

An empirical study for density peak clustering

Viet-Vu Vu*, Byeongnam Yoon**, Hong-Quan Do*, Hai-Minh Nguyen***, Tran-Chung

Dao***, Cong-Mau Tran****, Doan-Vinh Tran*****,

*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

****HaTinh University, HaTinh city, Vietnam

*****University of Education, Vietnam National University, Hanoi, Vietnam

vuvietvu@vnu.edu.vn

Abstract—Density Peak Clustering (DPC) is one of the most effective density-based clustering algorithms due to its ability to detect arbitrary clusters while being robust to noise. Since the first introduction in 2014, it has been cited a thousand times. This paper presents a comprehensive analysis of the DPC algorithm's performance on some UCI and Gaussian data sets. These used data sets have different properties such as intersecting clusters, unbalanced data, or different densities, such that not many clustering algorithms can perform well. From the obtained results, we aim to evaluate advantages and disadvantages of the algorithm and propose some open research directions.

Keywords—Density based clustering, density peak clustering, complex data.

I. INTRODUCTION

Clustering aims to split a data set $X = \{x_1, x_2, \dots, x_n\}$ with n elements into k clusters such that the elements in the same cluster are similar and the elements in different clusters are not. Although this is a classical problem of machine learning and has been studied since the 50s of the 20th centuries, it still attracts many researchers these days. The essence of the clustering is to discover hidden structures behind the data, detect the relationships between elements, and anomalies in the data. Due to the explosive growth of information technology, data become larger and more complicated; thereby it requires effective clustering algorithms to perform as a step of the data mining process.

Different clustering strategies have been reported over the years and can be categorized 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]). Density-based clustering algorithms like DBSCAN, and OPTICS can discover clusters of different sizes and shapes while being robust to noise. Dense regions in the data space are separated from those with lower density. However, most of these algorithms suffer from choosing their appropriate parameters for different datasets, for example the density threshold.

Subsequently, Rodriguez and Laio, 2014 [8] proposed a novel density-based clustering algorithm through fast search and density peaking (named as DPC). The algorithm attracts

many researchers' attention [10-16], and has a thousand citations till now. The DPC algorithm uses the local density of each point x_i and its relative distance, which represents the closest distance from the point to larger density points, to identify cluster centers for complex data distribution without specifying the number of clusters while being robust to noise. The non-center points then can be assigned to the cluster of the nearest higher density neighbor. Although the DPC algorithm is simple and effective, several issues are limiting its practical application. 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. DENSITY PEAK CLUSTERING

The density peak clustering, named DPC, was developed by Rodriguez and Laio, 2014 [8] based on two intuitive assumptions: (1) Cluster centers have higher densities than the surrounding regions; (2) Distances among cluster centers are relatively large. To utilize the assumptions, two important factors are computed for each data point x_i , namely the local density ρ_i and its relative distance δ_i . This calculation depends on the value d_{ij} which represents the distance between the point x_i and x_j (In [8], it is the Euclidean distance); and the cut-off distance d_c which represents the neighborhood radius of the data point.

On the one hand, the local density ρ_i of the point x_i is given by:

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

, where $\mathcal{L}(d) = 1$ if $d < 0$, otherwise $\mathcal{L}(d) = 0$. It can be seen that ρ_i denotes the number of data points whose distances to the point i are smaller than the cut-off distance. On the other hand, the relative distance δ_i of the data point x_i to their nearest point x_j with higher density ρ_j is defined as:

$$\delta_i = \begin{cases} \min_j(d_{ij}) & \text{if } \exists x_j \text{ s. t. } \rho_j > \rho_i \\ \max_j(d_{ij}), & \text{for the point with highest density} \end{cases} \quad (2)$$

Obviously, 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 $k\%$ of all nodes (normally 1-2% according to different datasets). It thus is defined by:

$$d_c = d_{N \cdot k\%} \quad (3)$$

, where N is the serial number of the last data point after the ascending order of all the distances d_{ij} , and it is also the total number of points.

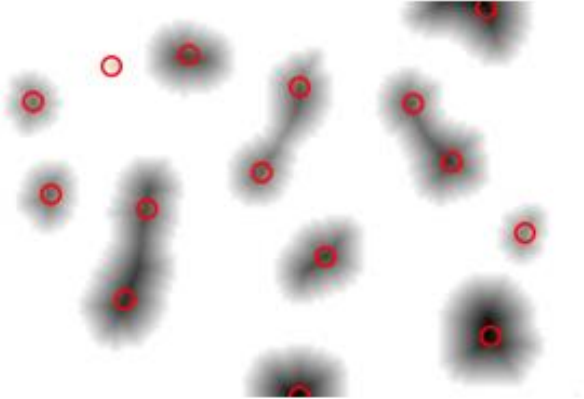


Figure 1. The idea of DPC is based on the concept of BLOB detection in which the boundaries where the water from two markers meets each other indicates where the BLOBs should be split [18].

