Hindawi
Scientific Programming
Volume 2020, Article ID 6458576, 12 pages
https://doi.org/10.1155/2020/6458576

Research Article

A New Clustering Algorithm and Its Application in Assessing the Quality of Underground Water

T. Vo-Van,[1] A. Nguyen-Hai,[2] M. V. Tat-Hong,[2] and T. Nguyen-Trang[3,4]

[1]*College of Natural Science, University of Can Tho, Can Tho City, Vietnam*
[2]*The Institute for Environment and Resources, Vietnam National University, Ho Chi Minh City, Vietnam*
[3]*Division of Computational Mathematics and Engineering, Institute for Computational Science, Ton Duc Thang University, Ho Chi Minh City, Vietnam*
[4]*Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam*

Correspondence should be addressed to T. Nguyen-Trang; nguyentrangthao@tdtu.edu.vn

Received 9 October 2019; Revised 7 January 2020; Accepted 4 February 2020; Published 7 March 2020

Guest Editor: Francisco Gomariz

Cluster analysis, which is to partition a dataset into groups so that similar elements are assigned to the same group and dissimilar elements are assigned to different ones, has been widely studied and applied in various fields. The two challenging tasks in clustering are determining the suitable number of clusters and generating clusters of arbitrary shapes. This paper proposes a new concept of "epsilon radius neighbors" which plays an essential role in the cluster-forming process, thereby determining both the number of clusters and the shape of clusters, automatically. Based on "epsilon radius neighbors," a new clustering algorithm in which the epsilon radius value is adapted to the characteristics of each cluster in the current partition is proposed. Recently, clustering has been widely applied in environmental applications, including underground water quality monitoring. However, the existing studies have simply applied conventional clustering techniques, in which the abovementioned two challenging tasks have not been solved already. Therefore, in this paper, the proposed clustering algorithm is applied in assessing the underground water quality in Phu My Town, Ba Ria-Vung Tau Province, Vietnam. The experimental results on benchmark datasets demonstrate the effectiveness of the proposed algorithm. For the quality of underground water, the new algorithm results in four clusters with different characteristics. Through this application, we found that the new algorithm might provide valuable reference information for underground water management.

1. Introduction

Cluster analysis is to discover the underlying structure of a dataset by partitioning the data into groups so that similar elements are assigned to the same group and dissimilar elements are assigned to different ones [1–5]. Recently, along with the development of big data, cluster analysis has been extensively studied and widely applied in various fields, such as physics, biology, economics, engineering, sociology, and data mining. [6]. For solving the problem of clustering, several approaches have been proposed in the literature, which includes: nonhierarchical clustering (k-means, k-means ++, etc. [7, 8] and other variances), hierarchical clustering [9], clustering for probability functions [1], or fuzzy clustering [10]. Among the abovementioned approaches, k-means clustering is the most well known and widely applied in various fields. However, the k-means algorithm and its extensions usually require a user-defined number of clusters that is often unknown in practice. (i) Furthermore, the k-means algorithm constructs spherical clusters, which is unsuitable for arbitrary-shaped clusters. (ii) The above two problems have been the major drawbacks of clustering so far, which lead to many difficulties and challenges in solving this problem [6].

For (i), to determine the suitable number of clusters, the most commonly used approach is running the clustering algorithm several times with different number of clusters each time, and evaluating them based on a number of

internal validity measures, such as S-index, F-index, Dunn index, and Xie-Beni index [11–14]. This approach can investigate the suitable number of clusters, but it repeats the clustering process many times to find the best number of clusters, thereby increasing the amount of time and space required, according to [6]. Moreover, the abovementioned evaluation indices are distance-based measures; therefore, they can only evaluate the qualities of spherical clusters and cannot be used for arbitrary-shaped clusters. In [15], Mavridis et al. proposed the algorithm PFClust (Parameter Free Clustering). The term "parameter free" means that the algorithm can automatically determine the number of clusters without requiring any user-defined parameters. For this purpose, PFClust performs an agglomerative algorithm on many subdatasets that are randomly sampled several times. Given an internal validity measure and a set of threshold corresponding to the number of clusters, the suitable threshold is then chosen based on the distribution of the given internal measure for all possible clustering results. In comparison to other conventional clustering algorithms, PFClust can result in a little better performance; however, it repeats the process of sampling and evaluating internal measures of the given thresholds in several times. Consequently, PFClust tends to be more time-consuming and expensive than other clustering methods. References [16–18] found the optimal partition by combining the metaheuristic optimization method and the clustering. These studies used the abovementioned internal validity measures as objective functions that need to be optimized to find the best clustering solution. It is well known that the metaheuristic optimization method, e.g., the genetic algorithm, results in an extreme computational cost, which reduces the efficiency of the algorithm. Furthermore, in spite of outputting the number of clusters and partitioning automatically, the metaheuristic optimization method requires a few of its own user-defined parameters that have effects on the optimal solution. As a result, avoiding the challenge of specifying the number of clusters, k leads to the challenge of specifying many other parameters. In [19], an automatic clustering algorithm was conducted using a function of force that can control the movements of the objects. The farther the distance, the weaker the force between two objects. In the end, each object converges to the center of the cluster it belongs to. Since the computing of force also requires a user-defined parameter denoted λ and the value of λ also has effects on the number of clusters, the attempt to overcome the problem of [16–18] of this algorithm is not too significant.

For (ii), DBSCAN [20], a density-based algorithm, is the most well-known method to construct arbitrary-shaped clusters. The algorithm utilizes two connectivity functions termed as density-reachable and density-connected, and each data instance is indicated as either a core point or a border point. The algorithm works to expand core points to form a cluster around itself. A drawback of DBSCAN is that when clusters of different densities exist, only particular kinds of noise points are captured [21]. Besides, two user-defined parameters regarding the minimum size of clusters and the radius need to be carefully turned. The other approaches, such as kernel k-means [22] and spectral clustering [23] can construct arbitrary-shaped clusters; these methods, however, also require a predefined number of clusters.

Because of the abovementioned drawbacks, an investigation of a new clustering method which can automatically determine the number of clusters and the clusters' shape is necessary. This paper proposes a new clustering method based on a new definition called "ε-radius neighbors" of a given point $\mathbf{x}_0$. ε-radius neighbors play a key role in constructing clusters with arbitrary shapes. When any new ε-radius neighbor is not found, the algorithm stops processing the current cluster and thereby the number of clusters is automatically determined. Furthermore, the radius ε can be adapted to specific cluster density, which is an advantage of the proposed methods in comparison with DBSCAN.

The quality of underground water depends on various factors, such as climate, characteristics of aquifers, pH, alkalinity, redox potential of the geological environment, initial sources, contamination due to human activities, and biological processes. The conventional methods of assessing the quality of groundwater are usually based on comparing the parameters representing water quality, which are collected by sensors, with the permitted standards. Clustering can help explain complex data matrix, analyze the similarities in water quality characteristics, and group them into clusters, thereby showing their general characteristics, as well as the causes that affect water quality. Therefore, clustering has been widely applied in environmental applications, including underground water quality monitoring. Some studies, for example, [24–28], have applied clustering in order to classify the water qualities in the whole region and design a future spatial sampling strategy in an optimal manner, which can reduce the number of sampling stations and associated costs. However, the abovementioned studies simply applied conventional clustering methods, such as hierarchical clustering with Ward distance, and k-means clustering. These methods, in general, have encountered the disadvantages, as mentioned in the previous parts. Therefore, in this paper, the proposed clustering algorithm is applied in assessing the underground water quality in Phu My Town, Ba Ria-Vung Tau Province, Vietnam. This application is expected to produce more reliable and valuable information so that the administrators can monitor underground water behavior.

