

Active learning for density peak clustering

Viet-Vu Vu^{*1}, Byeongnam Yoon^{**}, Cuong Le^{*}, Hong-Quan Do^{*}, Hai-Minh Nguyen^{***}, Chung Tran^{***}, Viet-Thang Vu^{****}, Cong-Mau Tran^{*****}, Doan-Vinh Tran^{*****}, Tien-Dung Duong^{*****}

^{*}VNU Information Technology Institute, Vietnam National University, Hanoi

^{**} Global IT Research Institute GIRI

^{***} Information and Communication Technology University, TNU

^{****}Hanoi University of Industry

^{*****}Ha Tinh University

^{*****} University of Education, Vietnam National University, Hanoi

^{*****} VietNam Academy of science and technology, Hanoi

¹Corresponding Author: vuvietvu@vnu.edu.vn

Abstract—Density peak clustering (DPC) is one of the most interesting methods in recent years. By using only one parameter, the algorithm can identify the number of clusters using peak estimation based on density and detect clusters in a single step. Although the DPC has lots of advantages, however, for some real applications, it is not easy to identify exactly the number of clusters because of the complex data distribution. In this paper, we propose a new active density peak clustering that aim to improve the clustering process for DPC. The main idea of our algorithm is adding a loop of active cluster centers selection for getting label from users. Experiments results show the effectiveness of our proposed solution.

Keywords—Clustering, Density peak clustering, Active Learning.

I. INTRODUCTION

The goal of clustering is to divide a set of objects into a finite and discrete set of clusters in which objects in the same cluster are similar while objects in different cluster are dissimilar follow some criteria of user domain [1,2]. Though clustering algorithms have long history, nowadays clustering topic has still attracted a lot of attentions because of the need of efficient data analysis tools in many applications such as web mining, spatial database analysis, GIS, textual document collection, image segmentation, etc. Given a data set of objects, clustering can detect the relation between data objects and the hidden structure of whole data objects and hence, it is the important tool in the data mining and knowledge discovery process.

In general, clustering methods can be grouped in some kinds including partition methods, hierarchical methods, density based method and model based methods [2]. Partition clustering aim to divide the data objects into k clusters using an objective function for optimizing. We can cite here some major algorithms such as K-means and Fuzzy C-Means. Hierarchical clustering (HC) algorithms usually organize data into a hierarchical structure according to the proximity matrix. The results of HC are usually depicted by a binary tree or dendrogram. From the organized dendrogram, we can decide which level of dendrogram may cut off to form the final clusters. The idea of density based clustering detects clusters

by identifying those regions where each region is a dense group of points. DBSCAN and SNN are two methods based on the density estimation concept. Model-based techniques assume that the distribution of data fit a model mathematic and the finding clusters are equal to an optimization process to fit between mathematic model and data distribution.

In 2014, Rodriguez and Laio introduced the density peak clustering method that is based on density estimation method [3]. The DPC has some advantage compared with DBSCAN [4] or other techniques as follow: it needs to use only one parameter and it can detect clusters with different shape and densities with noises. The algorithm has been attracted a lot of attention in research communities in which most of works try to improve clustering quality [5-7]. Several thousands of papers have cited the DPC algorithm up to date. In this paper, we focus on the two following problems:

- In real data set, the distribution of data may not follow some normal distributions. So it is not easy to match real cluster centers and peaks as DPC does.
- In each local region of data set, it has several peaks so these peaks may belong to the same cluster, this problem can generate some mistakes in the choosing cluster center process.

To tackle with these problems, in this paper we propose a new active density peak of clustering to improve clustering process by soliciting some labels from users to identify exactly the number of clusters for each data set. The experiments conducted on some UCI data sets show the effectiveness of our methods.

The rest of the paper is organized as follows: Section II discusses the related work. Section III presents our new active selection method, while section IV describes the experiments that have been conducted on benchmark datasets. Finally, section V concludes and discusses future research

II. RELATED WORKS

In this section, we will present the principle of density peak clustering and active learning method applied for clustering problem.

A. Density peak clustering and drawback

In 2014, Rodriguez and Laio proposed the density peak clustering which is the density based clustering [3]. With only one parameter, called d_c , the DPC uses the parameter to estimate the density of each data point and create a decision graph. From the decision graph, we can decide the number of clusters for each data set.

