

Density peak clustering evaluation

Viet-Vu Vu¹, Byeongnam Yoon², Dinh-Lam Pham^{1,2}, Hong-Quan Do³, Hai-Minh Nguyen⁴,
Tran-Chung Dao⁴, Thi-Hai-Yen Nguyen⁵, Doan-Vinh Tran⁶,
Thi-Huong-Ly Nguyen⁷, Viet-Thang Vu⁸

¹VNU Information Technology Institute, Vietnam National University, Hanoi, Vietnam

²Contents Convergence Software Research Institute, Kyonggi University, South Korea

³FPT University, Hanoi, Vietnam

⁴ICTU – University of Information and Communication Technology, Thai Nguyen University, Vietnam

⁵Hanoi Procuratorate University, Hanoi, Vietnam

⁶University of Education, Vietnam National University, Hanoi, Vietnam

⁷Nha Trang University

⁸Hanoi University of Industry

Correspondence: Viet-Vu Vu (vuvietvu@vnu.edu.vn)

Abstract—Density based clustering is one of the most tools used in clustering data analysis because of it can detect cluster with arbitrary shapes and noises. Some methods have been proposed in literature such as DBSCAN and density peak clustering. In density peak clustering (DPC), the results may directly depend on the method used in density estimation. In this paper, we conduct some experimentations with various data set to observe and analysis the results obtained by DPC when using some kinds of density estimation.

Keywords— Density based clustering, density peak clustering, density estimation.

I. INTRODUCTION

Given a data set $X = \{x_1, x_2, \dots, x_n\}$ in a space of d -dimensions; the aim of clustering algorithm is to divide X into k groups, called clusters, such that the points in the same cluster are similar and the points in different clusters are dissimilar. Although clustering analysis has been studied in nearly 70 years, it has attracted a lot of attention these days because of the need of new adaptation clustering method to real KDD applications.

Different clustering techniques have been proposed over the years such as Centroid-based (such as K-means [1], CLARA [2], Fuzzy C-Means [3]), Graph-based (such as SNN [4], CHAMELEON [5]), and Density-based (such as DBSCAN [6], OPTICS [7]).

In 2014, Rodriguez and Laio [8], introduced a new density-based clustering algorithm which based on the peak estimation. The main idea of DPC is using an estimation function to detect points having high density, those point called peaks. Each peak can be seen as a cluster center for each cluster, then the non-labeled point will be assigned the label as its nearest neighbor point that has a higher density. The DPC uses only one loop to detect clusters. The results of clusters obtained by DPC highly depends on the estimation of local density. The DPC is one of the most clustering methods that attract a lot of researches in recent years [10-14].

In order to extensively analyze the performance of DPC algorithm, as well as evaluate its advantages and disadvantages, in this work, we conduct a range of experiments on UCI and Gaussian data sets. From the obtained results, we furthermore propose some open research directions for improving the algorithm. A technical analysis of an algorithm on real-world data sets is also an important thing to be considered in order to help experimental researchers decide where it is appropriate and which algorithm should be used in a particular data mining process.

II. CLUSTERING BASED ON DENSITY PEAKS

The idea of density-based clustering is based on the observation: points in each cluster have higher densities than the surrounding regions. In general, algorithms like DBSCAN, OPTIC, or DPC usually use some concepts based on local density estimation. In DPC, for each data point x_i , we will define two factors as follows: the local density ρ_i and its relative distance δ_i . This calculation depends on the distance d_{ij} between the point x_i and x_j , for example, we can use the Euclidean distance here.

Firstly, the local density ρ_i of a point x_i is estimated by:

$$\rho(x_i) = \sum_{j \neq i} \mathcal{L}(d_{ij} - d_c) \quad (1)$$

where $\mathcal{L}(d) = 1$ if $d < 0$, otherwise $\mathcal{L}(d) = 0$ and a cut-off distance value d_c which represents the neighborhood radius of the data point. It can be seen that ρ_i denotes the number of data points whose distances to the point i are smaller than the d_c . In the case with small data set, we can use the equation follow for estimate the local density:

$$\rho(x_i) = \sum_{x_j \in X \setminus \{x_i\}} \exp\left(-\frac{(d(x_i, x_j))^2}{(d_c)^2}\right) \quad (2)$$

Secondly, the relative distance δ_i of the data point x_i to their nearest point x_j will be defined as the following equation:

$$\delta_i = \begin{cases} \min_j(d_{ij}) & \text{if } \exists x_j \text{ s.t. } \rho_j > \rho_i \\ \max_j(d_{ij}), & \text{for the point with highest density} \end{cases} \quad (3)$$

Intuitively, the selection of d_c has a great influence on the correctness of the clustering results. According to [8], d_c can be an adjustable parameter such that average number of neighbors is 1-2% according to different datasets.

