APPLYING SEMI-SUPERVISED FUZZY C-MEANS CLUSTERING ALGORITHM BASED ON COLLABORATIVE CLUSTERING MODEL FOR LAND COVER CLASSIFICATION FROM LANDSAT-7 IMAGERY

Dinh Sinh Mai[1,*], Tuan Kiet Nguyen[2], Chi Hieu Le[2], Le Hung Trinh[1]

[1]*Institute of Techniques for Special Engineering, Le Quy Don Technical University*

[2]*Military Topography Class Course 56, Le Quy Don Technical University*

Abstract

The rapid development of artificial satellites has led to an explosion of remote sensing data sources. Centralized storage of large data sources is becoming increasingly complex, and decentralized storage solutions on distributed systems are increasingly gaining attention. Traditional data mining techniques have become obsolete and are no longer suitable for solving large, multidimensional, distributed data problems. These datasets, for some reasons such as security, data transmission, privacy, etc., cannot be shared directly between computers but can only share information about cluster structure. This article presents a semi-supervised fuzzy c-means clustering algorithm based on the collaborative clustering model (CSFCM) on distributed systems applied to the problem of land cover classification from remote sensing data. The proposed model aims to solve the problem of land cover classification where remote sensing data is decentralized and stored on a distributed system of computers connected via the network. Experiments on four optical satellite image datasets show that the proposed method provides significantly better results in both classification quality and classification time compared to local clustering on individual datasets. This result suggests that developing collaborative model-based data analysis algorithms can help solve the problem of remote or distributed remote sensing image data analysis.

Keywords: *Land cover classification; remote sensing imagery; distributed systems; collaborative clustering.*

1. Introduction

With the rapid development of satellite science and technology, many remote sensing data sources are collected and stored (big data) [1]. From multiple sources and scales, complex structures and large volumes have led to an overload of centralized storage systems. The current solution for storing large data sources is to divide them into smaller datasets and store them in a distributed manner on a network of interconnected computers [2]. Data processing, therefore, requires the development of algorithms and

[*] Corresponding author, email: maidinhsinh@lqdtu.edu.vn

DOI: 10.56651/lqdtu.jst.v7.n02.876.sce

methods that enable decentralized data analysis on distributed systems [3]. This approach can have a significant impact on data clustering, especially when the datasets are related. If these datasets are related, clustering on one dataset can impact and influence clustering on other datasets. However, these datasets cannot be clustered centrally for many reasons, such as data privacy, security, transmission, etc. To address these challenges, it is necessary to develop solutions that effectively handle distributed data issues. This approach is important to overcome the limitations of centralized storage and ensure efficient data clustering.

Collaborative data clustering is a tool to find structural similarities between data samples located in multiple distinct regions based on the expansion of the objective function and the fuzzy clustering method of the fuzzy c-means clustering algorithm [4]. Pedrycz introduced collaborative fuzzy clustering to find structures and similarities between distinct datasets (distributed) [5]. There are two important characteristics of collaborative fuzzy clustering. One is that detailed information in datasets cannot be exchanged; only information about cluster structure can be exchanged. The second is to consider whether clustering on this dataset affects clustering on other datasets.

Nowadays, parallel and distributed computing is one of the research directions many scientists are interested in [6, 7]. Parallel and distributed computing is an important tool in reducing the execution time, and it can be suitable for detecting objects on the land's surface in real-time or near real-time from airborne and space-based platforms to support immediate decision-making. This paper [6] reviews recent advances in anomaly detection from hyperspectral remote sensing images and their implementation using parallel and distributed systems. Wu *et al.* provide a survey of state-of-the-art methods for processing remotely sensed big data and thoroughly study existing parallel implementations on distributed systems [7]. Feng *et al.* presented a study on applying distributed cloud computing architecture in hyperspectral remote sensing image classification based on big data on the Spark platform [8].