The remainder of this paper is organized as follows. Section 2 presents the study area, the data collection, and the proposed method. The results and discussion are presented in Section 3 in which Section 3.1 is the validation of the proposed algorithms for different datasets and Section 3.2 is the application in assessing the underground water quality in Phu My Town, Ba Ria-Vung Tau Province, Vietnam. Finally, Section 4 is the conclusion.

2. Materials and Methods

2.1. The Proposed Clustering Method. Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, $\mathbf{x}_i \in R^d$ be a set of n points, $\mathbf{x}_0 \in R^d$ be a given point, and ε be an arbitrarily positive integer. A set $S \subseteq X$ is called as ε-radius neighbors of $\mathbf{x}_0$ if

$$S = \{\mathbf{x}_i \in X: d(\mathbf{x}_i, \mathbf{x}_0) \le \varepsilon\}, \quad (1)$$

where $d(\mathbf{x}_i, \mathbf{x}_0)$ is the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_0$.

Obviously, ε-radius neighbors of $\mathbf{x}_0$ are located in a hypersphere of radius ε around $\mathbf{x}_0$. As a result, a cluster can be extended by searching on the dataset and adding new objects pertaining to any hypersphere of radius ε around the current objects. This process still depends on the value of ε. This parameter plays a role which is the same as the parameter ε in the well-known DBSCAN algorithm. The choice of this parameter has effects on the clustering result. A fixed value of ε has low generalization ability because different datasets and clusters with different densities in a dataset could require different values of ε. A natural strategy is simply to adapt ε using the current cluster density. For the sake of presentation, the set of pairwise distances in the current cluster is called the set of "historical extending." Based on the set of "historical extending" in the current cluster (samples), we can estimate the maximum "extending" of the entire cluster (population). In this case, two basic principles are as follows:

(1) We know that if data has the normal distribution with mean μ and standard deviation σ, then 95% of the data values belong to the interval $\mu \pm 2\sigma$. If the two abovementioned parameters are unknown and data is enough large, we can estimate them from the sample data. For example, the mean and the adjusted standard deviation of the sample can be selected as alternatives for μ and σ, respectively. Therefore, to estimate the maximum extending of the cluster (population), we can use the following formula:

$$\varepsilon = \max D = \overline{d} + 2s_d, \quad (2)$$

where $\overline{d}$ and s_d are the mean and adjusted standard deviation of "historical velocities" in the current-processing cluster (sample). Obviously, about 97.5% of the extending pertaining to the true cluster (population) must be less than the extending estimated by formula (2), and thus, this formula can be used to approximate the maximum extending of the true cluster (population).

(2) Let n be the sample size or the number of objects in the current-processing cluster and $\overline{d}$ and s_d be the sample mean and adjusted standard deviation, respectively. Assuming that the value of n is large enough or d has the normal distribution with the mean $\mu(d)$ and the variance s_d^2/n. Consequently, with a significant level of 0.05, the mean of d belongs to the interval $\overline{d} \pm (1.96s_d/\sqrt{n})$. As a result, the maximum of the mean extending can be directly estimated using the following formula:

$$\varepsilon = \overline{d} + \frac{1.96s_d}{\sqrt{n}}. \quad (3)$$

The maximum value of the confidence interval is then used as the representative extending of the cluster.

It can be observed from formulas (2) and (3) that, in the earlier processing stage, when the sample size is too small, the standard deviation and the adaptive extending must be large. Therefore, we can avoid unreasonable extending in the earlier processing stage when the current sample is not a good representation of the population. Meanwhile, in the later stage, the number of objects in the current-processing cluster or the sample size is large enough for maintaining a stable adaptive extending.

Based on formulas (2) and (3), we propose a new clustering method called adaptive radius clustering for automatically determining the number of clusters and clusters shapes. Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, $\mathbf{x}_i \in R^d$ be an original dataset of N objects. The new clustering algorithm is presented as the following pseudocode and in Figure 1.

Initialize $C_1^{\text{old}} = \varnothing$, $C_1^{\text{new}} = \varnothing$, and $\text{centroid}_{\text{new}} = \varnothing$, where C_1^{old} and C_1^{new} are current-processing cluster obtained before and after an update, respectively.

Step 1. Get the first three objects of the cluster using the formulas below:

$$\mathbf{v}_1 = \arg\min_{x_i \in X} \sum_{j=1}^{n} d(\mathbf{x}_i, \mathbf{x}_j), \quad (4)$$

$$\mathbf{v}_2 = \arg\min_{x_i \in X/\{v_1\}} d(\mathbf{x}_i, \mathbf{v}_1), \quad (5)$$

$$\mathbf{v}_3 = \arg\min_{x_i \in X/\{v_1, v_2\}} d(\mathbf{x}_i, \mathbf{v}_1), \quad (6)$$

which subject to

$$d(\mathbf{v}_2, \mathbf{v}_1) \le d(\mathbf{v}_3, \mathbf{v}_1) < \binom{2}{n}^{-1} \sum_{i \ne j} d(\mathbf{x}_i, \mathbf{x}_j). \quad (7)$$

Update

$$C_1^{\text{old}} = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}; C_1^{\text{new}} = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}. \quad (8)$$

In formulas (4), (5), and (6), argument "arg" of a function is the value that must be provided to obtain the function's result; hence $\mathbf{v}_1 = \text{argmin}_{x_i \in X} \sum_{j=1}^{n} d(\mathbf{x}_i, \mathbf{x}_j)$ is a point $\mathbf{x}_i$ in X such that the sum of distances between it and other points is the minimum. In other words, $\mathbf{v}_1$ is the centroid of the current dataset. Similarly, $\mathbf{v}_2$ is the nearest point of $\mathbf{v}_1$ and $\mathbf{v}_3$ is the nearest point of $\mathbf{v}_1$ when excluding $\mathbf{v}_2$. Formula (7) is defined to overcome the problem of bad initialization. For example, if $\mathbf{v}_2$ and $\mathbf{v}_3$ are two nearest neighbors of $\mathbf{v}_1$, but the corresponding distances are larger than the average of pairwise distances between points in the current dataset, then $\mathbf{v}_1$ will be considered as a single cluster and the current extending process will be stopped.