After calculating two factors above, the DPC algorithm will construct a decision graph. From the decision graph, we can choose the points with high values of both ρ_i and δ_i called peaks, then the cluster center can be identified from those peaks. Figure 1 shows an example of a decision graph. Users can select directly the peaks in this case. However, how we can choose the right number of peaks is also an interesting problem. We need more knowledge about the data set/domain before choosing peaks.

In the final step, the non-labeled points are assigned to the same cluster label as their nearest neighbor point according to the recursive equation:

$$\eta(x_j) = \eta(\sigma(x_j)) \quad (3)$$

, where $\eta(x_j)$ is the label of x_j and $\sigma(x_j)$ is the closest point of x_j that has greater density.

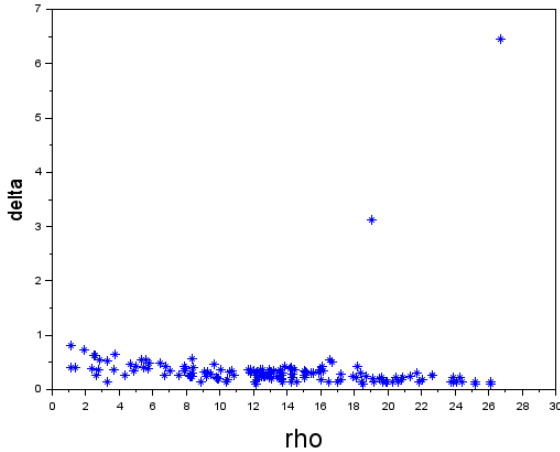


Figure 1. The decision graph for iris data set

The DPC algorithm is presented in Algorithm 1. The advantage of the DPC algorithm is that it uses only one parameter, and the step of assigning label needs to traverse the data one time. DPC algorithm can detect clusters of different sizes, shapes. However, when it deals with data set having different densities, the results of clustering may not correct. The main reason is that the peaks do not find correctly. In the figure 1, the Iris data set with 3 clusters, however, we see that there are only 2 peaks candidates. Another factor that can affect the results is the definition of local density. If the estimation of density is not adapted, the performance of clustering may

decrease. In the next section, we will conduct some experimentations for this problem.

Algorithm 1: DPC algorithm;

Input: a data set $X = \{x_1, x_2, \dots, x_n\}$, the cut-off distance value d_c ;

Output: a partition of X ;

Begin

1. Calculating ρ_i , $i = \overline{1, n}$, using Eq.1
2. Calculating δ_i , $i = \overline{1, n}$, using Eq.2
3. Constructing the decision graph based on ρ_i , and δ_i
4. Selecting peaks from the decision graph by users
5. For each unlabeled point x_j in the descending order of density, the label of x_j is assigned using Eq.3.
6. Detecting outliers [Optional]

End

III. EXPERIMENT RESULTS

A. Datasets

To evaluate the efficiency of DPC algorithm, we use 6 real-world datasets from UCI¹. These UCI datasets are widely used in the research community to compare and evaluate the clustering quality. Details of the datasets are given in Table 1.

TABLE 1. THE EXPERIMENTAL DATASETS

No.	Dataset	#Samples	#Features	#Clusters
1	Iris	150	4	3
2	Soybean	214	9	6
3	Ecoli	336	8	8
4	Yeast	1484	8	10
5	Protein	116	20	6
6	Glass	214	9	6

B. Evaluation Method

To compare two partitions of clustering, in this paper, we use two methods including Rand measure and F-measure [9, 15]. Given two partitions P and Q to compare, the idea of these methods is based on comparing their common and their different sets of pairs. The detail of these methods is shown in Table 2 and Table 3.