Research by O'Reilly *et al.* shows that a distributed anomaly detection model in many different network infrastructures can provide better results than a centralized model [9]. Li *et al.* proposed to build a distributed file system to manage remote sensing image data, taking ordinary files as the data model and TCP as the data transmission model [10]. Experiments show that the proposed distributed file system has stable read and write performance compared with existing systems. Wang *et al.* proposed an innovative distributed collaborative method (DCM) for training remote sensing image classification, showing that the proposed training method has better collaborative learning ability than the centralized model [11]. Obtaining a comprehensive view of the entire

flooded area is an urgent issue in flood disasters. Xie *et al.* proposed a near-real-time flood mapping system for automatic flood mapping with remote sensing image data and related computational algorithms exploited in a collaborative environment [12]. Li *et al.* designed and implemented a distributed parallel processing system for multi-source remote sensing data based on a distributed cluster platform [13]. The system connects several satellite data centers, serves several applications, and implements dynamic scaling integration for high-performance quantitative remote sensing products.

The article proposes an algorithm for classifying land cover objects from remote sensing images on a distributed system based on semi-supervised fuzzy c-means clustering [14, 15] and a collaborative clustering model [4, 5]. This approach can effectively solve decentralized data analysis problems, taking advantage of the power of multiple computers on a distributed computing system. To experiment with the proposed method, we use four optical remote-sensing image datasets stored on four computers connected to each other via the network. The experimental results show that the proposed method gives better results in both accuracy and running time compared to performing it individually on each dataset.

The article is organized into four sections: Section 1 is the introduction overview of the research content; Section 2 introduces some related knowledge; Section 3 presents results and discussion; Section 4 gives the conclusion.

2. Materials and methodology

2.1. Materials

Landsat multi-temporal satellite images, after being collected from the USGS database are pre-processed to remove spectral and geometric errors. Remote sensing data used in the study are Landsat-7 TM satellite images taken from central Hanoi and surrounding areas north of Hanoi [16], including image scenes on September 30, 2009 (Fig. 1). Satellite images are collected at a time not affected by weather. Experimental area coordinates from 104° 39' 01.9986" E, 21° 38' 13.7121" N to 106° 27' 53.6258" E, 20° 53' 43.6835" N. Satellite image size 1916 × 831 corresponding to 1,592,196 pixels. Landsat satellite image data is classified into six layers corresponding to six corresponding land cover class types as follows:

Class 1: Rivers, lakes, ponds.

Class 2: Vacant land, roads.

Class 3: Field, grass.

Class 4: Sparse forest, low trees.

Class 5: Perennial plants.

Class 6: Dense forest, jungle.

Fig. 1. Landsat image data in Hanoi area and surrounding areas on September 30, 2009.

2.2. Methodology

2.2.1. Semi-supervised fuzzy clustering

In data clustering problems, semi-supervised clustering is a hybrid technique between supervised and unsupervised clustering. The advantage of this technique is that it uses a very small amount of labeled data to improve the accuracy of the clustering results. This is very suitable for datasets that cannot be applied to supervised learning techniques due to difficulties in labelling or having very little labeled data. Many of these studies are semi-supervised c-means fuzzy clustering algorithms (SFCM) [14]. The objective function of the algorithm is supplemented with information about the labeled data.

The SFCM algorithm model is to optimize the following objective function:

$$J_m(U,V,X) = \sum_{i=1}^{c}\sum_{k=1}^{n} \mu_{ik}^{m}(d_{ik}^{2} + (v_i - v_i^{*})^2) \quad (1)$$

where v^* is the centroid computed from the labeled data, $U=[\mu_{ik}]_{cxn}$ is a fuzzy MF, $V=(v_1,v_2,...,v_c)$ is a vector of (unknown) cluster centers, $X=\{x_k, x_k \in R^M, k=1,...,n\}$, $d_{ik}=\|v_i-x_k\|$. With the following constraints:

$$m>1;\ 0\le \mu_{ik} \le 1;\ \sum_{i=1}^{c}\mu_{ik}=1;\ 1\le i\le c;\ 1\le k\le n \tag{2}$$

The objective function $J_m(U,V,X)$ reaches the smallest value when and only if:

$$v_i=\sum_{k=1}^{n}\mu_{ik}^m \frac{(x_k+v_i^*)}{2} / \sum_{k=1}^{n}\mu_{ik}^m \tag{3}$$

$$\mu_{ik}=1/\sum_{j=1}^{c}\left(\frac{[d_{ik}^2+(v_i-v_i^*)^2]}{[d_{jk}^2+(v_i-v_i^*)^2]}\right)^{1/(m-1)} \tag{4}$$

Equation (3), (4) can be obtained based on the Lagrange multiplier theorem with the constraints by objective function (2). SFCM algorithm will perform iterations according to Eq. (3), (4) until the objective function $J_m(U,V,X)$ reaches the minimum value.