After calculating the local density and the relative distance of all data points, the DPC algorithm establishes a decision graph. In [8], it has been shown that the point with high values of both ρ_i and δ_i is called a peak, then the cluster center can be selected from those peaks. In the final step, the non-center points are assigned to the same cluster as their nearest neighbor peak according to the recursive formula:

$$\eta(x_j) = \eta(\sigma(x_j)) \quad (4)$$

, 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. The idea of DPC is somewhat like the concept of BLOB detection algorithm [17]. The steps in DPC algorithm are summarized in Algorithm 1.

Algorithm 1: DPC Clustering Algorithm

Input: the dataset $X = \{x_1, x_2, \dots, x_n\}$, the cut-off distance d_c

Output: cluster label for each data point

Begin

1. Calculate ρ_i , $i = \overline{1, n}$, according to Eq.1
2. Calculate δ_i , $i = \overline{1, n}$, according to Eq.2
3. Establish a decision graph according to on ρ_i , and δ_i

4. Select k peaks from the decision graph
 5. For each unlabeled point x_j in descending order of density, the label of x_j is determined by Eq.4
- End*

Suppose that there are n points in the dataset. The time complexity of DPC depends on the following parts: (1) calculating the Euclidean distance between points takes $O(n^2)$; (2) computing the local density and relative distance for all points takes $O(n^2)$; (3) the assignment of non-center points to the clusters takes $O(n)$. Therefore, the total approximate time complexity of DPC is $O(n^2)$.

The advantage of the DPC algorithm is that it uses only one parameter, and the assignment of non-center points to the clusters only needs to traverse the data once. With the basis of the local density and the relative distance, DPC algorithm can detect clusters of different sizes, shapes, and densities.

III. EXPERIMENTS AND DISCUSSIONS

A. Datasets

To evaluate the efficiency of DPC algorithm, we use real-world datasets from UCI¹ and synthetic datasets Art1 and Art2 – both are with Gaussian distribution. 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	Zoo	101	16	7
5	Art1	400	2	4
6	Art 2	800	2	2

B. Evaluation Method

The Rand Index (RI) [9] is a measure of the similarity between two data clusterings. It is widely used in evaluation of clustering results. RI measure computes the agreement between the theoretical partition of each data set and the output partition of evaluated algorithms. This measure is based on $\frac{n(n-1)}{2}$ pairwise comparisons between the n points of a data set X . For each pair of points x_i and x_j in X , a partition assigns them either to the same clusters or to different clusters. Let us consider two partitions P_1 and P_2 , and let a be the number of decisions where the point x_i is in the same cluster as x_j in P_1 and P_2 . Let b be the number of decisions where the two points are placed in different clusters in both partitions. A total agreement can then be calculated as shown in Equation 5.

$$RI(P_1, P_2) = \frac{2(a+b)}{n(n-1)} \quad (5)$$

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

The value of RI is a number between 0 and 1. $RI = 0$ when the two clusterings have no similarities, and $RI = 1$ when the clustering are identical. In our experiments, we calculate RI in percentage. A higher the RI value indicates a better clustering performance.

C. Comparative Results

We first look at the Iris dataset in Figure 2, which is a 150-sample dataset. Two among three clusters of the dataset are very close to each other. In this case, only 2 peaks can be identified conveniently from the decision graph of Iris. Obviously, when clusters overlap, DPC algorithm finds it difficulty in determining the cluster centers.

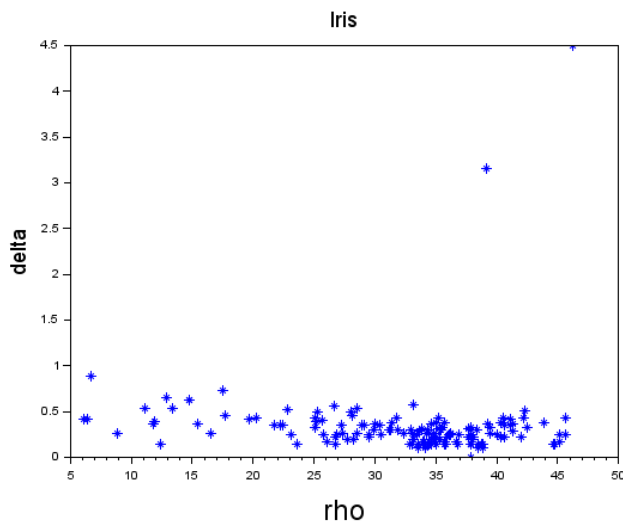


Figure 2. The decision graph of Iris dataset

The result for Ecoli dataset is shown in Figure 3. The dataset involves 8 clusters of hugely different sizes. One cluster can have a few points while another contains hundreds of points. Again, for those with hugely different in sizes, DPC has not proven its ability to determine peaks reasonably.