Firstly, the local density of a point x_i is defined as follow:

$$\rho_i = \sum_j X(d_{ij} - d_c) \quad (1)$$

in which d_c is a parameter, $X(x) = 1$ if $X < 0$ and $X(x) = 0$ otherwise. With the definition as the equation 1, $\rho(x_i)$ is the number of nearest neighbors of x_i within the d_c distance.

Secondly, the $\delta(x_i)$ is defined as the minimum distance between the point i and any other point with higher density (see equation 2).

$$\delta_i = \min_{j: \rho_j > \rho_i} (d_{ij}) \quad (2)$$

Thirdly, the decision graph is constructed in which the x-axis and the y-axis are respectively the ρ and δ values for every point in the data set. An example of decision graph is illustrated in figure 1. From the decision graph, we can choose some peaks to form center clusters.

Finally, the clusters will detect by using peaks with a propagation label process. The detail steps of the DPC are presented in Algorithm 1.

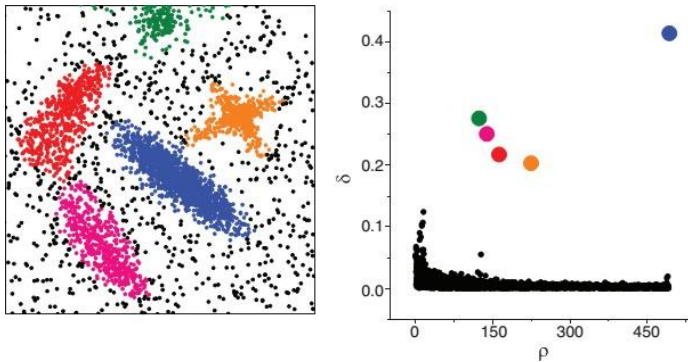


Figure 1. Data set (left) and the decision graph (right) [3]

Algorithm 1: Density peak clustering;

Input: a data set $X = \{x_1, x_2, \dots, x_n\}$, d_c

Output: a set of clusters

Step 1: Calculate $\rho(x_i)$ for each data point

Step 2: Calculate $\delta(x_i)$ for each data point

Step 3: Create decision graph

Step 4: Select k peaks from decision graph and label clusters named from 1 to k

Step 5: For each non-label x_j by the descending order of density, the label of x_j is defined recursively as follow:

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

Step 6: Output the results of clustering process

The DPC has shown the effectiveness compared with some clustering algorithms such as DBSCAN [4], CHAMELEON [11] because of it uses only one cluster and it can detect clusters with differences density with noises. Figure 2 shows another example that DPC can detect exactly the number of clusters. However, two drawbacks of DPC can be listed as follows:

- In real data set, the distribution of data may not follow some normal distribution. So it is not easy to match real cluster centers and peaks. In figure 3, an example of the decision graph for the real data set Iris¹. Iris data set has 3 clusters, however, in the decision graph we only choose 2 clusters centers.
- In each region of data set, it has several peaks so these peaks may belong to the same cluster.

These drawbacks mentioned above are also the research questions in this paper. To overcome these drawbacks, we propose to build an active learning algorithm that can collect exactly the number of cluster centers after some questions to users.

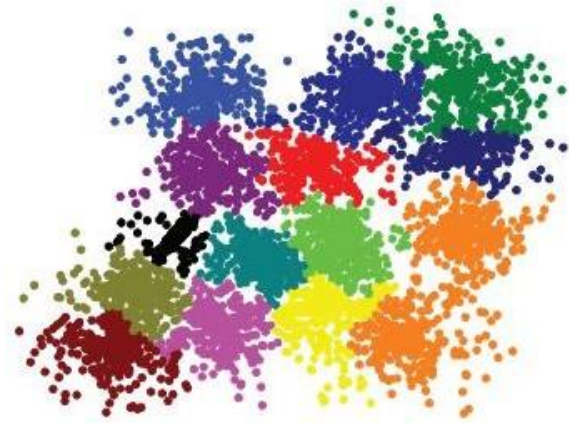


Figure 2. An example of overlap clusters [3]