TABLE 2. CATEGORIES IN PAIRING CLUSTERED OBJECTS

Clustering	Pairs in P	Pairs not in P
Pairs in Q	$a := \text{Pairs in both}$	$b := \text{Pairs in Q}$
Pairs not in Q	$c := \text{Pairs in P}$	$d := \text{Pairs in none}$

TABLE 3. THE F-MEASURE AND RAND MEASURE

precision	recall	F-measure	Rand
$\frac{a}{a+c}$	$\frac{a}{a+b}$	$\frac{2a}{2a+b+c}$	$\frac{a+d}{a+b+c+d}$

¹ <https://archive.ics.uci.edu/ml/index.php>

C. Results

The results of clustering used DPC are shown in the table 4. We have tried with two methods of local density evaluation as mentioned above; the maximum results of 2 method used are presented in the table 4. The Rand measure achieved the good performance, however the F-measure for Ecoli, Yeast, Protein and Glass is quite low. It can be explained by the fact that those data set are highly overlap between clusters. The DPC with propagation label method make a lot of errors. We need further research for improving the performance of DPC in this case. Moreover, the method that choosing the right peaks needs being enhanced.

TABLE 4. PERFORMANCE OF DPC METHOD

No.	Dataset	Rand measure	F-measure
1	Iris	82.4	75.0
2	Soybean	90.7	80.0
3	Ecoli	75.8	48.4
4	Yeast	61.6	25.3
5	Protein	80.1	36.3
6	Glass	69.9	32.8

IV. CONCLUSIONS

In this paper, we have conducted some experimentation used DPC method. The aim is to evaluate the performance of DPC using two kinds of local density estimation. Although the DPC has obtained some good results, however some problem needs to be studied as follows:

- DPC do not work well if the data set has some overlap clusters.
- The method for choosing peaks should improve. If the data set has different densities, DPC will make some mistakes.

We also propose some direction research in near future such as building hybrid method for DPC, i.e., combining DPC with some optimization methods, or develop some kind of semi-supervised DPC clustering algorithms.

ACKNOWLEDGMENTS

This research is funded by Hanoi University of Industry under grant number 26 - 2022 - RD/HD - DHCN for Viet-Thang Vu (Vu Viet Thang).

REFERENCES

- [1] F. S. Samaria and A. C. Harter, "Some methods for classification and analysis of multivariate observations," in Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, pp. 281-297, Berkeley, CA, USA, 1967.
- [2] L. Kaufman and P.J. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis," John Wiley & Sons, 1990.
- [3] Bezdek, James & Ehrlich, Robert & Full, William, "FCM-the Fuzzy C-Means clustering-algorithm," Computers & Geosciences, vol.10, pp. 191-203, 10.1016/0098-3004(84)90020-7, 1984.

- [4] Ertöz, L., Steinbach, M.S., & Kumar, V, "Finding Clusters of Different Sizes, Shapes, and Densities in Noisy," High Dimensional Data, SDM, 2003.
- [5] Karypis, George & Han, Eui-Hong & Kumar, Vipin, "CHAMELEON A hierarchical clustering algorithm using dynamic modeling," Computer, vol. 32, pp. 68-75, 10.1109/2.781637, 1999.
- [6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei 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), AAAI Press, pp. 226-231, 1996.
- [7] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander, "OPTICS: ordering points to identify the clustering structure," SIGMOD Rec. 28, pp. 49-60, 10.1145/304181.304187, 1999.
- [8] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," Science, vol. 344, no. 6191, pp. 1492-1496, 2014.
- [9] W.M. Rand, "Objective criteria for the evaluation of clustering methods," Journal of the American Statistical Association. American Statistical Association. 66 (336): 846-850, 1971.
- [10] Mehmood, R.; Zhang, G.; Bie, R.; Dawood, H.; and Ahmad, H., "Clustering by fast search and find of density peaks via heat diffusion," Neurocomputing, vol. 208, pp 210-217, 2016.
- [11] Wang, S.; Wang, D.; Li, C.; Li, Y.; and Ding, G., "Clustering by Fast Search and Find of Density Peaks with Data Field," Chin. J. Electron, vol. 25, pp. 397-402, 2016.
- [12] Du, M.; Ding, S.; and Jia, H. "Study on density peaks clustering based on k-nearest neighbors and principal component analysis," Knowl. - Based Syst, vol. 99, pp. 135-145, 2016.
- [13] Jiang, Z.; Liu, X.; and Sun, M., "A Density Peak Clustering Algorithm Based on the K-Nearest Shannon Entropy and Tissue-Like P System," Math. Probl. Eng, 2019.
- [14] Sieranoja, S.; and Fränti, P., "Fast and general density peaks clustering," Pattern Recognition Letter, vol. 128, pp. 551-558, 2019.
- [15] Elke Achtert, Sascha Goldhofer, Hans-Peter Kriegel, Erich Schubert, Arthur Zimek: Evaluation of Clusterings - Metrics and Visual Support. ICDE 2012: 1285-1288.



