

Active constraints selection based on density peak

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

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

^{**}Global IT Research Institute GIRI

^{***}ICTU – University of Information and Communication Technology, Thai Nguyen University, Vietnam

^{****}Ha Tinh University, Ha Tinh city, Vietnam

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

^{*****} People's Security Academy, Hanoi, Vietnam

^{*****}Hanoi University of Industry, Hanoi, Vietnam

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

¹Corresponding Author: vuvietvu@vnu.edu.vn

Abstract—Semi-supervised clustering which aims to integrate side information to improve the performance of clustering process, has received a lot of attentions in research community. Generally, there are two kinds of side information called seed (labelled data) and constraint (must-link, cannot-link). By integrating information provided by the user or domain expert, the semi-supervised clustering can produce expected results of users. In fact, clustering results usually depend on side information provided, so different side information will produce different results. In some cases, the performance of clustering may decrease if the side information is not carefully chosen. This paper addresses the problem of selecting good constraints for semi-supervised clustering algorithms. For this purpose, we propose an active learning algorithm for the constraints collection task, which relies on the min-max algorithm and peaks estimation based on density score. Experiments conducted on some real data sets from UCI show the effectiveness of our approach.

Keywords—Clustering, Semi-supervised clustering, Constraint, Density peak, Active Learning.

I. INTRODUCTION

Recently, semi-supervised clustering has received a lot of attentions in researcher communities [1, 2]. The advantage of semi-supervised clustering consists in possibility to use a small set of side information to improve clustering results. There are two kinds of side information including constraints and seeds. Constraints include must-link and cannot-link pairwise dependencies in which must-link constraint between two objects x and y means that x and y should be grouped in the same cluster and cannot-link constraint means that x and y should not be grouped in the same cluster. In real applications, we hypothesis that the side information is available or can be collected from users/experts.

When using constraints in semi-supervised clustering, a natural question rising is how we can have a good set of constraints before clustering. Many researches have been conducted on the problem of clustering with constraints [3-9] but most of the time the user is supposed to provide the algorithm with good constraints in a passive manner. One

alternative is to let the user actively choose the constraints. However, as some poorly chosen constraints can lead to a bad convergence of the algorithms [17] and as there is possibly $(n \times (n-1))/2$ ML or CL constraints in a dataset with n points, the choice of the constraints appears to be a crucial problem. This problem is fall in the active learning topic [10].

Figure 1 shows an example of constraints. In the figure must-link constraints are linked by the solid line while cannot-link constraints are linked in the dash lines.

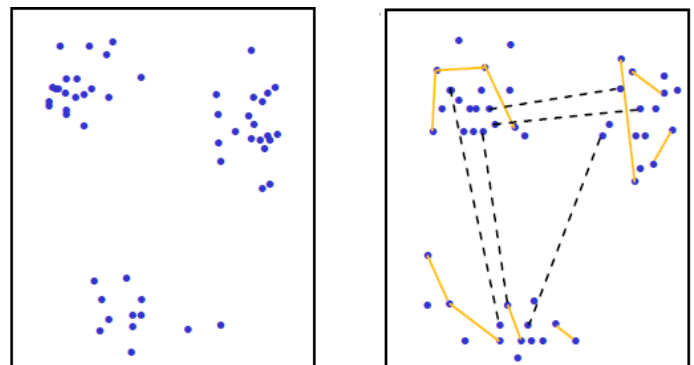


Figure 1. An example of data set (left) and constraints (right)

In this paper we propose a new active learning method for collecting constraints from users/experts. These collected constraints are applied for enhancing the results of semi-supervised clustering algorithms. The main idea of our new method is based on density peak estimation and min-max strategy. The experiments conducted on some UCI data sets show the effectiveness of our method.

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

II. RELATED WORKS

In this section, we will present some basic concepts using such as density peak detecting, active learning, and the min-max method. These methods will be used in our new algorithm in the next section.

A. Density peak detecting

In 2014, Rodriguez and Laio proposed the density peak concept in the context of clustering problem [11]. Using a parameter d_c , the local density of x_i is defined as follow:

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

in which $X(x) = 1$ if $X < 0$ and $X(x) = 0$ otherwise.

