A COLLABORATIVE POSSIBILISTIC FUZZY C-MEANS CLUSTERING APPROACH FOR MULTI-DIMENSIONAL DATA ANALYSIS

Viet Duc Do[1,3], *Dinh Sinh Mai*[2,*], *Long Thanh Ngo*[3]

Abstract

The rapid development of data acquisition technologies has led to an explosion of data sources. Many traditional data mining techniques and methods have become outdated and are no longer suitable for solving large, high-dimensional data problems. The paper proposes improving the collaborative possibilistic fuzzy clustering algorithm for multi-dimensional data analysis using random projection feature reduction. The random projection feature reduction technique allows for the preservation of relative distances after dimensional reduction, which can help reduce computational complexity while still ensuring the accuracy of the proposed algorithm compared to the algorithm before dimensionality reduction. The proposed algorithm implemented on the collaborative clustering model can help share information about cluster structure at data sites during computation, allowing problems to be performed where data is located on different computers in a network. Experiments performed on two multidimensional datasets downloaded from the UCI Machine Learning Repository library and remote sensing image data show that the proposed method yields significantly better results than some previously proposed methods. These experimental results demonstrate the potential of developing collaborative clustering models, combined with dimensionality reduction techniques, to tackle high-dimensional and distributed large data problems.

Index terms

Multi-dimensional data; possibilistic fuzzy c-means clustering; collaborative clustering; random projection; dimensionality reduction.

[1]National Defense Academy
[2]Institute of Techniques for Special Engineering, Le Quy Don Technical University
[3]Institute of Information and Communication Technology, Le Quy Don Technical University
[*]Corresponding author, email: maidinhsinh@lqdtu.edu.vn DOI: 10.56651/lqdtu.jst.v13.n02.924.ict

1. Introduction

Today, computer scientists face the challenge of handling large amounts of data [1]. Data mining techniques, particularly clustering, are widely recognized as reliable methods for discovering knowledge from this vast amount of data. Clustering is one of the primary data mining methods which aims at partitioning a specific dataset into a finite number of groups or clusters. Data samples in the same group often have more similarities than those in different groups. Although originally derived from data mining, clustering is widely used to solve various problems in other fields, such as bio-informatics, machine learning, networking, and pattern recognition [2].

When dealing with large data, certain key factors must be considered. Firstly, volume is crucial, as large data involves massive amounts of information. Although there is no fixed threshold for what can be considered "proposelarge" data, generally, it refers to data with significant volumes. Secondly, velocity is also important since data generated in large systems requires systems that can handle and respond to large amounts of incoming data quickly [3]. A clustering algorithm needs to analyze the relationships between data elements so quickly that the incoming data does not invalidate the analysis results before they are used. Variety, large data systems must handle a wide range of different incoming data types, including structured data (e.g., CSV data), semi-structured (e.g., HTML content), and unstructured data (e.g., videos and images) [4].

Traditional clustering methods based on Euclidean distances often focus on information from spectrum bands and allow only one pattern to belong to a single cluster, which will not describe all the data characteristics and lead to low accuracy of the clustering results. The clustering technique is based on fuzzy sets, which allow each data pattern to belong to many different clusters through the membership function value, which can handle data patterns whose boundaries are unclear and uncertain belonging to a specific cluster [5].

High-dimensional data often have high nonlinearity and overlap. Some recent studies have also shown that probabilistic fuzzy clustering has many advantages due to the combination of fuzzy and probabilistic information to describe data, especially noisy and high-dimensional data [6], [7]. Yu *et al.* proposed a suppressed possibilistic fuzzy c-means clustering algorithm based on shadow sets for noisy data with imbalanced sizes [8]. The method solves the clustering problem on imbalanced datasets. Wu *et al.* proposes a series of generalized multiplicative fuzzy possibilistic product partition clustering algorithms to enhance the ability to remove noise [9]. Farooq *et al.* present a fast and robust FCM (FRFCM) clustering algorithm that performs fast and robustly to noise for grayscale and color images [10]. A concept of uncertainty measures for probabilistic hesitant fuzzy information by comprehensively considering their fuzziness and hesitancy, and Fang *et al.* proposed some novel entropy and cross-entropy measures for them [11].

Large data, high-dimensional analysis problems often have difficulty processing centralized data due to the limitations of computer hardware. Data dimensionality

reduction, data compression,.... and/or parallel or distributed computing are common solutions. This underscores the necessity for a new approach. One such approach is using collaborative clustering techniques that harness the collective power of multiple computers in a network model. Additionally, data groups of the same type often exhibit similar characteristics, even in different datasets. Therefore, an approach that facilitates sharing features among data groups in different datasets can significantly enhance the quality of data clusters [12].

Collaborative data clustering is a tool to find structural similarities and similarities between data patterns located in many distinct regions, based on the objective function expansion and fuzzy clustering approach of the FCM algorithm [13] proposed by Professor Pedrycz. Pedrycz introduced collaborative fuzzy clustering as a tool to find structures and similarities between distinct datasets. Where details in datasets cannot be exchanged, only structural information can be exchanged [14]. At the same time, the fuzzy clustering in this dataset impacts the clustering in other datasets [15].

There are two features of collaborative fuzzy clustering. One is that detailed information in datasets cannot be exchanged, and only structural information can be exchanged [16]. The second is to consider whether fuzzy clustering in this dataset impacts clustering in other datasets [17]. Is the information about cluster structure in each dataset useful in clustering the remaining datasets? However, the FCC algorithm does not use additional information in the clustering process, which can help improve the accuracy of the clustering process [18].

From the issues pointed out above, it can be seen that the clustering problem is still difficult and needs to be researched and developed, especially for large data and high-dimensional data. In addition, dimensionality reduction is an extremely important step in optimizing the calculation for multi-dimensional or high-dimensional data. Although many dimensionality reduction techniques exist, such as principal component analysis, uniform approximation, and projection,... However, dimensionality reduction methods based on random projection have many advantages due to preserving the relative distance between data samples. In addition, solving large data problems is often difficult when working centrally, while computer networks have become very popular. Large data sources are also stored in many places (distributed). Therefore, developing methods that allow working with large data, many dimensions, and on many different computers will have many advantages and be in line with the development trend of data science.

The main contributions of the paper include proposing an improvement of the CPFCM algorithm based on random projection feature reduction technique (CPFCM-FR) and experimental setup for a multi-dimensional dataset placed on two connected computers. Experimental data were obtained from the UCI machine learning library [19], and satellite image data were downloaded from [20]. To evaluate the effectiveness of the proposed model, in this paper, we compare the experimental results of the collaborative possibilistic fuzzy clustering algorithm with algorithms before improvements, showing that the proposed method is more effective.

The paper is organized into five sections: Section 1 is the introduction; Section 2 introduces some related knowledge; Section 3 is the proposed method; Section 4 presents some experiments; and Section 5 gives the conclusions.