Assoc. Prof. Dr. **Viet-Vu Vu** received the B.S. degree in Computer Science from Hanoi University of Education in 2000, a M.S. degree in Computer Science from Hanoi University of Technology in 2004, and a Doctor Degree in Computer Science from Paris 6 University in 2011. He is a researcher at Information Technology Institute, Vietnam National University, Hanoi. His research interests include clustering, active learning, semi-supervised clustering, and E-government applications



Byeongnam Yoon (M'97) He became a Member of IEEE in 1997. He was born in Seoul Korea 1949. He got the PhD in computer science, Chungnam National University, Korea, 1997. He worked for the Sperryrand UNIVAC as a Computer Specialist 1974 - 1978, Samsung as a Manager of Telecommunications Section 1978 - 1982, Electronics & Telecommunications

Research Institute (ETRI) as a Principal Researcher 1982 - 1999, National Information society Agency (NIA) as a Senior Executive Director General 1999 - 2010, Kyonggi University (KGU) as an Associate Professor, Faculty of Computer Science 2010 - 2016. He is an invited professor at Information Technology Institute, Vietnam National University, Hanoi Vietnam since 2017, and Global IT Research Institute (GIRI) as a President 1999 - current. His research areas include Telecommunications, Internet, Software, Web programming & security, e-Government, Enterprise Architecture, Work Flow, BPM, System Work Method. SPICE, CMMI, etc.



Dinh-Lam Pham is a research professor in the Contents Convergence Software Research Institute at Kyonggi University, South Korea. He received B.S. degree and M.S. degree in computer science from Thai Nguyen University in 2008 and 2010, respectively, and received Ph.D. degree in Computer Science from Kyonggi University in 2021. His research interests include video retrieving system, deep learning, workflow systems, BPM, workflow-supported social and affiliation networks discovery and analysis, process mining, deep-learning-based process predicting.



MSc. Hong-Quan Do received a double M.S. degree in Information and Communication Technology from University of Science and Technology of Hanoi, Vietnam and The University of Rennes 1, France in 2015. He is a lecturer at FPT University, Hanoi, Vietnam. His research concentrates primarily on Clustering, Semi-supervised Clustering, Image processing and Recommender Systems. At the present, he has also been involved in some E-Government projects, and E-Commerce Recommendation applications.



Dr. Nguyen Hai Minh received a PhD degree in Computer science and Engineering from Kyungpook National University, Korea in 2014. He is a dean of Faculty of Information Technology - ICTU. His study focuses on Health Informatics, Health IT Standards and Medical Image Processing



M.IT Tran-Chung Dao received a M.IT degree in Information Technology from Manuel S. Enverga University Foundation, Philippines 2013. He is a lecturer at Thai Nguyen University Information and communication technology, Thai Nguyen, Viet Nam. His research concentrates primarily on Clustering and Image processing.



Thi-Hai-Yen Nguyen is now a lecturer of Hanoi Procuratorate University, Hanoi, Vietnam. Her interest topic is machine learning and information retrieval.



Associate Professor, **Dr. Doan-Vinh Tran** received a PhD degree in Informatics Education from the Russian Academy of Educational Sciences, Russia in 1997. Now, he is a lecturer at the Faculty of Educational Technology, University of Education, Vietnam National University, Hanoi. His research concentrates primarily on Informatics Education, Digital Education, applications VR/AR/MR/XR in Digital Education, Clustering and Image processing.



Thi-Huong-Ly Nguyen received her MS degree in information technology from Nha Trang University in 2019. Since 2019, she has been with the Faculty of Information Technology at Nha Trang University, Vietnam, where she is now a lecturer. Her main research interests are machine learning, data analysis, and data science.



Dr. Viet-Thang Vu received the B.S. degree in Information Technology from Mendeleev University of Chemical Technology of Russia, 2008, a M.S. degree in Information Technology, Le Quy Don University, VietNam, 2012, and a Doctor Degree in Computer Science, Moscow Institute of Physics and Technology (MIPT), Russia, 2019. He is a head of division information system at Hanoi University of Industry, Vietnam. His research interests include cyber-security, image processing and computer vision, machine learning, data mining.