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

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

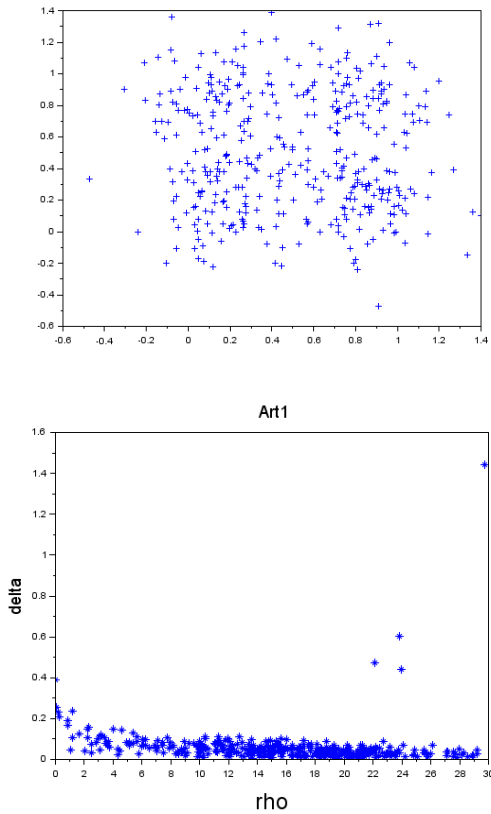


Figure 2. Date set Art1 (top) and the decision graph (bottom): four points in the right up corner of the figure can be seen as peaks of the decision graph

From two factors $\rho(x_i)$ and $\delta(x_i)$, we can build a decision graph in which the x-axis and the y-axis are respectively the ρ and δ values for whole data set. An example of decision graph is illustrated in figure 2. From the decision graph, we can choose some peaks as points with maxima local

density. Some improved version of DPC can be cited here such as these works in [12-14].

B. Active learning method

While active learning algorithms for supervised classification has been investigated for a long period of time [10], the problem of active learning for semi-supervised clustering was mentioned for the first time in the research on integrating prior knowledge in clustering proposed in 2001 [3]. The key idea behind active learning is that a machine learning algorithm can achieve greater accuracy with fewer training labels if it is allowed to choose the data from which it learns. An active learner may pose queries, usually in the form of unlabelled data instances to be labelled by an oracle (e.g., a human annotator). Active learning is well-motivated in many modern machine learning problems, where unlabelled data may be abundant or easily obtained, but labels are difficult, time-consuming, or expensive to obtain. Figure 3 present a general model for active learning applied for semi-supervised classification. The aims of active learning here is to build a machine learning model that can learn from unlabelled data and a small set of labelled data. Active learning for semi-supervised classification has been researching so far.

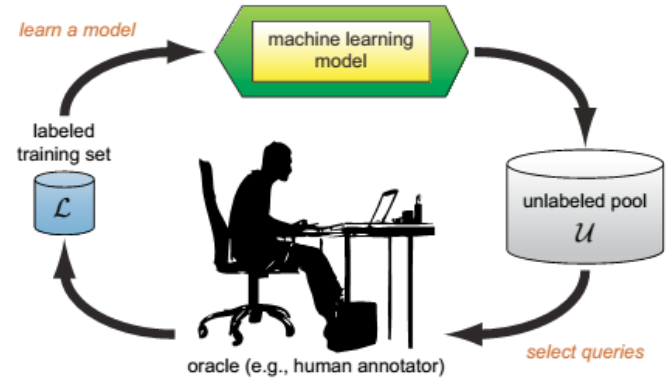


Figure 3. A general model of active learning [10]

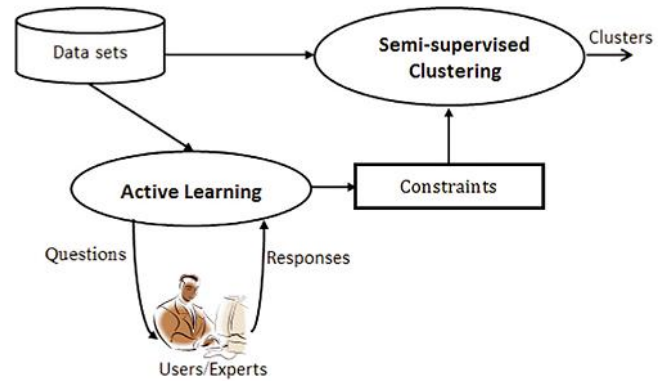


Figure 4. The schema of active learning for semi-supervised clustering [18]

The general idea has been rising when we work with the semi-supervised clustering problem. The situation is to choose a good set of constraints for semi-supervised clustering. As mentioned in the section 1, the choosing of constraints is the

crucial task because of a big number of candidates for examining. Figure 4 shows a schema for semi-supervised clustering.