2.2.2. Collaborative fuzzy clustering model on distributed systems

The idea of collaborative clustering is to locally cluster *P* subsets of data at computers, the cluster centroids obtained after clustering are shared among computers to calibrate the local cluster centroids. This process is repeated until all local cluster centroids do not change significantly, then stop and give the final clustering result.

The collaborative fuzzy clustering problem has the objective function that needs to be optimized as:

$$Q_{[ii]}=\sum_{k=1}^{N[ii]}\sum_{i=1}^{C}u_{ik}^2[ii]d_{ik}^2+\beta[ii/\ jj]\sum_{jj=1}^{P}\sum_{k=1}^{N[ii]}\sum_{i=1}^{C}(u_{ik}-\tilde{u}_{ik}[ii/\ jj])^2 d_{ik}^2 \tag{5}$$

The above objective function consists of two parts, the first part is similar to the objective function of the FCM algorithm [15]. The second part describes the collaboration information between datasets on computers. In the above objective function, d_{ik} is the distance between the k^{th} pixel to the i^{th} cluster center. The parameter $\beta[ii/\ jj]$ represents the cooperation coefficient between datasets. The larger the value of $\beta[ii/\ jj]$, the higher the cooperation level, and the value $\beta[ii/\ jj]=0$ represents that there is no cooperation between datasets *ii* and *jj*. $u_{ik}[ii]$ is the fuzzy partition matrix of object *k* into cluster *i* in

dataset *ii*. $\tilde{u}_{ik}[ii/jj]$ is called the collaboration fuzzy partition matrix of dataset *jj* into dataset *ii* and is calculated by the formula:

$$\tilde{u}_{ik}[ii/jj] = 1 \Big/ \sum_{j=1}^{c} \left(\frac{x_k[ii] - v_i[jj]}{x_k[ii] - v_j[jj]} \right)^2 \tag{6}$$

Using the Lagrange multiplier method to optimize the above objective function, we have the following formula for calculating the fuzzy partition matrix and cluster centroids on each data site:

$$u_{ik}[ii] = \frac{\left[a(v_i[ii] - x_k[ii])^2 + \beta[ii/jj](v_i[ii] - \tilde{v}_i\ [ii])^2 \right]^{-1/(m-1)}}{\sum_{j=1}^{c[ii]} \left[\frac{1}{a(v_j[ii] - x_k[ii])^2 + \beta[ii/jj](v_j[ii] - \tilde{v}_j\ [ii])^2} \right]^{1/(m-1)}} \tag{7}$$

$$v_i[ii] = \frac{\sum_{k=1}^{N[ii]} u_{ik}^m[ii] x_k[ii] + \beta[ii/jj](u_{ik}[ii] - \tilde{u}_{ik}[ii/jj])^m \tilde{v}_i\ [ii])}{\sum_{k=1}^{N[ii]} (u_{ik}^m[ii] + \beta[ii/jj](u_{ik}[ii] - \tilde{u}_{ik}[ii/jj])^m)} \tag{8}$$

The fuzzy partition matrix of the objective function must satisfy the constraint that the sum of the memberships of an element in clusters in the same dataset is equal to 1 as follows: $U = \{u_{ik} \in [0,1], \sum_{i=1}^{c} u_{ik}[ii] = 1, \forall k;\ 0 < \sum_{k=1}^{N[ii]} u_{ik}[ii] < N[ii], \forall i\}$.

The results performed on the data sites have collaborated to calculate the global centroid, which is sent back to the data sites for calculation in the next phase.