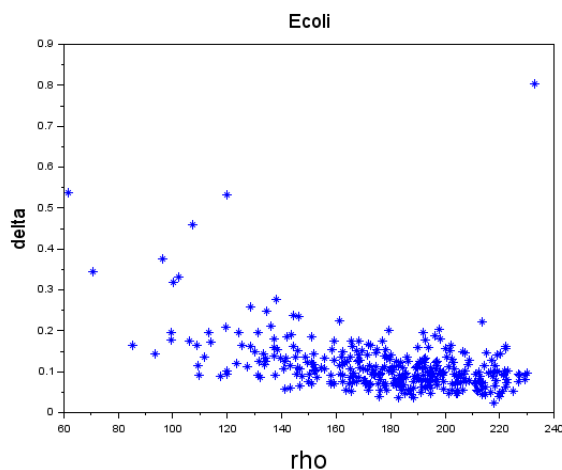


Figure 3. The decision graph of Ecoli dataset

The Soybean dataset consists of 4 clusters. From its decision graph as shown in Figure 4, it can be seen that if we determine 4 peaks at the top right position, there will be two points falling into the same cluster.

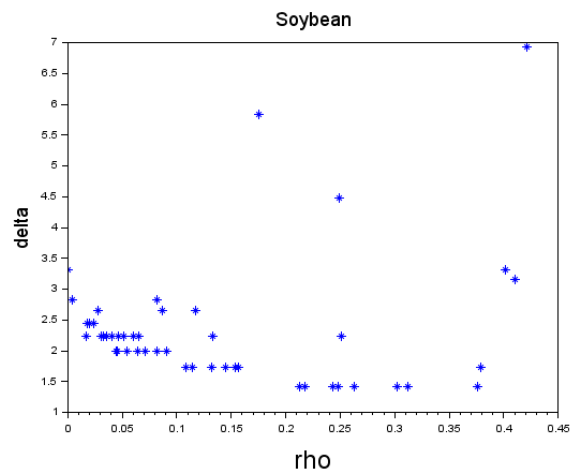


Figure 4. The decision graph of Soybean dataset

Similar to Soybean, it is not easy to determine all 7 cluster centers in the Zoo dataset as illustrated in Figure 5.

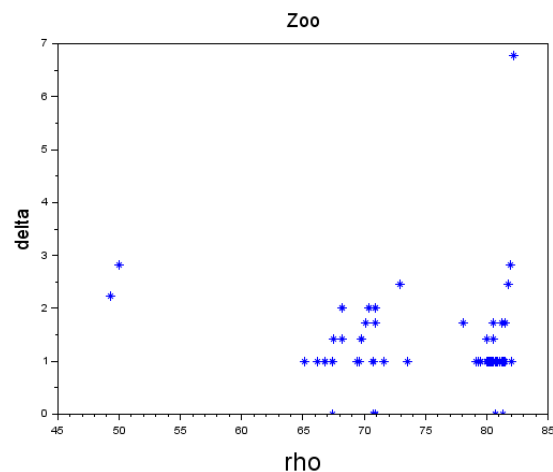


Figure 5. The decision graph of Zoo dataset

In the second experiment with the two datasets created according to Gaussian distribution, Art1 and Art2, we can see that the algorithm identifies the centers of clusters properly (Figure 6 and Figure 7). This claims the fact that the DPC algorithm works well for the datasets with Gaussian distribution.

Discussion:

From the analysis mentioned above, we can deeply understand the efficiency of DPC and also point out some comments as follows. First, it is difficult for DPC in determining the cluster centers where the dataset is unbalanced overlapped or the data has clusters in which there are local minima. Second, a combination with other algorithms to speed up the DPC algorithm can be considered, for example to reduce the time complexity of calculating the distance between

points. The study to apply different types of distances for each dataset can also be further considered. Finally, the research and development of semi-supervised clustering algorithms based on DPC is also a promising research direction.