2. Background

2.1. Possibilistic fuzzy c-means clustering

The possibilistic c-means algorithm (PCM) was proposed by Krishnapuram and Keller and was introduced to avoid the sensitivity of the FCM algorithm. Instead of using the fuzzy MFs such as FCM, PCM uses possibilistic MFs to represent typicality by τ_{ik}, the typicality matrix as $\mathrm{T} = [\tau_{ik}]_{cxn}$.

The PCM model is the constrained optimization problem:

$$\min \{J_\eta(T, V; X, \gamma) = \sum_{i=1}^{c}\sum_{k=1}^{n} \tau_{ik}^{\eta} d_{ik}^2 + \sum_{i=1}^{c} \gamma_i \sum_{k=1}^{n} (1-\tau_{ik})^{\eta}\} \tag{1}$$

where $\mathrm{T} = [\tau_{\mathrm{ik}}]_{\mathrm{cxn}}$ is a possibilistic MF, $\mathrm{V} = (\mathrm{v}_1, \mathrm{v}_2, ..., \mathrm{v}_\mathrm{c})$ is a vector of cluster centers, $\gamma_i > 0$ is a user-defined constant. With the following constraints:

$$\eta > 1; 0 \leq \tau_{ik} \leq 1; \sum_{k=1}^{n} \tau_{ik} = 1; 1 \leq i \leq c; 1 \leq k \leq n \tag{2}$$

Krishnapuram and Keller also suggest using the results of the FCM algorithm as a good way to initialize the PCM algorithm, and the parameter γ_i should be calculated according to the following equation:

$$\gamma_i = K \sum_{k=1}^{n} \mu_{ik}^{\eta} d_{ik}^2 / \sum_{k=1}^{n} \mu_{ik}^{\eta} \tag{3}$$

where μ_{ik} is the fuzzy membership from the results of the FCM algorithm, K is a user-defined constant (usually selected by 1).

FCM and PCM are the most popular approaches to fuzzy clustering and possibilistic clustering. However, they suffer from drawbacks such as high noise sensitivity and difficulty working with overlapping data. The PFCM algorithm [21] is a hybrid algorithm between FCM and PCM, inheriting the advantages of both FCM and PCM. The PFCM algorithm has two types of MFs, including the fuzzy MF in the FCM algorithm and the possibilistic MF in the PCM algorithm.

PFCM model is the constrained optimization problem:

$$J_{m,\eta}(U, T, V, X, \gamma) = \sum_{i=1}^{c}\sum_{k=1}^{n} (a\mu_{ik}^{m} + b\tau_{ik}^{\eta}) d_{ik}^2 + \sum_{i=1}^{c} \gamma_i \sum_{k=1}^{n} (1-\tau_{ik})^{\eta} \tag{4}$$

where $\mathrm{X} = \{\mathrm{x_k}, \mathrm{x_k} \in \mathrm{R^M}, \mathrm{k} = 1,...,\mathrm{n}\}$ and $\mathrm{U} = [\mu_{\mathrm{ik}}]_{\mathrm{cxn}}$ is a fuzzy partition matrix, which contains the fuzzy membership degree, $\mathrm{T} = [\tau_{\mathrm{ik}}]_{\mathrm{cxn}}$ is a typicality partition

matrix, which contains the possibilistic membership degree, $\mathrm{V} = (\mathrm{v}_1, \mathrm{v}_2, ..., \mathrm{v}_\mathrm{c})$ is a vector of cluster centers, m is the weighting exponent for fuzzy partition matrix and η is the weighting exponent for the typicality partition matrix. $\gamma_i > 0$ are constants given by the user.

Subject to the constraints:

$$\begin{aligned} &m, \eta > 1; a, b > 0; 0 \leq \mu_{ik}, \tau_{ik} \leq 1; \\ &\sum_{i=1}^{c} \mu_{ik} = 1; \sum_{k=1}^{n} \tau_{ik} = 1; 1 \leq i \leq c; 1 \leq k \leq n \end{aligned} \tag{5}$$

The objective function $J_{m,\eta}(U, T, V, X)$ reaches the smallest value with the constraints (5) if and only if:

$$v_i = \left(\sum_{k=1}^{n} (a\mu_{ik}^m + b\tau_{ik}^\eta) x_i / \sum_{k=1}^{n} (a\mu_{ik}^m + b\tau_{ik}^\eta) \right) \tag{6}$$

$$\mu_{ik} = 1 / \sum_{j=1}^{c} \left(d_{ik}^2 / d_{jk}^2 \right)^{2/(m-1)} \tag{7}$$

$$\tau_{ik} = 1 / \left(1 + \left(b d_{ik}^2 / \gamma_i \right)^{1/(\eta-1)} \right) \tag{8}$$

where, the constraints (5), Equations (6) and (7) achieved in the same way as FCM algorithm, Equation (8) achieved in the same way as PCM algorithm.

2.2. Collaborative fuzzy clustering

The model of structural information exchange or collaboration between datasets is depicted in Figure 1. Given the dataset $X = x_1, x_2, ..., x_N$, there are P sub-datasets (data site) including $D[1], D[2], ..., D[P]$, where each sub-dataset contains $N[1], N[2], ..., N[P]$ data samples in the same attribute space X. In each data site $D[ii]$, the data is divided into C clusters. The clustering results in each dataset affect the clustering in the remaining regions [22].

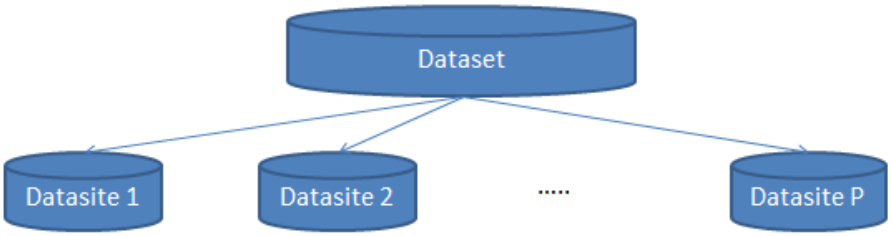

Fig. 1. The model of collaborative learning between data sites.

In this process, the data sites do not directly exchange detailed data but only share structural information, is the cluster center vector $v[ii]$ (with $ii = 1, .., P$). When

collaboratively clustering results, looking at the overall level of data sites will be better than clustering results based only on local data at each data site.

The collaborative fuzzy clustering problem whose objective function needs to be optimized is:

$$Q_{[ii]} = \sum_{k=1}^{N[ii]} \sum_{i=1}^{C} u_{ik}^2[ii] d_{ik}^2 + \beta \sum_{jj=1}^{P} \sum_{k=1}^{N[ii]} \sum_{i=1}^{C} (u_{ik} - \tilde{u}_{ik}[ii/jj])^2 d_{ik}^2 \tag{9}$$

The first part of the objective function is similar to the FCM algorithmic objective function. The second part of the objective function shows that the optimization in the collaborative process decreases the difference between the partitioning matrices.