Now, we will present some methods for active collecting constraints from users that can apply for semi-supervised clustering algorithms.

In [15], a method named FFQS (Farthest First Query Selection) is introduced for collecting constraints applied for constraints K-means. The FFQS contains two steps: Explore and Consolidate. The Explore step aims to build a set of CL constraints such that at the end of this step, at least one CL appears between every pair of clusters. In this case, we need to know the number of clusters before clustering process having done; the set of collected cannot-link constraints forms a skeleton of the clusters. After that, an iteration process will be used for choosing the farthest data point from existing skeleton to form query from users. The next step randomly picks a point not in the skeleton and queries it against each point in the skeleton, until the user decides to stop labelling the constraints.

In [16], Grira et al proposed an active constraint method for semi-supervised clustering. The key idea of the proposed method is based on the measure the border points in the clustering process. However, this method is only designed for the Fuzzy C-Means method.

In [17], Vu et al. introduced a method based on density score named ASC. The key idea of the method is to select the points in the sparse regions to form the query candidates to soliciate label from users. One measure of the quality of constraints set has been proposed based on the k-nearest neighbor graph. The method can collect constraints for all kind of semi-supervised clustering. However, some constraints may be not generated correctly follow the propagation process of the method.

Very recently, Abin and Vu proposed a querying informative constraint method based on density of data distributions [18]. They used three key assumptions about the usefulness of constraints for constructing the method as follows. (1) It provides helpful information about the boundary points of clusters or, (2) it provides helpful information about the boundary between different clusters or, (3) It provides helpful information about the skeleton of clusters. It provides helpful information about the skeleton of clusters. Based on these assumptions, a candidate constraint can be considered as a useful constraint if at least one of the conditions is present. The experiments conducted for some data sets show the better results compared with some existing methods.

C. The min-max method

Given a data set X with n point, the min-max method can find a subset Y with T points of X which can cover the distribution of X [15]. For this aim, the min-max method uses an iterative approach to build Y as the following steps:

- Step 1: $Y = \{y_1\}$, y_1 is randomly chosen from X
- Step 2: For step t ($t \leq T$), y_t is identified as follow:
 $y_t = \operatorname{argmax}_{x \in X} (\min\{d(x; y_i)\}); i = 1, \dots, t-1; Y = Y \cup \{y_t\}$;

Figure 5 shows an example of the min-max method, in this case we collect 7 data point for the set Y . We can see from the figure, the collected set Y can cover whole data set and thus it will be helpful in the collecting points to form user questions.

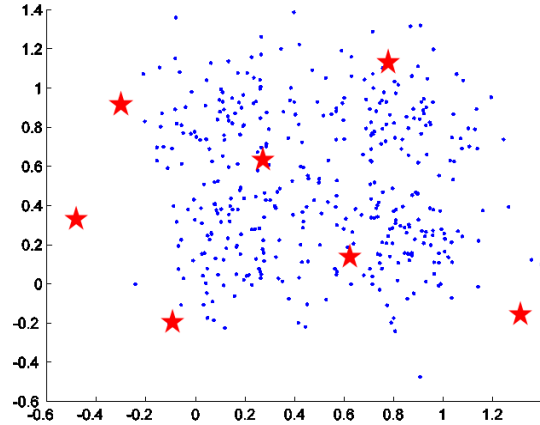


Figure 5. An example of min-max method: star points are collected by min-max method.

III. THE PROPOSED METHOD

Given a data set X with n points, the key idea of our algorithm is to apply density peaks estimation method and min-max method for building an active learning strategy. Firstly, we calculate local density and for every point of data and construct a decision graph. Secondly, we choose the peaks based on the decision graph, the set of chosen peaks are as an initial skeleton. Thirdly, we build an active learning process in which at each step, a new point is chosen following the min-max method using the skeleton and the point is used in forming user questions for getting label from users. The detail step of our algorithm is presented in Algorithm 1.