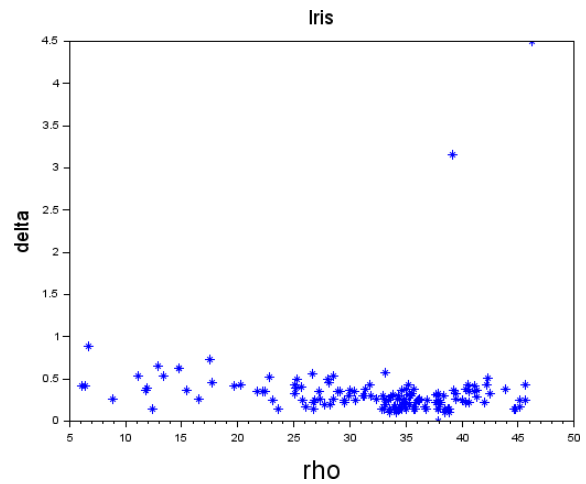


Figure 3. The decision graph of iris data set

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

B. Active learning

Active learning is a sub-field of machine learning topic. The aim of active learning is to develop algorithms that can learn new feedbacks at each step in a question-answering process [8]. In the context of semi-supervised clustering, active learning aims to select the most useful constraints so that they not only boost the clustering performance but also minimize the number of questions to the user/expert [9].

For active learning methods, we assume that the users/experts are always ready for answering the question proposed by active learning process. For example, in [9], Basu et al introduced a method for collecting constraints in which at each step, the algorithm will propose a question about the relating between two points and the users will answer the pair of point is must-link or cannot-link; in [10], an active learning method applied for K-means clustering has presented, the idea is using K-means to partition data set into large number of clusters, after that queries will be asked for each pair of clusters and the process will finish when users have been satisfied about the final results.

III. THE PROPOSED METHOD

The key different of our algorithm compared with the original DPC is the adding of a loop for soliciting label from users to identify cluster centers. This is the active loop based on the uncertainty strategy. To this aim, a loop (step 5 to step 8 in Algorithm 2) is added to form a user/question process. At each step of the loop, we will choose the point so that the multiple of rho and delta values is maxima and solicit users to get its label. The loop process will stop by users; it means that we have collected enough cluster centers. After that, the strategy as in the DPC is applied for clustering process and gets final results. The detail of algorithm is presented in Algorithm 2.

Algorithm 2: Active learning for density peak clustering;

Input: a data set $X = \{x_1, x_2, \dots, x_n\}$, d_c

Output: a set of clusters

Step 1: Calculate $\rho(x_i)$ for each data point

Step 2: Calculate $\delta(x_i)$ for each data point

Step 3: Create decision graph

Step 4: Set of labels $L = \{\}$; set of center points $C = \{\}$;

Step 5: Repeat

Step 6: Select a point x_k with maxima of $\delta \cdot \rho$ to get label from user;

Step 7: $L = L \cup \{\text{label}(x_k)\}$; $C = C \cup \{x_k\}$;

Step 8: Until (user stop = true);

Step 9: For each non-label x_j by the descending order of density, the label of x_j , noted $\eta(x_j)$, is defined recursively as follow: $\eta(x_j) = \eta(\sigma(x_j))$ in which $\sigma(x_j)$ is the point nearest of x_j with higher density.

Step 10: Output the results of clustering process

IV. EXPERIMENTATIONS

A. Experiment setup

To evaluate our new algorithm, we have used four data sets from UCI machine learning and two data sets are generated follow the Gaussian distribution. The details of these data sets are presented in Table 1 in which n, m, and k respectively are the number of data points, the number of features, and the number of clusters.

TABLE 1. MAIN CHARACTERISTICS OF THE DATASETS USING IN THE EXPERIMENTATIONS

No	Data	n	m	k
1	Iris	150	4	3
2	Soybean	214	9	6
3	Ecoli	336	8	8
4	Zoo	101	16	7
5	Art1	400	2	4
6	Art 2	800	2	2

To evaluate the clustering results we have used the Rand Index (RI) measure, which is widely used for this purpose in different researches [12]. The RI calculates the agreement between the true partition (P_1) and the output partition (P_2) of each data set.

To compare two partitions P_1 and P_2 , let u be the number of decisions where x_i and x_j are in the same cluster in both P_1 and P_2 . Let v be the number of decisions, where the two points are put in different clusters in both P_1 and P_2 . The RI measure is calculated using the equation as follow:

$$RI(P_1, P_2) = \frac{2(u + v)}{n(n-1)}$$

The value of RI is in the interval $[0 \dots 1]$; $RI = 1$ when the clustering result corresponds to the ground truth or user expectation. In our experimentation, we use the Rand Index in percentage. The higher the RI, the better the result of clustering.