In the above objective function, $u_{ik}[ii]$ is the matrix that partitions the object k into cluster i in data site ii. $\tilde{u}_{ik}[ii/jj]$ is called the collaborative partitioning matrix of data site jj onto data site ii (with $ii, jj = 1, .., P$), and is calculated by the Equation (9):

$$\tilde{u}_{ik}[ii/jj] = \frac{1}{\sum_{j=1}^{c} \left(\frac{x_k[ii] - v_i[jj]}{x_k[ii] - v_j[jj]} \right)^2} \tag{10}$$

Parameter $\beta[ii/jj]$ represents the degree of cooperation between data sites. The larger the value, the higher the degree of cooperation, and the value $\beta[ii/jj] = 0$ represents between datasets without cooperation. d_{ik} is the distance from the k^{th} object to the i^{th} cluster center in the same data site.

Using the Lagrange method to optimize the above objective function, the equation for calculating the partition matrix and cluster center is as follows:

$$u_{rs}[ii] = \frac{1}{\sum_{j=1}^{c} d_{rs}^2 / d_{js}^2} \left[1 - \sum_{j=1}^{c} \frac{\beta[ii/jj] \sum_{jj=1, jj \neq ii}^{P} \tilde{u}_{js}[ii/jj]}{(1 + \beta[ii/jj](P-1))} \right] + \frac{\beta[ii/jj] \sum_{jj=1, jj \neq ii}^{P} \tilde{u}_{rs}[ii/jj]}{(1 + \beta[ii/jj](P-1))} \tag{11}$$

$$v_{rt}[ii] = \frac{\sum_{k=1}^{N[ii]} u_{rk}^2[ii] x_{kt} + \beta[ii/jj] \sum_{jj=1, jj \neq ii}^{P} \sum_{k=1}^{N[ii]} (u_{rk}[ii] - \tilde{u}_{rk}[ii/jj])^2 x_{kt}}{\sum_{k=1}^{N[ii]} u_{rk}^2[ii] + \beta[ii/jj] \sum_{jj=1, jj \neq ii}^{P} \sum_{k=1}^{N[ii]} (u_{rk}[ii] - \tilde{u}_{rk}[ii/jj])^2} \tag{12}$$

The partition matrix of the objective function must satisfy the constraint that the total membership of an element in the clusters in the same dataset is equal to 1 as follows:

$$\mathrm{U} = \{ \mathrm{u_{ik}} \in [0, 1], \sum_{\mathrm{i}=1}^{\mathrm{c}} \mathrm{u_{ik}[ii]} = 1, \forall \mathrm{k}; 0 < \sum_{\mathrm{k}=1}^{\mathrm{N[ii]}} \mathrm{u_{ik}[ii]} < \mathrm{N[ii]}, \forall \mathrm{i} \} \tag{13}$$

2.3. Random projection feature reduction

Formally, a dimensionality reduction technique can be defined as follows. Given a dataset of d-dimensions, we find a function such that $f : R^d \rightarrow R^k$, with $k < d$. The function f projects the original d-dimensional data to k-dimensional data with the constraint $k < d$. Most dimensionality techniques share two common properties [23], but an ensemble of two properties in a dimensionality reduction method would produce a state-of-the-art technique for reducing high-dimensional data.

The dimensionality reduction used in this paper is a random projection. The main idea behind this random projection is from a popular lemma named Johnson-Lindenstrauss (JL) lemma [24].

The lemma states: given a finite set $X \subset R^d$ of size $|X| = K$, there exists a linear map $f : R^d \rightarrow R^k$ with $k = O(\varepsilon^{-2} * \log K)$ such that:

$(1 - \varepsilon)\|x - y\|_2 \leq \|f(x - y)\|_2 \leq (1 + \varepsilon)\|x - y\|_2$, for all $x, y \in R^d$.

This means that when we have a set of high-dimensionality, rather than points in Euclidean space, it can be linearly embedded into a space of lower dimensions. The projection also preserves the distance between points. It does not specify a method for identifying the value of k; instead, it merely states that such a dimension does exist. The reduced data are obtained by multiplying the vector of the original data with a random matrix: $X_d \times R$ to produce a new vector Y with the new reduced dimensions.

3. Proposal methods

This section presents a collaborative possibilistic fuzzy c-means clustering algorithm based on random projection feature reduction (CPFCM-FR). Given a set of data samples in *d*-dimensions X_{nd}, there exists a linear transformation down to *k*-dimensions in which the Euclidean relative distances between data samples in the new space are approximately constant using the projection multiplication of the input data matrix with a random *k*-dimensional matrix R_{dk} to get *k*-dimensional output data $Y_{nk} = X_{nd} * R_{dk}$, n is the number of data samples.

According to the famous work of Johnson-Lindenstrauss lemma on random projection [23], high-dimensional data into some lower dimension according to the following equation:

$$k = \varepsilon^{-2} \log(d) \tag{14}$$

in which, *k* is the new dimension, *d* is the original dimension, and ε is a constant. The lemma shows that the distance between samples is negligibly changed within $1 \pm \varepsilon$. To compute the matrix *R*, in the paper [24] shows an independent random distribution of $R = r_{ij}, i = 1, ..., d; j = 1, ..., k$ as follows:

$$r_{ij} = \begin{cases} -1 & p = 1/2 \\ 1 & p = 1/2 \end{cases} \quad ; \quad r_{ij} = \sqrt{3} \ * \begin{cases} -1 & p = 1/6 \\ 0 & p = 2/3 \\ 1 & p = 1/6 \end{cases} \tag{15}$$

This formula is derived based on the Johnson-Lindenstrauss lemma to preserve the distance between data samples after changing the number of data dimensions [24], with *p* being the probability distribution for the corresponding data sample. Once the dimensionality reduction is done, the dataset in the new dimension will be used for the clustering step.

The objective function of the CPFCM algorithm has two parts, including a component representing the fuzzy membership function and a component representing the collaboration between data sites. It can help to describe relationships between data sites and between data samples within each data site more closely than using only fuzzy membership function information.