Algorithm 1: Active constraint selection based on density peaks;

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

Output: a set of constraints

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 skeleton peaks from the decision graph

Step 5: Repeat

Step 6: Select the point x_k that follows the min-max method;

Step 7: $t = 1$;

Step 8: Repeat

Step 9: Question to users for getting label of (p_t, x_k) ;

Step 10: $t = t + 1$;

Step 11: Until $(\text{label}(p_t, x_k) = \text{CL})$ or $(\text{user_stop} = \text{true})$;

Step 12: Until $(\text{user stop} = \text{true})$;

Step 13: Output the set of constraints;

The complexity of our algorithm is the process of the calculating $p(x_i)$ which is $O(n^2)$. However, for low dimensional data, we can use some data structure such as R-Tree that can reduce the complexity of the algorithm as $O(n \log n)$.

IV. EXPERIMENTATIONS

A. Experiment setup

To evaluate our new algorithm, we have used 6 data sets from UCI machine learning¹. These UCI data sets have been chosen because they facilitate the reproducibility of the experiments and because some of them have already been used in semi-supervised clustering algorithms. The details of these data sets are shown in Table 1 in which n , m , and k respectively are the number of data points, the number of attributes, and the number of clusters.

We use the MCSSGC [19], a semi-supervised graph-based clustering algorithm, to measure the effectiveness of constraints collected by our algorithm.

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	Protein	115	20	6
5	Zoo	101	16	7
6	Thyroid	215	5	3

To estimate the clustering efficiency, we have used the Rand Index (RI) measure, which is widely used for this purpose in different researches [20]. The RI calculates the agreement between the theoretical partition (P_1) and the output partition (P_2) of each data set by the evaluated clustering algorithm.

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 following equation:

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

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

B. Experiment results

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

Table 2 presents the Rand Index measure obtained by MCSSGC using constraints collected by our method and constraint randomly chosen. From the table 2, we can see that with the same constraints, the MCSSGC obtains the better results compared with using constraint collected randomly. It can be explained by the fact that our algorithm can collect constraints in the good strategy based on cluster centres and min-max method. Intuitively, our method can choose the candidates which cover the whole data set and let the clustering algorithm produces the good results.

TABLE 2. RESULTS OF RAND INDEX MEASURE (%)

No	Data	Number of constraints	MCSSGC+ Random	MCSSGC+ proposed
1	Iris	140	91.4	93.7
2	Soybean	60	98.9	100
3	Ecoli	200	90.1	91.5
4	Protein	120	81.9	83.3
5	Zoo	100	98.3	98.8
6	Thyroid	180	80.9	81.3

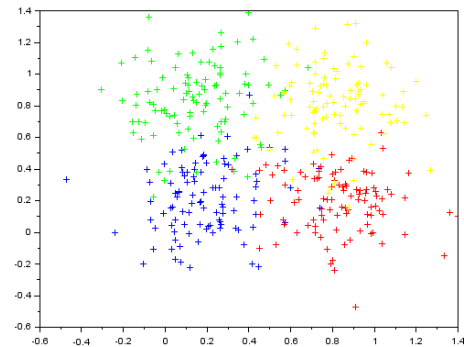


Figure 6. The visualisation of art1 data set

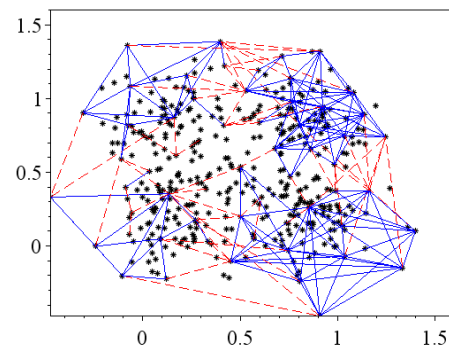


Figure 7. The constraints collected by our method: must-link and cannot-link constraints are illustrated by solid lines and dashed lines respectively

Figure 7 presents an example about the constraints collected by our method for the data set Art1 (see figure 6). The data set Art1 is a data set which has four clusters with some overlaps between clusters. As we can see from figure 7, our method can collect constraints in the border regions between clusters that are very useful for semi-supervised clustering algorithms in the clustering process, this is an evident showing the effectiveness of our method.