In the abovementioned formulas, d is the Euclidean distance between any two d-dimensional points. In some illustration below, for the sake of visualization, $\mathbf{x}$ will be chosen as a 2-dimensional point ($\mathbf{x}_1$ and $\mathbf{x}_2$) so that we can draw the scatter plot of data. In fact, $\mathbf{x}_1$ and $\mathbf{x}_2$ not only can be the coordinates but also can be other informations such as height, weight, Ca^{2+}, Mg^{2+}, and Na^{+}. Furthermore, $\mathbf{x}$ can be a

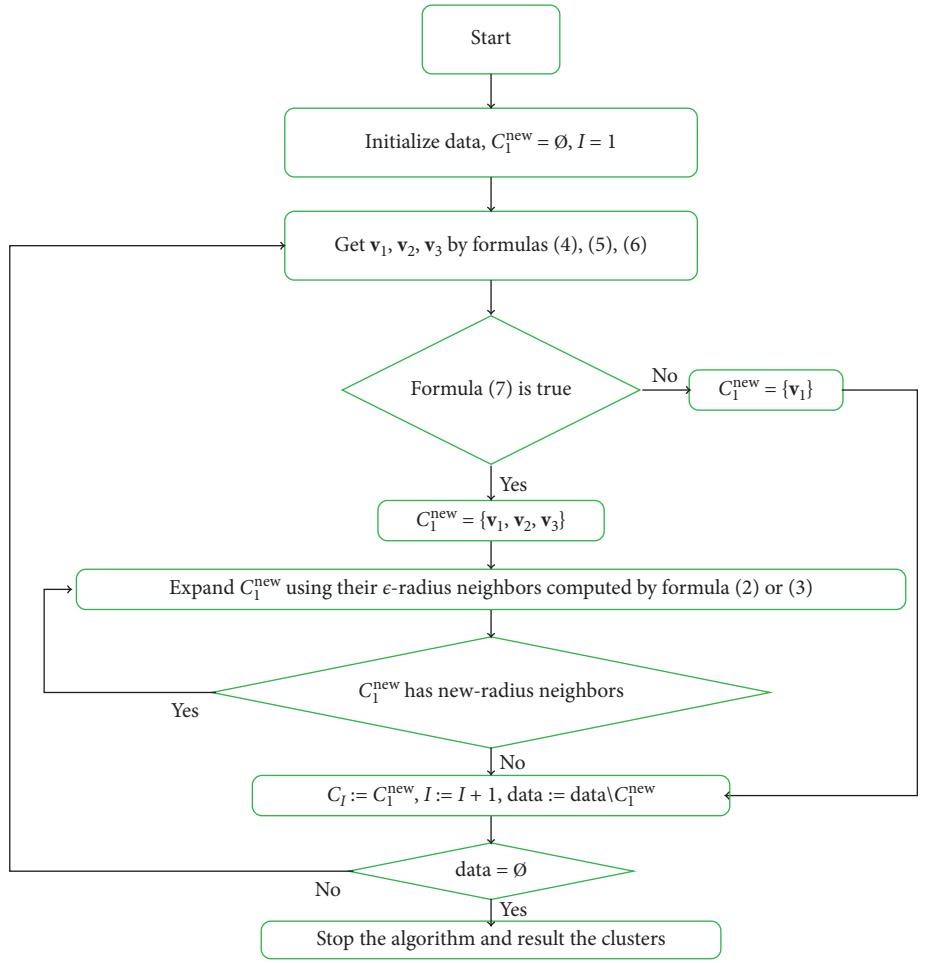


Figure 1: The flowchart of proposed algorithm, where I is the number of iterations.

d-dimensional vector, in general. Certainly, we can calculate Euclidean distance between two d-dimensional points $\mathbf{x}$ and $\mathbf{y}$ using the following formula:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}. \tag{9}$$

In addition, because variables measured at different scales do not contribute equally when calculating the distance, the data are normalized into [0, 1] interval using the following formula:

$$z_{ij} = \frac{x_{ij} - \min_i(x_{ij})}{\min_i(x_{ij}) - \min_i(x_{ij})}, \tag{10}$$

where x_{ij} is the value of variable j ($j = \overline{1, d}$) at the point i ($i = \overline{1, n}$), z_{ij} is the normalized value of variable j at point i and $\min_i(x_{ij})$ and $\max_i(x_{ij})$ are the minimum and maximum value of variable j, respectively.

Step 2. For each $\mathbf{v}_i \in C_1^{new}$, compute the adaptive ε-radius and the corresponding ε-radius neighbors S_i using Definition 1 and either formula (2) or formula (3); update C_1^{new} and centroid$_{new}$ by the following formulas:

$$C_1^{new} := C_1^{new} \cup S_i,$$
$$\text{centroid}_{new} := \frac{\text{centroid}_{new}}{v_i}. \tag{11}$$

In this step, formulas (2) and (3) are utilized to compute the adaptive ε-radius and the corresponding ε-radius neighbors S_i. Note that, the two abovementioned formulas are now just some options that need to be tested. In the numerical results, after applying both, the best option will be selected in the application.

Step 3. If $C_1^{new}/C_1^{old} \neq \emptyset$, then $C_1^{old} := C_1^{new}$ and centroid$_{new}$: = centroid$_{new} \cup C_1^{new}/C_1^{old}$. Repeat Step 2 and Step 3 until centroid$_{new} = \emptyset$, then stop the current-processing cluster.

Step 4. Repeat the three steps above until all objects are assigned to their clusters.

The main idea of the proposed algorithm is that from a number of points initialized using formulas (4), (5), and (6)

subject to (7), the cluster can automatically expand based on formulas (2) or (3). When the cluster does not extend more, the abovementioned process will repeat over the rest of the data until all points in the data are assigned to a specific cluster. With formulas (2) or (3), the ε-radius neighbor can adapt to different cluster densities; hence, the proposed algorithm can determine the number of clusters and find clusters of arbitrary shapes in cases of both balanced and imbalanced cluster densities. This is an advantage of the proposed algorithm in comparison to conventional methods, such as k-means, k-medoids, and DBSCAN.

2.2. Study Area and Data Used. The clustering method proposed above will be applied in assessing the underground water quality in Phu My Town, Ba Ria-Vung Tau Province, Vietnam. The study area and data used are described as follows.

2.2.1. Study Area. Phu My town has a natural area of 33,825 hectares and a population of 137,334 people. To the east, it borders Chau Duc district, Ba Ria-Vung Tau province. To the West, it borders Can Gio district, Ho Chi Minh City, and Vung Tau City, Ba Ria-Vung Tau province. To the South, it borders Ba Ria City, Ba Ria-Vung Tau province, and to the North is the Long Thanh district, Dong Nai province. Phu My town is located in the climate region of the Southern Delta, Vietnam, with a tropical climate and is influenced mainly by the northeast and southwest monsoon. There are two distinct seasons in a year, dry season and rainy season. The first lasts from December to April with an average annual temperature of 26.3 Celsius, and the second is between May and November with an average annual rainfall of 1356.5 mm.

Phu My town is the most concentrated industrial area and is one of the most developed areas in Ba Ria-Vung Tau province, Vietnam. To serve economic development, the demand for water in this area is quite high, but the sources of surface water from rivers and lakes do not meet the demand. According to the 2012 survey data of the Department of Natural Resources and Environment of Ba Ria-Vung Tau province, the total volume of underground water exploitation in this town had accounted for 18,608,430 m^3/year (mainly from Phu My-My Xuan water station and Toc Tien Water Plant). Groundwater exploitation has been reported to be mainly in the Pleistocene aquifer, which is composed of coarse-grained soil of Cu Chi Formation, Thu Duc Formation, and Trang Bom Formation with the main minerals: fluorite-apatite, feldspar, gypsum, tourmaline, montmorillonite, ilmenite, and some other impurities.