From the above idea, the proposed algorithm aims to eliminate redundant (unnecessary) attributes and increase the rigor of the objective function. Thus, it can help improve the accuracy of the data clustering results. The new objective function of the proposed algorithm on each data site is added with possibilistic values τ_{ik} with fuzzy parameter m and possibilistic parameter η, specifically as follows:

$$\begin{aligned} Q(U,V,P) = & \sum_{k=1}^{N[ii]} \sum_{i=1}^{C} (a u_{ik}^{m} + b\tau_{ik}^{\eta}) x_i[ii] d_{ik}^2 \\ & + \beta[ii/jj] \sum_{jj=1}^{P} \sum_{k=1}^{N[ii]} \sum_{i=1}^{C} (u_{ik} - \tilde{u}_{ik}[ii/jj])^m d_{ik}^2 \end{aligned} \tag{16}$$

in which, a, b are coefficients representing the weights of the fuzzy and possibilistic membership functions. C is the number of clusters of the dataset, P is the number of data sites and $ii, jj = 1, ..., P$. $N[ii]$ is the number of data samples of data site ii. $\beta[ii/jj]$ is the collaboration coefficient between two data sites ii and jj. $\tilde{u}_{ik}[ii/jj]$ is the collaborative partitioning matrix of data site jj onto data site ii. u_{ik} and τ_{ik} are the fuzzy and possibilistic membership function values of the k^{th} data sample (data site ii) for the i^{th} cluster, $k = 1, ...N[ii]$, $i = 1, ..., C$.

Before each phase of collaboration, we calculate $\tilde{v}$ as the new prototypes from the prototypes communicated with all remaining data sites; the number of items $\tilde{v}$ is the same as the number of clusters of data site ii by using the PFCM algorithm.

The factor β is the arithmetic average of $\beta[ii/jj]$, the interaction level $\beta[ii/jj]$ between two data sites ii and jj, at a given collaboration stage, can be defined as:

$$\beta = \frac{\sum_{jj=1, jj \neq ii}^{P} \beta[ii/jj]}{P-1} \tag{17}$$

The initial dataset is divided into data sites, and clustering data sites is carried out independently. The resulting cluster prototypes after each iteration are shared from one data site to another. The proposed model is shown in Figure 2. The values $v_i = v_{ij}$ are shared on each data site as supporting information for the clustering process.

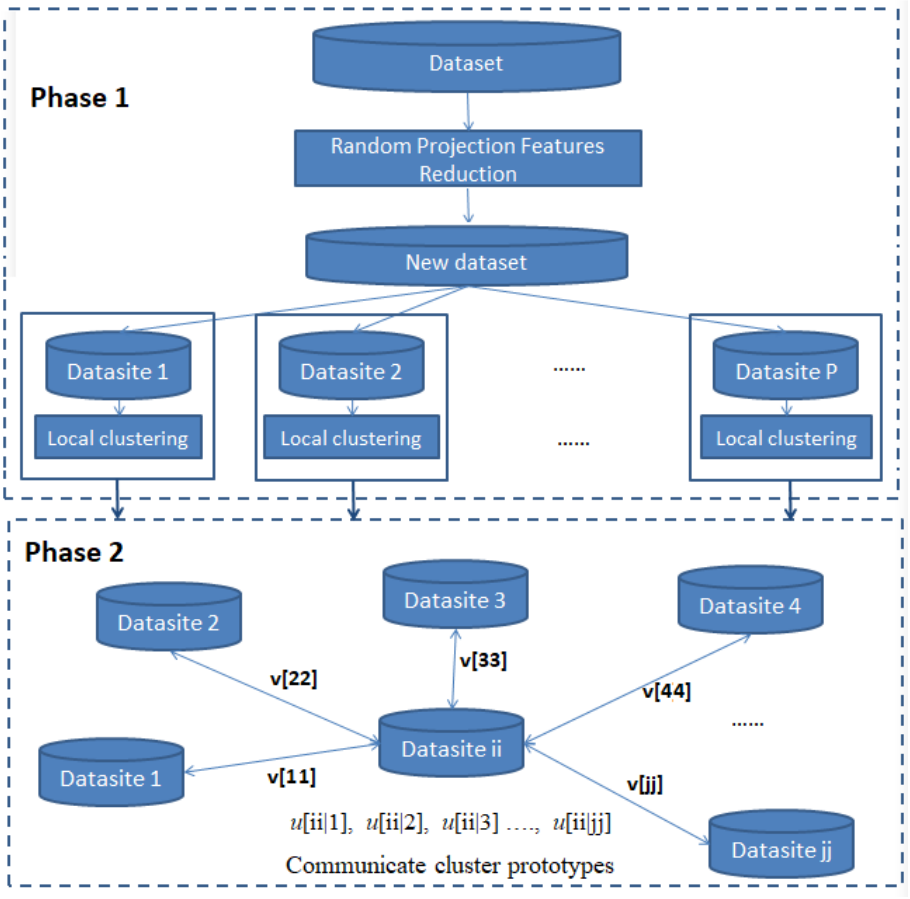

Fig. 2. The model of collaborative possibilistic fuzzy c-means clustering.

In the objective function (16): $u_{ik}[ii]$ is the fuzzy partition matrix of the data site ii^{th}. $\beta[ii/jj]$ is the parameter that represents the degree of cooperation of the data site jj with the data site ii and has the value domain $[0, 1]$. The value $\beta[ii/jj] = 0$ shows that the data site ii and data site jj without cooperation, $\beta[ii/jj]$ is calculated according to the following equation:

$$\beta[ii/jj] = min[1, \frac{J[ii]}{\tilde{J}[ii/jj]}] \tag{18}$$

with $\tilde{J}[ii/jj] = \sum_{k=1}^{N[ii]} \sum_{j=1}^{C} \tilde{u}_{ik}^2[ii/jj](x_k - v_i[jj])^2$, $\tilde{u}_{ik}[ii/jj]$ is the cooperative partitioning matrix of data site jj on data site ii and is calculated by the equation:

$$\tilde{u}_{ik}[ii/jj] = 1 \Big/ \sum_{j=1}^{C} \left(\frac{(x_k[ii] - v_i[jj])}{(x_k[ii] - v_j[jj])} \right)^2 \tag{19}$$

where, d_{ik}^2 is the distance between the k^{th} data sample in the data site $D[ii]$ and the

centroid v_{ij} of the i^{th} cluster in this same data site: $d_{ik}^2 = \sum_{j=1}^{M} (x_{kj} - v_{ij})^2$ and $d_{ik}[jj]$ is the distance between the k^{th} data sample in the data site $D[ii]$ and the i^{th} cluster center $v_{ij}[jj]$ in the data site $D[jj]$: $d_{ik}^2[jj] = \sum_{j=1}^{M} (x_{kj} - v_{ij}[jj])^2$. The cluster centroid at data sites is calculated as follows:

$$\tilde{v}[jj] = \frac{\sum_{j=1}^{C} v_j}{C} \tag{20}$$

We confirm the use of the technique of Lagrange multipliers for the objective function at each site. For any data sample $k, k = 1, 2, ..., N[ii]$, we reformulate the objective function to be in the form:

$$\begin{aligned} Q_{[ii]}(U, V, \lambda) &= \sum_{k=1}^{N[ii]} \sum_{i=1}^{C} (au_{ik}^m + b\tau_{ik}^\eta)(v_i[ii] - x_k[ii])^2 \\ &+\beta \sum_{k=1}^{N[ii]} \sum_{i=1}^{C} (u_{ik}[ii] - \tilde{u}_{ik}[ii])^m (v_i[ii] - \tilde{v}_i[ii])^2 + \sum_{k=1}^{N[ii]} \lambda_k \sum_{i=1}^{C} (1 - u_{ik}[ii])^m \end{aligned} \tag{21}$$