V. CONCLUSION

In this paper, we proposed an efficient active learning algorithm for collecting constraints from users/experts. Our method is based on the idea of density peaks estimation method and min-max strategy. Experiment conducted on UCI data sets show the effectiveness of our proposed method. In the near future, we will continue to extend our method for some specific kinds of data and apply for some real KDD applications.

ACKNOWLEDGMENT

This research has been done under the research project QG.21.58 “[Researching and developing clustering integrating constraints and deep learning algorithms]” of Vietnam National University, Hanoi.

REFERENCES

- [1] Rui, X.; Wunsch, D. Survey of clustering algorithms. *IEEE Trans. Neural Netw.* 2005, 16, 645–678..
- [2] Xu, D.; Tian, Y. A Comprehensive Survey of Clustering Algorithms. *Ann. Data Sci.* 2015, 2, 165–193.
- [3] S. Basu, I. Davidson, and K. L. Wagstaff, *Constrained Clustering: Advances in Algorithms, Theory, and Applications*, Chapman and Hall/CRC Data Mining and Knowledge Discovery Series, 1st edn., 2008.
- [4] Sugato Basu, Mikhail Bilenko, Raymond J. Mooney: A probabilistic framework for semi-supervised clustering. *KDD 2004*: 59–68.
- [5] Nizar Grira, Michel Crucianu, Nozha Boujemaa: Semi-Supervised Fuzzy Clustering with Pairwise-Constrained Competitive Agglomeration. *FUZZ-IEEE 2005*: 867–872
- [6] Carlos Ruiz, Myra Spiliopoulou, Ernestina Menasalvas Ruiz: Density-based semi-supervised clustering. *Data Min. Knowl. Discov.* 21(3): 345–370 (2010)
- [7] Rajul Anand, Chandan K. Reddy: Graph-Based Clustering with Constraints. *PAKDD (2) 2011*: 51–62.
- [8] Toon van Craenendonck, Sebastijan Dumancic, Hendrik Blockeel: COBRA: A Fast and Simple Method for Active Clustering with Pairwise Constraints. *IJCAI 2017*: 2871–2877.
- [9] Hongjing Zhang, Tianyang Zhan, Sugato Basu, Ian Davidson: A framework for deep constrained clustering. *Data Min. Knowl. Discov.* 35(2): 593–620 (2021)
- [10] Burr Settles. *Active Learning Literature Survey*. Computer Sciences Technical Report 1648, University of Wisconsin–Madison. 2009.
- [11] Rodriguez, A.; Laio, A. Clustering by fast search and find of density peaks. *Science*, 2014, 344, 1492.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] Sugato Basu, Arindam Banerjee, Raymond J. Mooney: Active Semi-Supervision for Pairwise Constrained Clustering. *SDM 2004*: 333–344
- [16] Nizar Grira, Michel Crucianu, Nozha Boujemaa: Active semi-supervised fuzzy clustering. *Pattern Recognit.* 41(5): 1834–1844 (2008)
- [17] Viet-Vu Vu, Nicolas Labroche, Bernadette Bouchon-Meunier: Improving constrained clustering with active query selection. *Pattern Recognit.* 45(4): 1749–1758 (2012).
- [18] Ahmad Ali Abin, Viet-Vu Vu, A density-based approach for querying informative constraints for clustering. *Expert Syst. Appl.* 161: 113690 (2020)
- [19] Viet-Vu Vu, Hong-Quan Do, Graph-based Clustering with Background Knowledge. *SoICT 2017*: 167–172.
- [20] 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 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 area includes Telecommunications, Internet, Software, Web programming & security, e-Government, Enterprise Architecture, Workflow, BPM. System Work Method. SPICE, CMMI, etc.



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.



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.



Mrs. Thi-Nhuong Phi received the B.S. degree in Electronics and Telecommunications Engineering from Thai Nguyen University in 2007, a M.S. degree in Electronics - Telecommunications Technology from College of Technology Confers – Vietnam national university, Hanoi in 2010. She is a teacher at People's Security Academy. Her research interests include technology, engineering,

and information security.



Dr. Viet-Thang Vu received the B.S. degree in Computer Science from Mendelev 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



MSc. Dzong-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.