sciences*, vol. 666, pp. 120439, 2024.

[13] S.M. Taheri and Gholamreza Hesamian, “A generalization of the wilcoxon signed-rank test and its applications,” *Statistical Papers*, vol. 54, pp. 457–470, 2013.