Equation (21) is the objective function on the data sites used to compute the convergence on the data sites. After computing the derivative with respect to the elements of the partition matrix with $\sum_{i=1}^{C} u_{ik}[ii] = 1,\ i = 1, 2, ..., C; k = 1, 2, ..., N[ii]$. We will get each data site's membership function matrix.

$$u_{ik}[ii] = \frac{\left[a(v_i[ii] - x_k[ii])^2 + \beta(v_i[ii] - \tilde{v}_i\ [ii])^2\right]^{-1}}{\sum_{j=1}^{c[ii]} \left[\frac{1}{a(v_j[ii] - x_k[ii])^2 + \beta(v_j[ii] - \tilde{v}_j\ [ii])^2}\right]^{1/(m-1)}} \tag{22}$$

Similarly to the above, to calculate the centroid of clusters, we calculate the derivative of the objective function (21) with the parameter being the centroid v_i.

$$v_i[ii] = \frac{\sum_{k=1}^{N[ii]} (au_{ik}^m[ii] + b\tau_{ik}^\eta[ii])x_k[ii] + \beta(u_{ik}[ii] - \tilde{u}_{ik}[ii])^m \tilde{v}_i\ [ii])}{\sum_{k=1}^{N[ii]} (au_{ik}^m[ii] + b\tau_{ik}^\eta[ii] + \beta(u_{ik}[ii] - \tilde{u}_{ik}[ii])^m)} \tag{23}$$

To calculate the typicality partition matrix, which contains the possibilistic membership degree, with $i = 1, 2, ..., C; k = 1, 2, ..., N[ii]; ii = 1, ..., P$, use the following equation:

$$\tau_{ik} = 1/\left(1 + (b(v_i[ii] - x_k[ii])^2/\gamma_i)^{1/(\eta-1)}\right) \tag{24}$$

In Algorithm 1, phase 1 performs local clustering on data sites as an initial step to get results used for Phase 2. Phase 2 implements collaboration between data sites.

Algorithm 1: The CPFCM-FR algorithm

Input: Dataset X, ε, and initialize the parameters $m = 2$ and $\eta = 2$, the maximum number of iterations $T_{max} = 100$, the number of data sites P, the number of elements in each data site ii is $N[ii]$, the number of clusters in each data site ii is $c[ii]$, the number of attributes of the data element is n, the data item in each data site $X[ii]$, $a = b = 1$.
Output: Clustering results.
Begin
Phase 1: Features reduction
    1.1 Random projection feature reduction
    1.2 Put *P* data sites on different computers
    1.3 Locally clustering: Run PFCM algorithm for each data site
Phase 2: Collaboration
    2.1 REPEAT
        2.1.1 $t++$
        2.1.2 Communicate cluster prototypes from each computer to all others
        2.1.3 For each computer D[ii]
            a. Compute induced partition matrices for data site D[ii]
            b. Repeat
                + Compute local partition matrices $u^{(t)}$ by Equation (22)
                + Compute local cluster prototypes $v^{(t)}$ by Equation (23)
                + Compute typicality partition matrix $\tau^{(t)}$ by Equation (24)
            Until the objective function is minimized (Stop condition 1)
        2.1.4 End for
    2.2 UNTIL $max((v^{(t)} - v^{(t-1)}) < \varepsilon$ OR $(t = T_{max})$ (Stop condition 2)
End.

Evaluation indicators

To evaluate the classification quality of the algorithm on experimental datasets, the paper uses several indicators to measure the quality of clusters. Partition coefficient (PC) [25] and partition entropy (PE) [26] are measures used to evaluate the quality of clustering results. The SC index is the ratio operating form of intra-cluster compactness and inter-cluster separation, and the XB index defines the inter-cluster separation as the minimum squared distance between cluster centers and the intra-cluster compactness as the mean squared distance between each data object and its cluster centers [27]. In the collaboration clustering model, we have not only the fuzzy membership function u_{ik} but the possibilistic membership function τ_{ik}. Where n_i is the number of data samples belonging to the i^{th} cluster. A large value of the PC index, while small PE, XB, and SC, indicates good clustering quality.

$$PC = \frac{1}{N}\sum_{i=1}^{C}\sum_{k=1}^{N}(u_{ik}^2 + \tau_{ik}^2) \tag{25}$$

$$PE = -\frac{1}{N}\sum_{i=1}^{C}\sum_{k=1}^{N}(u_{ik}\ln u_{ik} + \tau_{ik}\ln \tau_{ik}) \tag{26}$$

$$XB = \frac{\frac{1}{N}\sum\limits_{i=1}^{C}\sum\limits_{j=1}^{N}(u_{ij}^2 + \tau_{ij}^2)\|v_i - x_j\|^2}{\min\limits_{i\neq j}\|v_i - v_j\|^2} \tag{27}$$

$$SC = \sum_{i=1}^{C}\frac{\sum\limits_{j=1}^{n}(u_{ij} + \tau_{ij})\|v_i - x_j\|^2}{n_i * \sum\limits_{i=1}^{n}\|v_i - v_j\|^2} \tag{28}$$

In addition, we also sample clusters/classes from the original experimental dataset to evaluate the accuracy of the clustering results. The equation calculates the accuracy of the clustering results as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{29}$$

where *TP* is the number of correctly classified data, *FN* is the number of incorrectly misclassified data, *FP* is the number of incorrectly classified data, and *TN* is the number of correctly misclassified data. The better the algorithm is, the higher the *TPR* value is, and the smaller the *FTR* value is encountered.

4. Experimental results and discussion

In the experiments, we use two computers with the same configuration: Intel Core i7 2.9 GHz CPU, Windows 10 operating system, and the graphic card NVIDIA with a device memory size of 8Gb and 16Gb RAM. The algorithm is experimentally installed using the CUDA library in the C++ programming environment.

In the experimental part, the paper compares the clustering results of the proposed method with the algorithms CFCM [16], [17], CFCM-FR [28], CPFCM [29], and proposed algorithm CPFCM-FR. The experimental parameters are set together as follows: Fuzzy and possibilistic parameter $m = 2$ and $\eta = 2$; $a = b = 1$, the maximum number of iterations $T_{max} = 100$, stopping condition $\epsilon = 10^{-6}$. These parameters were selected based on previously published studies such as [16], [17], [29]. The running time of the algorithms is calculated from reading the data to giving the final result, including the dimensionality reduction step (if any). For each cluster we take 100 samples to evaluate the clustering quality.