2.2.2. Data Used. The dataset has been provided by the Department of Natural Resources and Environment of Ba Ria-Vung Tau Province. The groundwater samples in the Middle-Upper Pleistocene (qp_{2-3}) aquifer and Upper Pleistocene (qp_3) aquifer, which consist of 11 variables, have been collected from 17 monitoring wells. The locations of 17 monitoring wells are shown in Figure 2, and the detailed dataset is presented in Table 1.

In this study, the contribution of variables is the same when calculating distance, that is, the proposed method considers the equal importance for each chemical parameter. In case in which some chemical parameters are more important than the others, the proposed method can be performed by using the weighted Euclidean distance instead of using the standard Euclidean distance. Also, note that, in this application, well's location is not considered as a variable, that is, the wells will only be grouped by their chemical parameters. The algorithm thereby will not be too focused on location, but more on chemical properties. Naturally, if wells in the same region have the same chemical properties, they will be assigned to the same cluster. As a result, we have wells sorted by locations. In contrast, through the clustering results, we can still identify wells that are in the same region, but have different chemical properties, or wells that are in different regions, but have similar chemical properties. In such cases, the corresponding explanation will also be provided.

3. Results and Discussion

3.1. Numerical Example. In this section, a simple dataset is used in order to illustrate the proposed algorithm in detail. The dataset consists of 20 bivariate points presented in Table 2; the normalized data points are presented in Figure 3.

Using formulas (4), (5), and (6), we found the three initial points $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$ of the first cluster, which are represented by red in Figure 4. It can be seen from Figure 4 that the distance between these three points is really small in comparison with the distance between all points; therefore, condition (7) is satisfied and we can use these three points for extending the cluster.

Now, we use the points in the processing cluster to build up the cluster itself. For example, in Figure 5, starting from the green point, $\mathbf{v}_2$, using formula (3), we calculate the adaptive radius and determine the three new ε-neighbors, based on the circle formed. After that, the processing cluster will be extended by adding these three new points, and the point $\mathbf{v}_2$ will no longer be used to extend the cluster in the next steps. Using another point in the processing cluster, for example, the green point in Figure 6, we also calculate the adaptive radius and determine the new ε-neighbors, based on the circle formed.

Repeat the abovementioned process until the processing cluster cannot be extended more, that is, all points in the processing cluster have been used for the extending process and we cannot find any new points linked to them, as shown in Figure 7.

Figure 7 completely determines the first cluster; we can repeat the abovementioned process for the remainder of the dataset and obtain the final partition, as shown in Figure 8.

3.2. Experiments in Benchmark Datasets. Section 3.1 step-by-step illustrated the proposed algorithm. In this section, to test the partitioning performance of the proposed algorithm

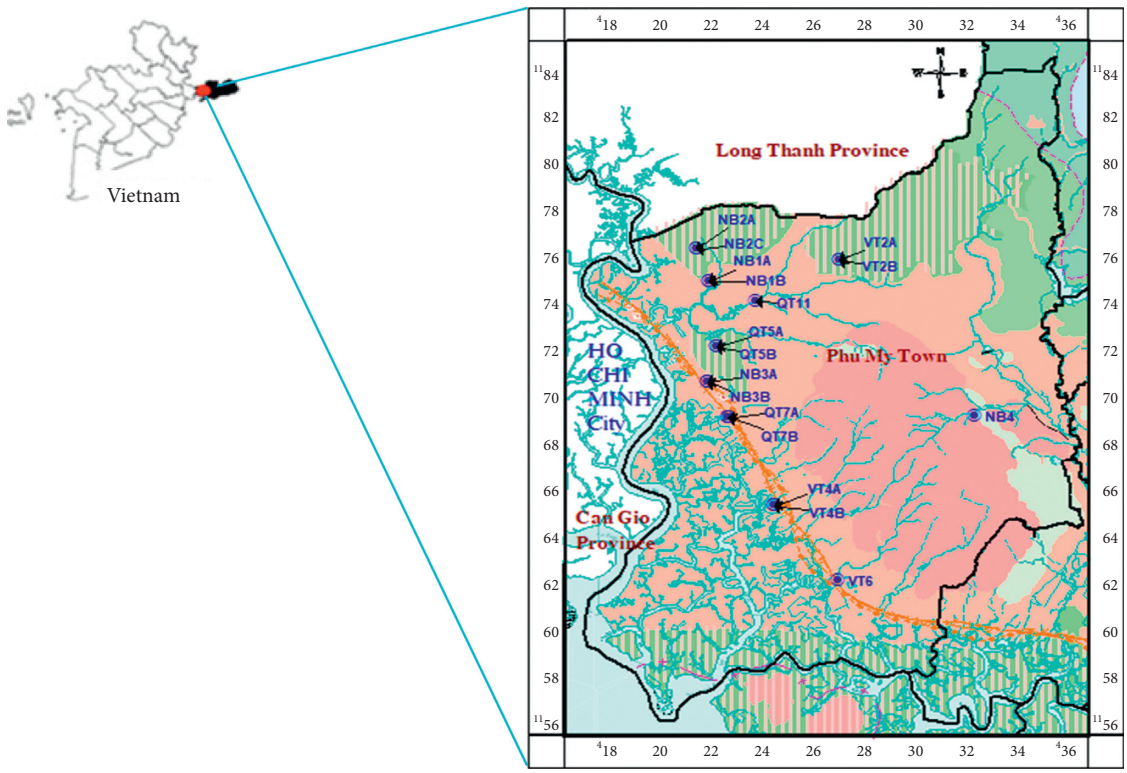


Figure 2: The position of wells.

Table 1: Concentration of chemical parameters (mg/l) collected at wells.

ID	Na^+	K^+	Ca^{2+}	Mg^{2+}	NH_4^+	Al^{3+}	HCO_3^-	Cl^-	SO_4^{2-}	NO_3^-	NO_2^-
NB3A	4.19	1.47	17.03	0.61	0.00	0.00	54.92	8.15	2.40	1.20	0.00
NB3B	6.56	3.85	2.00	0.49	0.00	0.00	6.10	17.73	2.40	0.64	0.00
QT5B	6.57	4.07	17.03	0.61	0.00	0.08	54.92	10.64	9.61	1.49	0.00
QT7B	192.73	9.00	22.04	19.46	2.31	8.51	0.00	375.77	81.65	1.15	0.00
NB2C	8.29	2.90	35.07	1.82	0.00	0.00	103.73	17.73	7.20	11.24	0.00
NB1B	4.14	2.32	1.60	0.24	0.00	0.00	12.20	7.80	2.40	0.41	0.00
VT4B	277.65	17.60	26.05	31.62	2.24	1.67	0.00	514.03	115.27	1.44	0.00
VT6	33.79	7.25	10.02	0.61	0.00	0.00	24.41	49.63	19.21	2.88	7.77
NB4	11.00	1.44	21.04	0.61	0.24	0.00	67.12	14.18	9.61	0.91	0.01
QT5A	10.43	1.22	1.40	0.36	0.00	0.00	6.10	16.66	2.40	7.91	0.00
QT7A	644.44	57.90	100.20	118.56	36.10	0.00	494.26	946.52	528.33	6.23	7.28
NB1A	5.00	5.38	2.00	0.49	0.00	0.00	18.31	8.86	2.40	0.59	0.00
NB2A	3.86	3.61	11.02	0.97	3.72	0.00	48.82	7.09	3.84	0.85	0.32
VT4A	82.73	5.76	54.11	21.89	3.65	0.00	85.43	223.34	31.22	0.81	0.00
QT11	4.89	1.65	1.00	0.24	0.00	0.00	12.20	7.80	0.96	1.32	0.00
VT2B	6.88	1.94	9.02	0.61	0.11	0.00	24.41	12.41	2.40	15.14	0.01
VT2A	4.33	2.14	5.61	1.09	0.04	0.00	18.31	8.15	3.36	11.34	0.01