After performing clustering on the data sites, the global cluster center is calculated by summing the local cluster center values as follows:

$$V_i = \frac{\sum_{ii=1}^{C[ii]} v_i[ii]}{C[ii]} \tag{9}$$

where $V = \{V_i, i = 1,...,C\}$ is the global cluster centroid of the entire dataset, $v[ii] = \{v_i[ii], ii = 1,...,C[ii]\}$ is the local cluster centroid of the ii^{th} sub-dataset. $C[ii]$ is the number of clusters of the ii^{th} sub-dataset. In this study, the number of clusters (classes) in the sub-datasets is the same.

A peer-to-peer distributed system type follows a decentralized organization. Each device can operate as both the client and server. Computer network applications use a peer-to-peer system to organize processors that communicate with each other but

maintain separate local memory bases. The programs and servers in this network all have the same privileges, access and functions and communicate at the same level without a hierarchy. The computers used in the experiment have the same configuration, Intel Core i5 2.99 GHz, Windows 10 operating system, 16GB RAM, NVIDIA graphics card with 4.0GB device memory and Visual C^{++} programming environment. During the experiment, the connection, sending and receiving of data between computers using TCP/IP protocol is used to identify and communicate between computers.

2.3. Proposed method

Figure 2 shows the proposed model on a distributed computer system. The initial data is divided into 4 small datasets and placed on 4 computers connected to each other via a network.