4.1. Experiment 1

In experiment 1, we test a large dataset downloaded from the UCI machine learning library [19]. The details of the experimental dataset: A cybersecurity dataset containing nine different network attacks on a commercial IP-based surveillance system and an IoT network. The dataset consists of 27,170,754 observations with 115 features. The dataset includes reconnaissance, MitM, DoS, and botnet attacks. A total of 9 network capture datasets are as follows: 1. OS Scan (scans the network for hosts and their operating systems to reveal possible vulnerabilities); 2. Fuzzing (searches for vulnerabilities in the camera's web servers); 3. Video Injection (injects a recorded video clip into a live video stream); 4. ARP MitM (intercepts all LAN traffic via an ARP poisoning attack); 5. Active Wiretap (intercepts all LAN traffic via active wiretap); 6. SSDP Flood (overloads the DVR by causing cameras to spam the server); 7. SYN DoS (disables a camera's video stream by overloading its web server); 8. SSL Reneg (disables a camera's video stream by sending many SSL renegotiation packets to the camera); 9. Mirai (infects IoT with the Mirai malware).

For each attack (network capture) above, we provide (1) a CSV of the features used in our paper where each row is a network packet, (2) the corresponding labels [benign, malicious], and (3) the original network capture in truncated pcap format. We randomly divided the Kitsune Network Attack dataset into two subsets (2 data sites for two computers), in which one machine is set up in a central location, and the computers are connected via the local network. The data is clustered into 10 clusters (one benign packet cluster and nine attack packet clusters).

Table 1. Accuracy of classification results and computation time

No.	Algorithm	PC	PE	XB	SC	Accuracy	Time
1	CFCM	0.8367	0.6328	0.9843	0.6783	0.8589	21m18s
2	CFCM-FR	0.8324	0.6331	0.9844	0.6800	0.8588	**13m35s**
3	CPFCM	0.8858	**0.4984**	**0.6739**	0.5872	0.9045	26m41s
4	CPFCM-FR	**0.8942**	0.5006	**0.6739**	**0.5871**	**0.9068**	16m29s

Table 1 is the clustering results of four algorithms, CFCM, CFCM-FR, CPFCM, and CPFCM-FR, on the entire cybersecurity dataset. The higher the PC and Accuracy values give the better the clustering quality, whereas the smaller the PE, XB, SC, and running time values give the better the quality. Overall, the proposed algorithm CPFCM-FR gives better clustering results in most metrics. Specifically, the CPFCM-FR algorithm achieves the best values in the indices PC, XB, SC, and Accuracy. The PE index reaches 0.5006, higher than 0.4984 of the CPFCM algorithm, but the difference is insignificant. The running time of the CPFCM-FR algorithm also gives the result of 16m29s and is much faster than the CPFCM algorithm at 26m41s.

With the algorithm execution time, it can be seen that when clustering on a dataset without attribute dimensionality reduction, the running time is slower than when data dimensionality is reduced. Specifically, the clustering time on the CFCM algorithm is

21m18s, while the CFCM-FR algorithm is 13m35s. Similarly, the running time on the CPFCM algorithm is 26m41s, while the CPFCM-FR algorithm is 16m29s.

Table 1 also shows that the values of PC, PE, XB, SC, and Accuracy indices of CFCM and CFCM-FR algorithms show worse clustering quality than CPFCM and CPFCM-FR algorithms. This is because the CPFCM algorithm is an improvement of the CFCM algorithm and has higher accuracy than CFCM. Furthermore, dimensionality reduction based on the random projection feature reduction technique helps preserve distance, so the distance measurements on the new dataset are not significantly affected after dimensionality reduction. Therefore, the CPFCM-FR algorithm not only improves clustering quality compared to CFCM but also reduces the algorithm's running time.

The above results show that the proposed algorithm CPFCM-FR gives better clustering results than the algorithm before improvement. The accuracy of clustering results compared to when not using the random projection feature reduction is nearly equivalent. However, the calculation time is significantly reduced.

4.2. Experiment 2

The test data in experiment 2 is the Sentinel-2A satellite image in two areas (two data sites); the spatial resolution of the image is 10 m, and the number of spectral bands is 12. It is a free and good-quality satellite image [20]. The data were taken in the Hanoi central and Vinh Phuc areas, all north of Vietnam. The authors choose satellite images taken when there are no clouds to avoid being affected by clouds (Figure 3). The computers are connected via the local network.

Fig. 3. Experimental data: Hanoi central and Vinh Phuc areas.

The total pixel count of all 2 data sites is 1,280,000 pixels. These labeled data are used to calculate the additional membership function values for the semi-supervised

solution. Using the proposed algorithm, classified into six classes of objects describing six types of land covers: Class 1: Rivers, lakes, ponds. Class 2: Vacant land, roads, etc. Class 3: Field, grass. Class 4: Sparse forest, low trees. Class 5: Perennial plants. Class 6: Dense forest, jungle.

Figure 4 results from classifying satellite image data sites on two computers using the proposed algorithm. The resulting image shows six land cover layers from satellite image data.

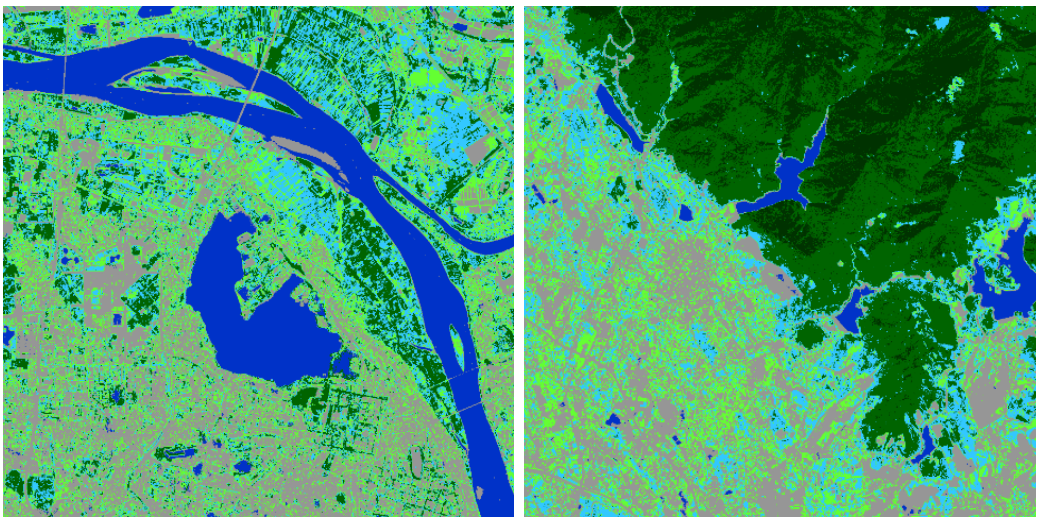

Fig. 4. Experimental results on the proposed algorithm: Hanoi central and Vinh Phuc areas.