and compare it with other methods, and the proposed algorithm is implemented on different datasets with different characteristics.

The tested datasets can be downloaded from (https://cs.joensuu.fi/sipu/datasets/), which include

(i) Spiral: a dataset with spiral-shaped clusters

(ii) Aggregation: a dataset with different cluster shapes

(iii) Compound: a compound dataset with different cluster shapes and densities

(iv) Gauss: a dataset simulated in [6] with three Gaussian clusters

The tested algorithms include

(i) ARC1: the proposed method with the adaptive radius defined according to formula (3).

(ii) ARC2: the proposed method with the adaptive radius defined according to formula (4).

(iii) k-mean, DBSCAN: two popular clustering algorithms. The k-means requires an initial number of

Table 2: Illustrated data.

Data	X_1	X_2	Data	X_1	X_2
1	42	72	11	41	58
2	44	71	12	41.5	59
3	46	73	13	42.5	59
4	47	72	14	43	60
5	49	71	15	45	61
6	51	71	16	45.5	61
7	52	70	17	47	61
8	54	69	18	48	61
9	55	68	19	49	61
10	57	67	20	50	60

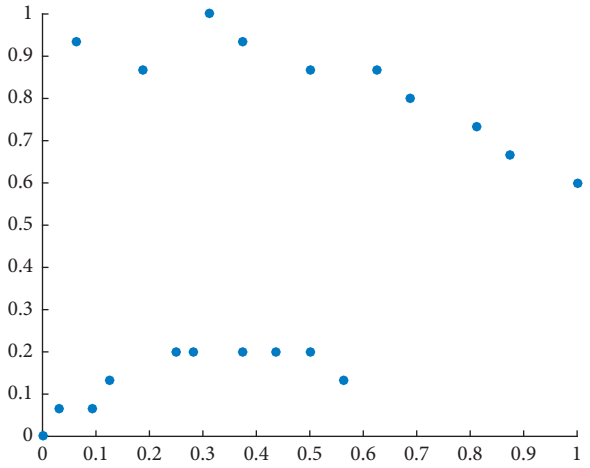


Figure 3: Illustrated data points.

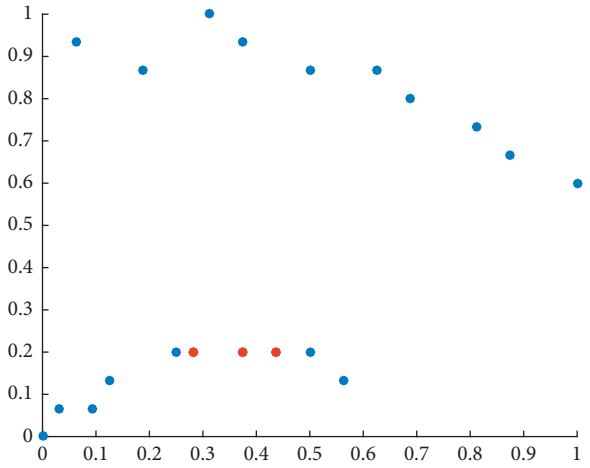


Figure 4: The initial prototypes.

clusters and results in the spherical clusters, while the DBSCAN is a density-based clustering algorithm that is suitable for clusters of arbitrary shapes.

(iv) SU: an automatic clustering algorithm recently presented by [19] for determining the number of clusters, automatically.

In this paper, the Adjusted Rand Index, ARI [29, 30], is employed to evaluate the performance of the five compared

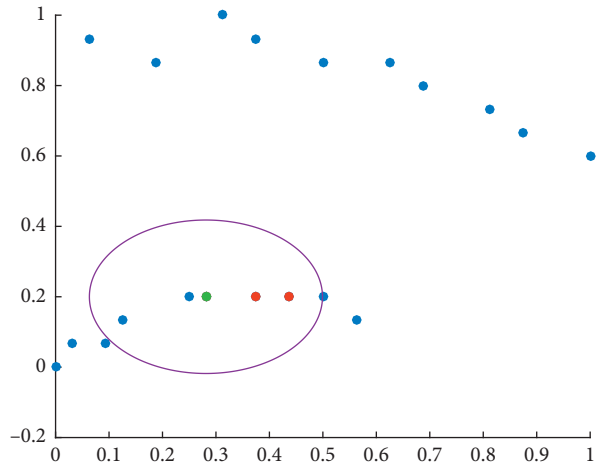


Figure 5: Determining the adaptive radius and new neighbors.

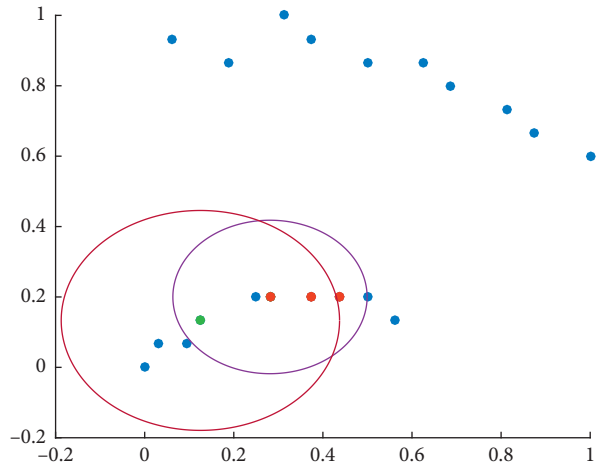


Figure 6: The processing cluster is extended at the green point.