B. Experiment results

Table 2 presents the Rand Index measure obtained by two methods DPC and ADPC. From the table, we can see that, with some questions proposed for users, we can collect exactly the number of clusters and hence the quality of clustering process had been improved. Some details explanations will be made as follows.

For the Iris data set, it has three clusters in which two clusters are overlaps, so DPC can not detect exactly the number of clusters as mentioned in the figure 3. In contrary, by soliciting label from users for data points with high density, we can collect the cluster centers and the clustering results will be improved. For Soybean data set, the explanation is similar to Iris, DPC cannot detect four clusters, using ADPC,

we can detect four cluster centers after five questions and the result is enhanced. For Ecoli data set, it is an imbalanced data with cluster size from 2 to 142, so we need 45 questions to collect the labels for cluster centers and ADPC obtained the RI with 89.8% while DPC cannot detect the number of clusters and ADPC obtained the good results compared with DPC. The results can be explained by the fact that real data sets always have complex distribution and the peak may appear in many regions in each cluster. Using decision graph to identify the cluster centers may have some troubles. Figure 4 shows the decision graph of Ecoli data set. It can be seen from the figure that it is not easy to choose the good cluster centers without using domain experts.

TABLE 2. RAND INDEX MEASURE FOR DPC AND ADPC

No	Data	DPC	ADPC	Number of Questions
1	Iris	88.6	92.7	4
2	Soybean	90.7	92.8	5
3	Ecoli	70.3	89.8	45
4	Zoo	90.5	100	28
5	Art1	88.3	88.3	4
6	Art 2	96.1	96.1	2

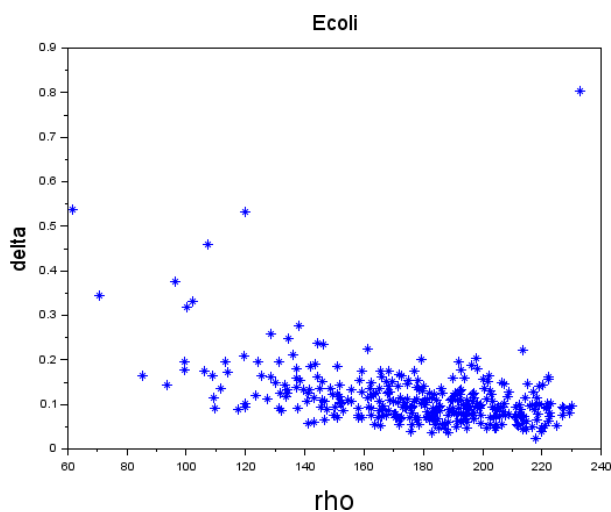


Figure 4. The decision graph of Ecoli data set

For the Art1 and Art2 data set, these are the data sets following the Gaussian distribution (see figure 5). So the DPC and ADPC can detect exactly the number of clusters and the accuracy in clustering results and the number of questions for active learning process is exactly equal to the number of cluster. This also can explain further for the UCI data sets that the real data set may not follow the normal/Gaussian distributions. In this case, it is not easy to detect exactly the

number of clusters. The active learning is one solution to overcome this drawback when we tackle with real applications.

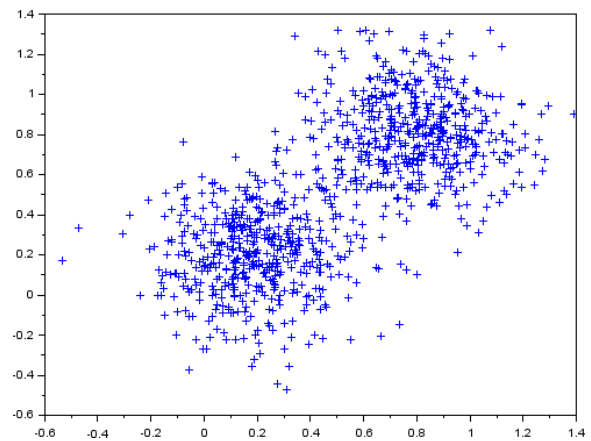


Figure 5. The plot of Art2 data set

V. CONCLUSION

In this paper, a new active learning clustering based on density peak clustering has been proposed. By soliciting label from users after a small set of questions, our algorithm can collect a good set of cluster centers and can detect clusters in straightforward. Experiments conducted on both artificial data and UCI data show the effectiveness of our algorithm and hence providing a new clustering technique for real applications. In the near future, some improvement of density peak clustering will be examined and applied to some applications such as image data or text data.