Table 2. Accuracy of classification results and computation time

No.	Algorithm	PC	PE	XB	SC	Accuracy	Time
1	CFCM	0.8914	0.5276	0.7823	0.7635	0.9132	8m27s
2	CFCM-FR	0.8986	0.5288	0.7822	0.7722	0.9127	**5m41s**
3	CPFCM	0.9367	**0.4992**	**0.7089**	0.5182	0.9306	10m33s
4	CPFCM-FR	**0.9389**	0.5003	0.7091	**0.5099**	**0.9311**	6m18s

Table 2 shows classification quality indicators and implementation time when clustering on the satellite image dataset. The PC, SC, and Accuracy indices show that the proposed algorithm CPFCM-FR gives better clustering quality than the three algorithms CFCM, CFCM-FR, and CPFCM. While the PE and XB indices achieved the best with 0.4992 and 0.7089 for the CPFCM algorithm, this value is not significantly higher than the 0.5003 and 0.7091 of the CPFCM-FR algorithm. Although the running time of the CPFCM-FR algorithm is 6m18s, which is not the lowest value, it is significantly higher than the 10m33s of the CPFCM algorithm. The CPFCM-FR algorithm gives the highest accuracy with an Accuracy index value of 0.9311, followed by the CPFCM algorithm of 0.9306. The Accuracy index on the two algorithms, CFCM and CFCM-FR, gives lower values of 0.9132 and 0.9127, respectively.

The algorithms' running times are significantly different. While the CFCM-FR algorithm gives the fastest running time, 5m41s, the CPFCM algorithm gives the slowest running time, 10m33s. The running time of the CFCM algorithm is 8m27s, and that of the CPFCM-FR algorithm is 6m18s. This result also shows that dimensionality reduction does not significantly affect the clustering quality but can also significantly reduce the algorithm execution time.

The clustering results in the two datasets shown in Table 1 and Table 2 show that the proposed algorithm CPFCM gives better clustering results than the CFCM algorithm. Similarly, when using the random projection feature reduction technique, the CPFCM-FR algorithm also gives better clustering results than the CFCM-FR algorithm. In addition, the calculation time on the datasets after feature reduction is smaller than clustering on the original dataset.

5. Conclusions

The paper presented an improvement of the collaborative possibilistic fuzzy clustering algorithm based on random projection feature reduction for multi-dimensional analysis. Experimental results show that the proposed method can significantly reduce the computation time while the accuracy does not change significantly. Random projection feature reduction techniques can help preserve the relative distance between data samples. This makes the distance measure between data samples and data samples to cluster centers on the dataset before and after dimensionality reduction preserved. Comparing the clustering results of the proposed method with some other methods shows that the proposed method gives better results in both accuracy and algorithm execution time as indicated by the PC, PE, XB, SC, Accuracy indices, and running time. This confirms that when dimensionality reduction techniques are incorporated, the clustering quality of the CPFCM-FR algorithm remains unchanged or is little affected. In contrast, the algorithm execution time is significantly reduced. This result shows the potential for developing collaborative clustering models combined with data dimensionality reduction techniques that preserve distance for high-dimensional big data analysis problems.

In the future, we will experiment with multiple computers on the Internet to be able to solve high-dimensional big data problems where decentralized data resides in many different places.

Acknowledgments

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number **105.99-2023.12**. This research is also funded by Le Quy Don Technical University under the grant number **24.1.53**.

References

[1] W. Wang, "Big data, big challenges," in *IEEE International Conference on Semantic Computing*, 2014, pp. 1–6. DOI: 10.1109/ICSC.2014.65

[2] M. A. Mahdi, K. M. Hosny, and I. Elhenawy, "Scalable clustering algorithms for big data: A review," *IEEE Access*, vol. 9, pp. 80 015–80 027, 2021. DOI: 110.1109/ACCESS.2021.3084057

[3] S. Yinghua, P. Witold, C. Yuan, W. Xianmin, and G. Adam, "Hyperplane division in fuzzy c-means: Clustering big data," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 11, pp. 3032–3046, 2020. DOI: 10.1109/TFUZZ.2019.2947231

[4] X. Du, Y. He, and J. Z. Huang, "Random sample partition-based clustering ensemble algorithm for big data," in *IEEE International Conference on Big Data*, 2021, pp. 5885–5887. DOI: 10.1109/BigData52589.2021.9671297

[5] S. Azzouzi, A. Hjouji, J. EL-Mekkaoui, and A. E. Khalfi, "A novel efficient clustering algorithm based on possibilistic approach and kernel technique for image clustering problems," *Applied Intelligence*, vol. 53, p. 4327–4349, 2023. DOI: 10.1007/s10489-022-03703-0

[6] D.-W. Jia and Z.-Y. Wu, "Effect of fuzzy failure criterion on probabilistic seismic risk analysis under multidimensional performance limit state," *Journal of Building Engineering*, vol. 52, 2022. DOI: 10.1016/j.jobe.2022.104438

[7] H. Yu, L. Jiang, J. Fan, and R. Lan, "Double-suppressed possibilistic fuzzy gustafson–kessel clustering algorithm," *Knowledge-Based Systems*, vol. 276, 2023. DOI: 10.1016/j.knosys.2023.110736

[8] H. Yu, H. Li, X. Xu, Q. Gao, and R. Lan, "Suppressed possibilistic fuzzy c-means clustering based on shadow sets for noisy data with imbalanced sizes," *Applied Soft Computing*, 2024. DOI: 10.1016/j.asoc.2024.112263

[9] C. Wu and M. Li, "Generalized multiplicative fuzzy possibilistic product partition c-means clustering," *Information Sciences*, vol. 670, 2024. DOI: 10.1016/j.ins.2024.120588

[10] F. Anum and K. H. Memon, "Kernel possibilistic fuzzy c-means clustering algorithm based on morphological reconstruction and membership filtering," *Fuzzy Sets and Systems*, vol. 477, 2024. DOI: 10.1016/j.fss.2023.108792

[11] B. Fang, "Some uncertainty measures for probabilistic hesitant fuzzy information," *Information Sciences*, vol. 625, pp. 255–276, 2023. DOI: 10.1016/j.ins.2022.12.101

[12] M. S. Mahmud, J. Z. Huang, S. Salloum, T. Emara, and K. Sadatdiynov, "A survey of data partitioning and sampling methods to support big data analysis," *Big Data Mining and Analytics*, vol. 3, no. 2, pp. 85–101, 2020. DOI: 10.26599/BDMA.2019.9020015

[13] J. C. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm," *Computers Geosciences*, vol. 10, no. 2, pp. 191–203, 1984. DOI: 10.1016/0098-3004(84)90020-7

[14] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Wu, "An efficient k-means clustering algorithm: analysis and implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 881–892, 2002. DOI: 10.1109/TPAMI.2002.1017616

[15] L. Zhu, F.-L. Chung, and S. Wang, "Generalized fuzzy c-means clustering algorithm with improved fuzzy partitions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 3, pp. 578–591, 2009. DOI: 10.1109/TSMCB.2008.2004818