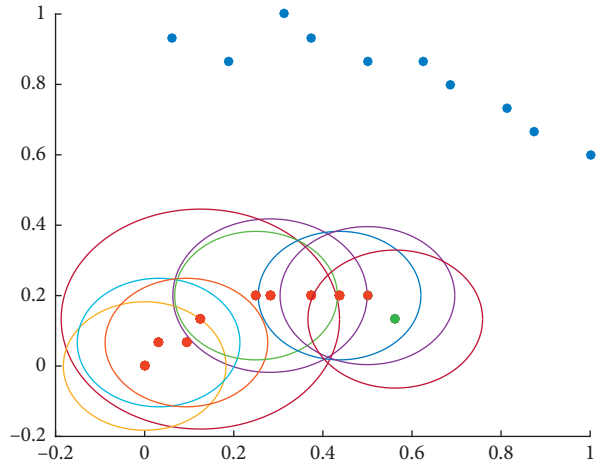


Figure 7: Determining the first clustering.

methods. ARI is an external measure that can make the comparison between the partition produced by a clustering algorithm (P) and the actual partition (Q), where "ground-truth" labeling is known. Particularly, given P and Q, the formulation of ARI is defined as follows:

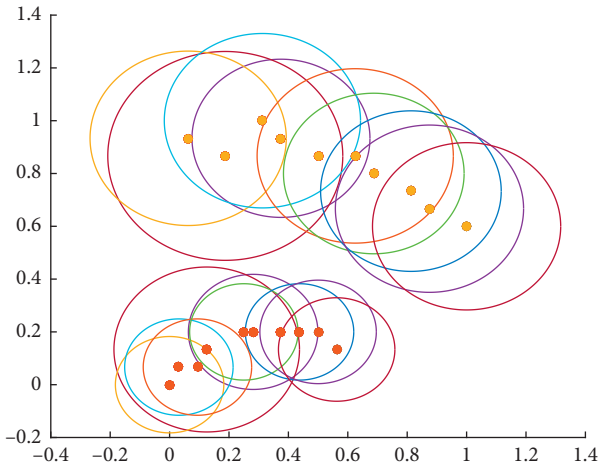
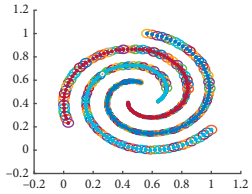
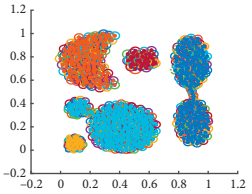
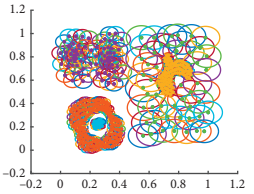
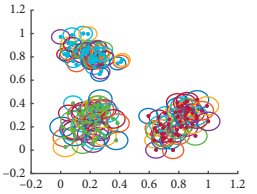
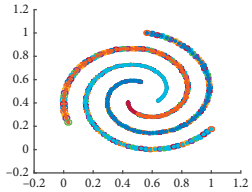
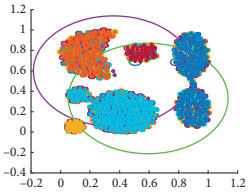
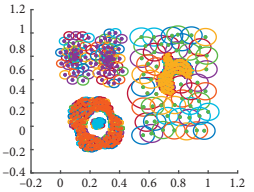
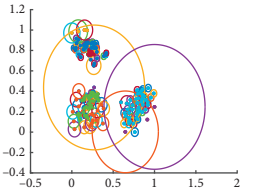
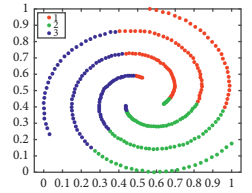
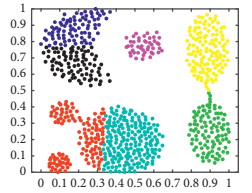
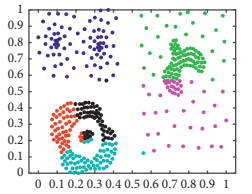
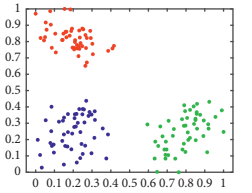
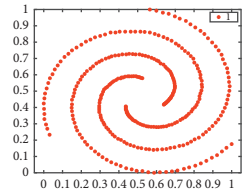
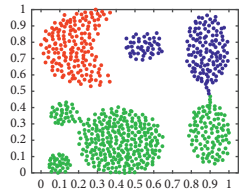
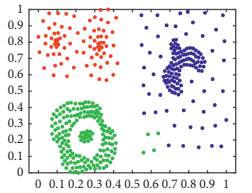
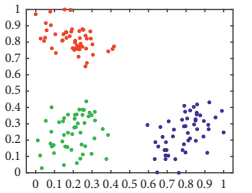
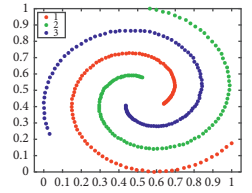
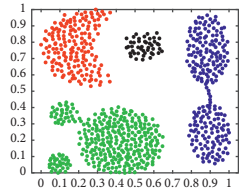
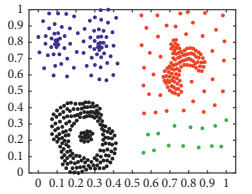
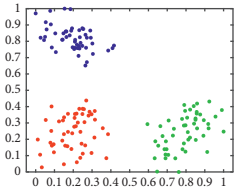


Figure 8: Determining the two clusters.

Table 3: The clusters formed the by the tested algorithms.

	Spiral	Aggregation	Compound	Gauss
ARC1				
ARC2				
k-means				
SU				
DBSCAN				

$$\mathrm{ARI} = \frac{a - (a+c)(a+b)/(a+b+c+d)}{((a+c)+(a+b))/2 - (a+c)(a+b)/(a+b+c+d)}, \quad (12)$$

where a is the number of pairs of elements in the same cluster in P and Q, b is the number of pairs of elements in the same cluster in P, but in different clusters in Q, c is the number of pairs of elements in a different cluster in P, but in the same cluster in Q, and d is the number of pairs of elements in a different cluster in both P and Q. The closer the ARI is to 1, the better the clustering result is (it can be seen from formula (12) that when P and Q are the same, $b = c = 0$ and ARI = 1).

Table 3 intuitively presents the clustering results of the five tested algorithms on the four used datasets.

Remarks:

(i) For the nonspherical clusters, the performance of the DBSCAN is better than that of SU and k-means algorithms. This result is reasonable because DBSCAN can easily group the data points into arbitrary shape clusters, based on the density and the connection rather than the distance between them. ARC2 algorithm, in general, is quite efficient in terms of ARI and outperforms the DBSCAN on two of the three datasets. Meanwhile, the ARC1 achieves the largest ARI values, which indicates the best performance in terms of clustering accuracy.

(ii) For the spherical or Gaussian clusters, most of the methods render good performance, in which ARC1, SU, and DBSCAN are the proper methods. The k-means algorithm also provides the best result, for $k = 3$; however, when k is randomly changed and does not satisfy $k = 3$, this method shows poor performance. Tables 3 and 4 also show that the ARC2 performs better than the k-means; however, it is not good enough for the Gaussian clusters.

(iii) In summary, it can be claimed that ARC1 is an effective algorithm. Specifically, the ARC1 can automatically determine the number of clusters and has notably larger ARI values or notably better clustering results for any given dataset.