ACKNOWLEDGMENT

REFERENCES

- [1] Xu, D.; Tian, Y. A Comprehensive Survey of Clustering Algorithms. *Ann. Data Sci.* 2015, 2, 165–193.
- [2] Rui, X.; Wunsch, D. Survey of clustering algorithms. *IEEE Trans. Neural Netw.* 2005, 16, 645–678.
- [3] Rodriguez, A.; Laio, A. Clustering by fast search and find of density peaks. *Science*, 2014, 344, 1492.
- [4] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *KDD 1996*: 226–231.
- [5] Mehmood, R.; Zhang, G.; Bie, R.; Dawood, H.; Ahmad, H. Clustering by fast search and find of density peaks via heat diffusion. *Neurocomputing* 2016, 208, 210–217.
- [6] Wang, S.; Wang, D.; Li, C.; Li, Y.; Ding, G. Clustering by Fast Search and Find of Density Peaks with Data Field. *Chin. J. Electron.* 2016, 25, 397–402.
- [7] Du, M.; Ding, S.; Jia, H. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowl. -Based Syst.* 2016, 99, 135–145.
- [8] Burr Settles. *Active Learning Literature Survey*. Computer Sciences Technical Report 1648, University of Wisconsin–Madison. 2009.

- [9] Sugato Basu, Arindam Banerjee, Raymond J. Mooney: Active Semi-Supervision for Pairwise Constrained Clustering. SDM 2004: 333-344
- [10] Toon van Craenendonck, Sebastijan Dumancic, Hendrik Blockeel: COBRA: A Fast and Simple Method for Active Clustering with Pairwise Constraints. IJCAI 2017: 2871-2877.
- [11] George Karypis, Eui-Hong Han, Vipin Kumar, Chameleon: Hierarchical clustering using Dynamic Modeling, Computer, 32(8): 68 75 (1999).
- [12] 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.



Assoc. Prof. Dr. Viet-Vu Vu received the B.S. degree in Computer Science from Ha Noi University of Education in 2000, a M.S. degree in Computer Science from Ha Noi 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 area includes Telecommunications, Internet, Software, Web programming & security, e-Government, Enterprise Architecture, WorkFlow, BPM. System Work Method. SPICE, CMMI, etc.



Dr. Cuong Le is lecturer in School of Applied Mathematics and Informatics (SAMI), Hanoi University of Sciences and Technology (HUST) from 1998 to 2016. Since 2016 until now he in Information Technology Institute (ITI), Vietnam National University, Hanoi (VNU, Hanoi) since 2016. He got his PhD at HUST. His research interests lie in the area of information security, mathematics computation and quaternion and clifford analysis as well as PDE.



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 researcher at Information Technology Institute, Vietnam National University, Hanoi. 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.



Dr. Viet-Thang Vu received the B.S. degree in Computer Science from Mendeleev University of Chemical Technology of Russia in 2008, a M.S. degree in Computer Science from Le Quy Don Technical University in 2012, and a Doctor Degree in Computer Science from Moscow Institute of Physics and Technology in 2019. He is a teacher at Information Technology faculty,

Hanoi university of industry. His research interest is in the area of: clustering, computer vision, network security.



M.S Cong-Mau Tran received the B.S. degree in Computer Science from Hanoi University of Education in 2000, a M.S. degree in Computer Science from University of Technology – Vietnam National University, Hanoi in 2011. Now, he is a lecturer at the Faculty of Engineering – Technology, Ha Tinh University, Vietnam. His research interested include data mining, clustering, image processing and its applications.



Assoc. Prof. 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.



MSc. Dzung-Tien Duong received a M.S. degree in Information and Computing Technology from Peter the Great St.Petersburg Polytechnic University, Saint Petersburg, Russia in 2013. Now, he is a researcher at Vietnam Academy of science and technology, Hanoi, Vietnam. His research concentrates primarily on Clustering, Semi-supervised Clustering, semi-super fuzzy clustering.