[16] W. Pedrycz, "Collaborative fuzzy clustering," *Pattern Recognition Letters*, vol. 23, no. 14, pp. 1675–1686, 2002. DOI: 10.1016/S0167-8655(02)00130-7

[17] W. Pedrycz and P. Rai, "Collaborative clustering with the use of fuzzy c-means and its quantification," *Fuzzy Sets and Systems*, vol. 159, no. 18, pp. 2399–2427, 2008. DOI: 10.1016/j.fss.2007.12.030

[18] D. S. Mai and L. T. Ngo, "Semi-supervised fuzzy c-means clustering for change detection from multispectral satellite image," in *IEEE International Conference on Fuzzy Systems*, 2015, pp. 1–8. DOI: 10.1109/FUZZ-IEEE.2015.7337978

[19] D. Aha, "Machine learning repository," 2024. [Online]. Available: https://archive.ics.uci.edu/datasets

[20] USGS, "The united states geological survey (usgs)," 2024. [Online]. Available: https://earthexplorer.usgs.gov/

[21] R. P. Nikhil, P. Kuhu, J. M. Keller, and J. C. Bezdek, "A possibilistic fuzzy c-means clustering algorithm," *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 4, p. 517–530, 2005. DOI: 10.1109/TFUZZ.2004.840099

[22] Y. Shen and W. Pedrycz, "Collaborative fuzzy clustering algorithm: Some refinements," *International Journal of Approximate Reasoning*, vol. 86, pp. 41–61, 2017. DOI: 10.1016/j.ijar.2017.04.004

[23] J. A. Lee and M. Verleysen, "Two key properties of dimensionality reduction methods," in *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 2014, pp. 163–170. DOI: 10.1109/CIDM.2014.7008663

[24] W. Johnson and J. Lindenstraus, "Extensions of lipshotz mapping into hilbert space," *Contemporary Mathematics*, pp. 189–206, 1984. DOI: 10.1090/conm/026/737400

[25] J. C. Bezdek, "Numerical taxonomy with fuzzy sets," *Journal of Mathematical Biology*, vol. 1, p. 57–71, 1974. DOI: 10.1007/BF02339490

[26] J. C. Bezdek†, "Cluster validity with fuzzy sets," *Journal of Cybernetics*, vol. 3, no. 3, pp. 58–73, 1973. DOI: 10.1080/01969727308546047

[27] H.-Y. Wang, J.-S. Wang, and G. Wang, "A survey of fuzzy clustering validity evaluation methods," *Information Sciences*, vol. 618, pp. 270–297, 2022. DOI: 10.1016/j.ins.2022.11.010

[28] T. H. Dang, V. D. Do, D. S. Mai, L. T. Ngo, and L. H. Trinh, "Features reduction collaborative fuzzy clustering for hyperspectral remote sensing images analysis," *Journal of Intelligent Fuzzy Systems*, vol. 45, no. 5, pp. 7739–7752, 2023. DOI: 10.3233/JIFS-230511

[29] V. D. Do, D. S. Mai, and L. T. Ngo, "Approaching semi-supervised collaborative learning model for remote sensing image analysis," in *2022 RIVF International Conference on Computing and Communication Technologies*, 2022, pp. 548–553. DOI: 10.1109/RIVF55975.2022.10013798

Manuscript received 5-8-2024; Accepted 27-12-2024.

■

Viet Duc Do is a research at National Defense Academy, Vietnam. He received the B.S. (2009), M.S. (2014) degrees in GeoInformatics and Computer Science from Le Quy Don Technical University. Now, he is a PhD. Student at Le Quy Don Technical University. His research interests are fuzzy clustering, remote sensing image processing techniques, pattern recognition, high performance computing. Email: minhducmta@gmail.com

Dinh Sinh Mai is a lecturer at Institute of Techniques for Special Engineering, Le Quy Don Technical University. He received the B.S. (2009), M.S. (2013) and PhD. (2021) degrees in GeoInformatics and Computer Science from Le Quy Don Technical University. His research interests are fuzzy clustering, remote sensing image processing techniques, pattern recognition and geographic information system technologies. Email: maidinhsinh@lqdtu.edu.vn

Long Thanh Ngo received the M.Sc, Ph.D degrees in Computer Science from Le Quy Don Technical University, in 2003 and 2009, respectively. He is an Associate Professor at the Institute of Information and Communication Technology, Le Quy Don Technical University. His current research interests include computational intelligence, type-2 fuzzy logic, pattern recognition and image processing. Email: ngotlong@lqdtu.edu.vn

TIẾP CẬN PHÂN CỤM C-MEANS MỜ KHẢ NĂNG CỘNG TÁC CHO PHÂN TÍCH DỮ LIỆU NHIỀU CHIỀU

Đỗ Viết Đức, Mai Đình Sinh, Ngô Thành Long

Tóm tắt

Sự phát triển nhanh chóng của các công nghệ thu thập dữ liệu đã dẫn đến sự bùng nổ các nguồn dữ liệu. Nhiều kỹ thuật và phương pháp khai phá dữ liệu truyền thống đã trở nên lỗi thời và không còn phù hợp để giải quyết các vấn đề dữ liệu lớn, dữ liệu nhiều chiều. Bài báo này đề xuất cải thiện thuật toán phân cụm mờ khả năng cộng tác để phân tích dữ liệu nhiều chiều bằng cách sử dụng kỹ thuật giảm chiều dựa trên phép chiếu ngẫu nhiên (CPFCM-FR). Kỹ thuật này cho phép bảo toàn khoảng cách tương đối sau khi giảm chiều, có thể giúp giảm độ phức tạp tính toán trong khi vẫn đảm bảo độ chính xác của thuật toán được đề xuất so với thuật toán trước khi giảm chiều. Thuật toán đề xuất triển khai trên mô hình phân cụm cộng tác có thể giúp chia sẻ thông tin về cấu trúc cụm tại các vị trí dữ liệu khác nhau (data site) trong quá trình tính toán. Mô hình cộng tác cho phép giải quyết các vấn đề khi dữ liệu nằm phân tán trên các máy tính khác nhau trong hệ thống mạng. Các thực nghiệm được thực hiện trên hai tập dữ liệu nhiều chiều được tải xuống từ thư viện học máy UCI và dữ liệu ảnh viễn thám cho thấy phương pháp được đề xuất mang lại kết quả tốt hơn đáng kể so với một số phương pháp được đề xuất trước đây. Các kết quả thực nghiệm này cũng minh chứng cho tiềm năng phát triển các mô hình phân cụm cộng tác, kết hợp với các kỹ thuật giảm chiều, để giải quyết các vấn đề dữ liệu lớn, nhiều chiều, và phân tán.

Từ khóa

Dữ liệu đa chiều; phân cụm c-means mờ khả năng; phân cụm cộng tác; phép chiếu ngẫu nhiên; giảm chiều.