3.3. Application for Underground Water Quality Assessment. In this section, we cluster the samples of groundwater quality parameters provided by the Department of Natural Resources and Environment of Ba Ria-Vung Tau Province. The study area and data used have been presented in Section 2. The clustering results in Figure 9 showed that the 17 monitoring wells are classified into 4 groups based on the water quality characteristics:

(i) Cluster 1: NB3A, QT5B, NB4

(ii) Cluster 2: NB3B, NB1B, NB1A, QT11

(iii) Cluster 3: QT7B, NB2C, VT4B, VT6, QT5A, NB2A, VT4A, VT2B, VT2A

(iv) Cluster 4: QT7A

Table 4: The ARI of the comparative methods on the four datasets.

	Spiral	Aggregation	Compound	Gauss
ARC1	**1.0000**	**0.8089**	**0.9438**	**1.0000**
ARC2	0.9253	0.8035	0.9438	0.7034
k-means	0.0924	0.5906	0.5890	0.6968
SU	0.0000	0.5638	0.7257	**1.0000**
DBSCAN	**1.0000**	0.7338	0.7568	**1.0000**

A comparison of some parameters among clusters is shown in Figure 10. We have the following remarks:

(i) Cluster 4 consists of only 1 well, QT7A, with very high parameter values. This result demonstrates that the water quality in this well is really bad compared to the remaining clusters. In addition, it can be seen from Table 1 and Figure 10(a) that QT7A has more salt ions (Mg^{2+}, Na^{+}, K^{+}, Ca^{2+}, HCO_3^{-}, NH_4^{+}, Cl^{-}, SO_4^{2-}, and Nitrite) compared to the remaining clusters. According to National Technical Regulation on Groundwater Quality of Vietnam, the permitted standard for Cl^{-} is 250 mg/l and for SO_4^{2-} is 400 mg/l. Therefore, the Cl^{-} and SO_4^{2-} values of QT7A exceed the permitted standards 3.78 and 1.3 times, respectively. This demonstrates that QT7A may be overaffected by saline intrusion because this well is located near the saline boundary. Additionally, it can be seen in Figure 9 that two wells QT7A and QT7B are located in the same region, but they belong to different clusters. Actually, they are both contaminated wells, but they have different depths, representing separate aquifers. As a result, QT7A exhibits a higher level of contamination than QT7B.

(ii) For the three remaining clusters, it can be seen from Figures 9 and 10(b) that Cluster 1 consists of three wells, with high HCO_3^{-} values. To our knowledge, the two wells, NB3A and QT5B, are located near My Xuan B1 industrial zone, and the well NB4 is located near Toc Tien landfill. As a result, those wells may be contaminated by the waste discharge process of the abovementioned industrial zone and landfill.

(iii) Cluster 2 consists of four wells with relatively good quality. In this cluster, most of the parameter values are lower than those of other clusters and are within safe ranges. It can be concluded that the wells of Cluster 2 are not affected by agricultural activities as well as saline intrusion.

(iv) Cluster 3 consists of eight wells with higher values of Mg^{2+}, Na^{+}, K^{+}, Ca^{2+}, Cl^{-}, and SO_4^{2-} compared to those of Cluster 1 and Cluster 2. Especially, Cl^{-} value exceeds the permitted standard at 2/8 wells. This indicates a number of wells in Cluster 3, which are located near the coast as well as salinity boundaries, are capable of being affected by salinity intrusion. In addition, as shown in Figure 10(b), in Cluster 3, the average value of NO_3^{-} is higher than that of Cluster 1 and Cluster 2. This demonstrates

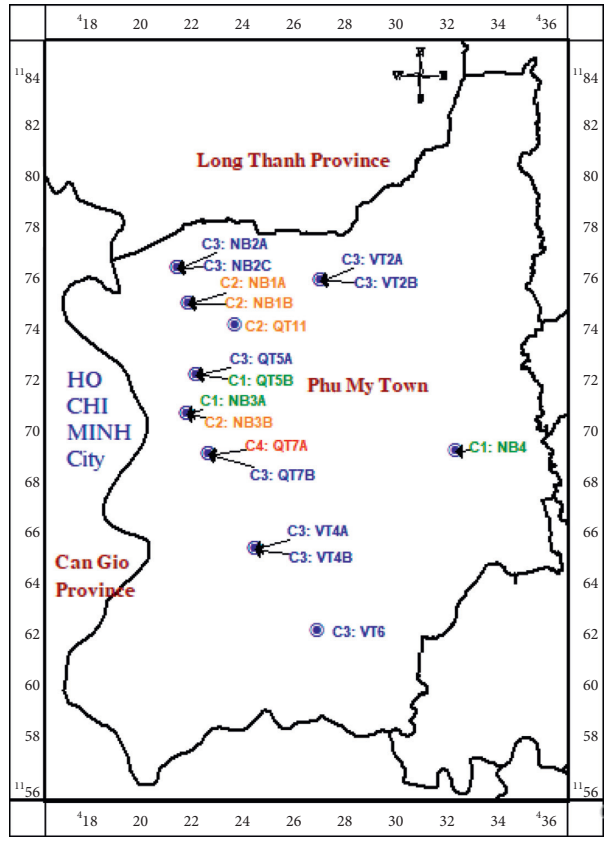


Figure 9: The clustering result for 17 wells.

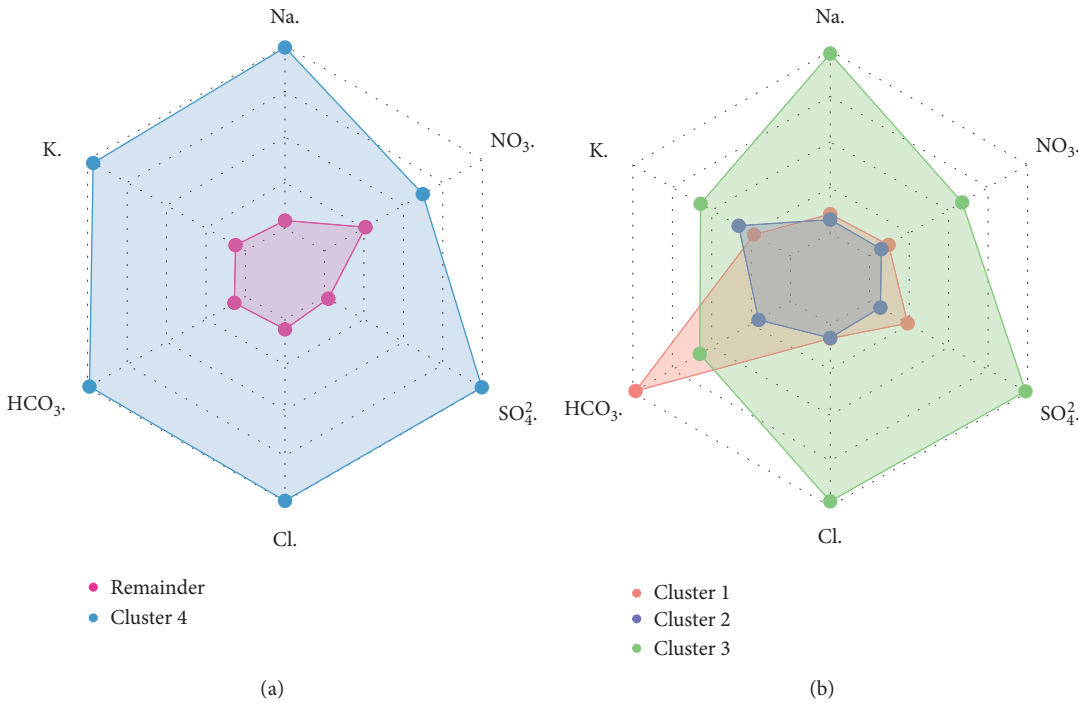


Figure 10: Comparing the parameters among clusters. (a) Cluster 4 and the remaining clusters. (b) Cluster 1, Cluster 2, and Cluster 3.

that agricultural activities taking place around the monitoring area served as large contributors to the underground water quality of this cluster. In particular, well NB2C, VT2B, and VT2A are located near the industrial planting area. Meanwhile, well VT6, which is located near the aquaculture area,

may be seriously affected by organic matter from the residual feed; therefore, the NO_3^- value reaches 7.77 times higher than the permitted standard.