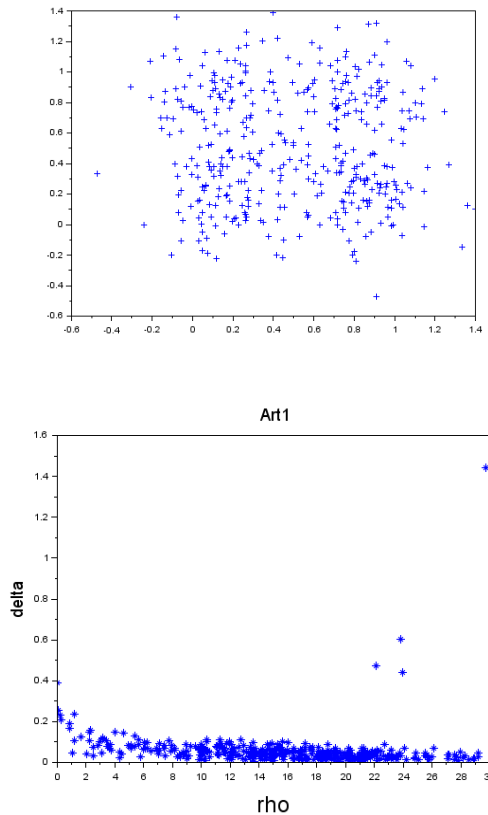


Figure 6. The Art1 dataset and its decision graph

Table 2 presents some clustering results according to Rand Index measure. Basically, DPC algorithm has better results than K-means in most dataset, except Ecoli. In the case of Ecoli, which is an unbalanced dataset, DPC cannot find the exact cluster centers, so that it leads to clustering performance degradation. K-Means, on the other hand, assumes that the number of clusters is known, thus it solves the case of Ecoli more effectively.

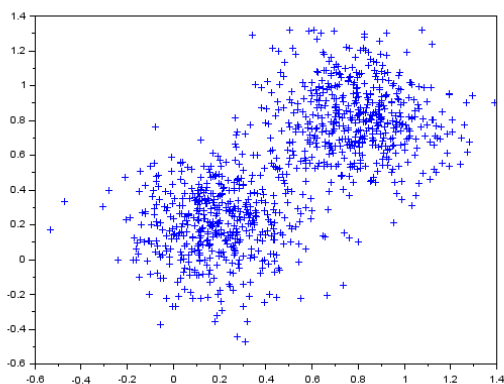


Figure 7. The Art2 dataset and its decision graph

TABLE 2. PERFORMANCE COMPARISON BETWEEN DPC AND K-MEANS

No.	Dataset	DPC	K-Means
1	Iris	88.6	85.1
2	Soybean	90.7	85.3
3	Ecoli	70.3	76.5
4	Zoo	90.5	84.3
5	Art1	88.3	87.5
6	Art 2	96.1	95.2

IV. CONCLUSIONS

In this paper, we have studied and evaluated the DPC algorithm. A range of experiments were conducted not only on read datasets but also on synthetic data in order to verify its performance. The obtained results show that DPC outperform K-means in terms of the Rand index, however, its advantages and disadvantages have also been depicted. To sum up, we highlight several issues that are limiting its practical application and propose some directions of further development of DPC as follows:

- First, it is difficult for DPC in determining the cluster centers where the dataset is unbalanced overlapped or the data has clusters in which there are local minima.
- Second, a combination with other algorithms to speed up the DPC algorithm can be considered, for example to reduce the time complexity of calculating the distance between points. The study to apply different types of distances for each dataset can also be further considered.
- Finally, the research and development of semi-supervised clustering algorithms based on DPC is also a promising research direction.

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] Tomi Kinnunen, Ilja Sidoroff, Marko Tuononen, and Pasi Fränti, "Comparison of clustering methods: A case study of text-independent speaker modeling," Pattern Recognit. Lett, vol. 32, pp. 1604-1617, 2011.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] Bai, L.; Cheng, X.; Liang, J.; Shen, H.; and Guo, Y., "Fast density clustering strategies based on the k-means algorithm," Pattern Recognition, vol. 71, pp. 375-386, 2017.
- [15] 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.
- [16] Sieranoja, S.; and Fränti, P., "Fast and general density peaks clustering," Pattern Recognition Letter, vol. 128, pp. 551-558, 2019.
- [17] Martin ter Haak, Machine learning for blob detection in high-resolution 3D microscopy images, Master thesis, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, 2018.
- [18] Martin ter Haak, Machine learning for blob detection in high-resolution 3D microscopy images, Master thesis, 2018.



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 include Telecommunications, Internet, Software, Web programming & security, e-Government, Enterprise Architecture, Work Flow, 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.