4. Conclusion

Based on the definition of epsilon radius neighbors, this paper has proposed a new clustering algorithm that can automatically determine the number of clusters and can find clusters with different sizes, shapes, and densities. The radius or extending is adapted to the current-processing cluster and has good generalization ability. The proposed algorithm is tested on benchmark datasets and is then applied to underground water quality assessment in Phu My Town, Ba Ria-Vung Tau province, Vietnam. For the experiments with many datasets, the ARC1 algorithm exhibits a better performance than the other tested algorithms in terms of the Adjusted Rand index. The ARC2 algorithm performs better than the conventional clustering algorithms in the case of nonspherical clusters but worse in the case of spherical clusters. For the underground water quality assessment in Phu My Town, Ba Ria-Vung Tau province, Vietnam, the proposed algorithm indicated that there are four clusters of water quality that represent different source contributions.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the Vietnam National University Ho Chi Minh City (VNU-HCM) under Grant no. C2018-24-01.

References

[1] T. Vo Van and T. Pham-Gia, "Clustering probability distributions," *Journal of Applied Statistics*, vol. 37, no. 11, pp. 1891–1910, 2010.

[2] T. VoVan and T. NguyenTrang, "Similar coefficient for cluster of probability density functions," *Communications in Statistics—Theory and Methods*, vol. 47, no. 8, pp. 1792–1811, 2018.

[3] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[4] A. K. Jain, "Data clustering: 50 years beyond *K*-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[5] Z. Xie, R. Dong, Z. Deng et al., "A probabilistic approach to latent cluster analysis," in *Proceedings of the Twenty Third International Joint Conferences on Artificial Intelligence (IJCAI)*, Beijing, China, August 2013.

[6] T. VoVan and T. Nguyen Trang, "Similar coefficient of cluster for discrete elements," *Sankhya B*, vol. 80, no. 1, pp. 19–36, 2018.

[7] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, Oakland, CA, USA, June 1967.

[8] D. Arthur and S. Vassilvitskii, "*k*-means++: the advantages of careful seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035, Society for Industrial and Applied Mathematics, New Orleans, LA, USA, January 2007.

[9] G. Karypis, E.-H. Eui-Hong Han, and V. Kumar, "Chameleon: hierarchical clustering using dynamic modeling," *Computer*, vol. 32, no. 8, pp. 68–75, 1999.

[10] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: the fuzzy *c*-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2-3, pp. 191–203, 1984.

[11] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *Journal of Cybernetics*, vol. 4, no. 1, pp. 95–104, 1974.

[12] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841–847, 1991.

[13] J. C. Bezdek and N. R. Pal, "Cluster validation with generalized Dunn's indices," in *Proceedings of the 1995 Second New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems*, pp. 190–193, Dunedin, New Zealand, November 1995.

[14] R. Babuška, *Fuzzy Modeling for Control*, Springer Science & Business Media, Berlin, Germany, 2012.

[15] L. Mavridis, N. Nath, and J. B. Mitchell, "PFClust: a novel parameter free clustering algorithm," *BMC Bioinformatics*, vol. 14, no. 1, p. 213, 2013.

[16] L. E. Agustı, S. Salcedo-Sanz, S. Jiménez-Fernández, L. Carro-Calvo, J. Del Ser, and J. A. Portilla-Figueras, "A new grouping genetic algorithm for clustering problems," *Expert Systems with Applications*, vol. 39, no. 10, pp. 9695–9703, 2012.

[17] S. Das, A. Abraham, and A. Konar, "Automatic clustering using an improved differential evolution algorithm," *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, vol. 38, no. 1, pp. 218–237, 2008.

[18] T. Vo-Van, T. Nguyen-Thoi, T. Vo-Duy, V. Ho-Huu, and T. Nguyen-Trang, "Modified genetic algorithm-based clustering for probability density functions," *Journal of Statistical Computation and Simulation*, vol. 87, no. 10, pp. 1964–1979, 2017.

[19] W.-L. Hung and J.-H. Yang, "Automatic clustering algorithm for fuzzy data," *Journal of Applied Statistics*, vol. 42, no. 7, pp. 1503–1518, 2015.

[20] M. Ester, H.-P. Kriegel, J. Sander et al., "A density-based algorithm for discovering clusters in large spatial databases with noise," *KDD*, vol. 96, no. 34, pp. 226–231, 1996.

[21] C. Dobbins and R. Rawassizadeh, "Towards clustering of mobile and smartwatch accelerometer data for physical activity recognition," *Informatics*, vol. 5, no. 2, p. 29, 2018.

[22] G. F. Tzortzis and A. C. Likas, "The global kernel *k*-means algorithm for clustering in feature space," *IEEE Transactions on Neural Networks*, vol. 20, no. 7, pp. 1181–1194, 2009.

[23] I. S. Dhillon, Y. Guan, and B. Kulis, "Kernel *k*-means: spectral clustering and normalized cuts," in *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD'04*, pp. 551–556, ACM, New York, NY, USA, August 2004.

[24] M. O. Mavukkandy, S. Karmakar, and P. S. Harikumar, "Assessment and rationalization of water quality monitoring network: a multivariate statistical approach to the Kabbini

River (India)," *Environmental Science and Pollution Research*, vol. 21, no. 17, pp. 10045–10066, 2014.

[25] U. N. Kura, F. M. Ramli, N. W. Sulaiman, S. Ibrahim, A. Aris, and A. Mustapha, "Evaluation of factors influencing the groundwater chemistry in a small tropical island of Malaysia," *International Journal of Environmental Research and Public Health*, vol. 10, no. 5, pp. 1861–1881, 2013.

[26] S. Shrestha and F. Kazama, "Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji river basin, Japan," *Environmental Modelling & Software*, vol. 22, no. 4, pp. 464–475, 2007.

[27] M. Varol and B. Ş r, "Assessment of nutrient and heavy metal contamination in surface water and sediments of the upper Tigris River, Turkey," *Catena*, vol. 92, pp. 1–10, 2012.

[28] Q. Yang, J. Zhang, Y. Wang, Y. Fang, and J. Martin, "Multivariate statistical analysis of hydrochemical data for shallow ground water quality factor identification in a coastal aquifer," *Polish Journal of Environmental Studies*, vol. 24, 2015.

[29] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.

[30] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.