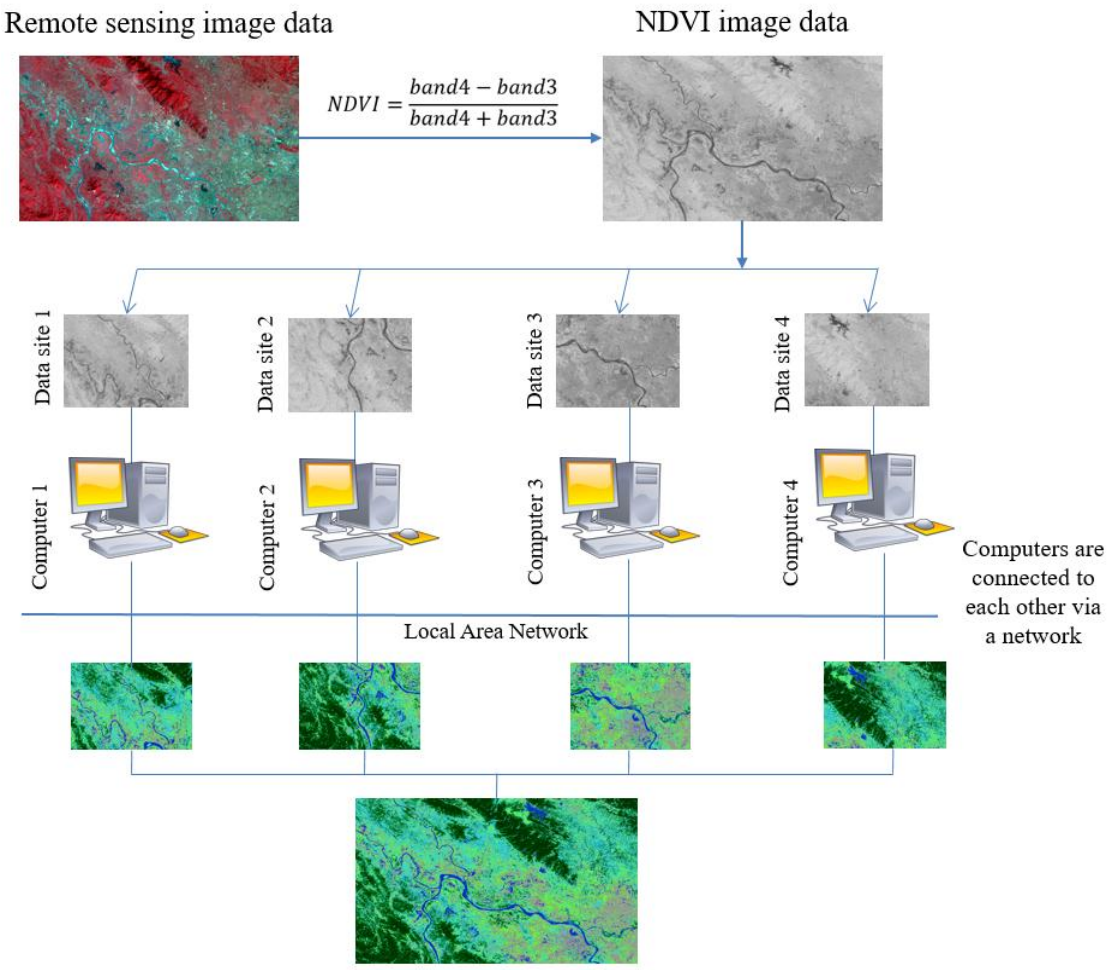


Fig. 2. Proposed model on the distributed system.

Given a remote sensing image dataset $X = x_1, x_2, ..., x_n$ with $x_i, i = 1, ..., n$ being the pixels. The dataset *X* is divided into *P* sub-datasets $D[1], D[2], ..., D[P]$, where each sub-dataset contains $N[1], N[2], ..., N[P]$ pixels. The *P* subsets are stored on *P* different computers and are connected to each other via a network (Fig. 2).

The proposed CSFCM algorithm consists of 2 phases, phase 1 is the local clustering step using the semi-supervised fuzzy c-means clustering algorithm (SFCM), phase 2 uses the results in phase 1 to perform the collaboration step between data sites on computers in the network.

The steps to implement the CSFCM algorithm are detailed as follows:

Algorithm 1: The CSFCM algorithm

Input: Dataset X, ε, and initialize the parameters $m, \eta; a, b; \epsilon$, β, the maximum number of iterations T_{max}, the number of data site *P*, the number of items in each data site ii is $N[ii]$, the number of clusters in each data site ii is $c[ii]$, the data item in each data site $X[ii]$; $t = 0$.

Output: Clustering results.

Begin

Phase 1: Locally clustering

1.1 Put P data sites on different computers.

1.2 Run SFCM algorithm for each data site (at each computer).

Phase 2: Collaboration

2.1 **REPEAT**

2.1.1 $t = t + 1$.

2.1.2 Compute global cluster centroid $V = \{V_i, i = 1, ..., C\}$ by Eq. (9).

2.1.3 Send the global cluster centroid result *V* to all computers.

2.1.4 At each computer $D[ii]$: $t[ii] = 0$

a. **Repeat**

+ $t[ii] = t[ii] + 1$;

+ Compute local fuzzy membership matrices $u^{(t[ii])}[ii]$ by Eq. (7).

+ Compute local cluster centroids $v^{(t[ii])}[ii]$ by Eq. (8).

b. **Until** $max(v^{(t[ii])}[ii] - v^{(t[ii]-1)}[ii]) < \varepsilon$ **or** $t[ii] \geq T_{max}$ (Stop condition 1)

2.2 **UNTIL** $max(V^{(t)} - V^{(t-1)}) < \varepsilon$ **OR** $t \geq T_{max}$ (Stop condition 2)

End.

The collaborative process is to share the local clustering results on the sub-datasets and recalculate the global centroid for the entire dataset. Next, use the global centroid to cluster locally on the sub-datasets, and the results are shared again to calculate the global cluster centroid. This process is repeated until the cluster centroid changes insignificantly or the number of iterations reaches the maximum, then stop and return the clustering results. From this clustering result, use the labeled data to assign clusters to the corresponding land cover classes.

2.4. Accuracy assessment

In this study, we use some cluster quality evaluation indexes such as Partition Coefficient (PC), Partition Entropy (PE) [17]. The larger the PC and the smaller the PE, the better the clustering result.

$$PC = \frac{1}{N}\sum_{i=1}^{C}\sum_{k=1}^{N}u_{ik}^{2} \tag{10}$$

$$PE = -\frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{N}u_{ik}\ln u_{ik} \tag{11}$$

The article also uses labeled data to evaluate the accuracy of experimental results. The accuracy of the classification results is calculated by the formula:

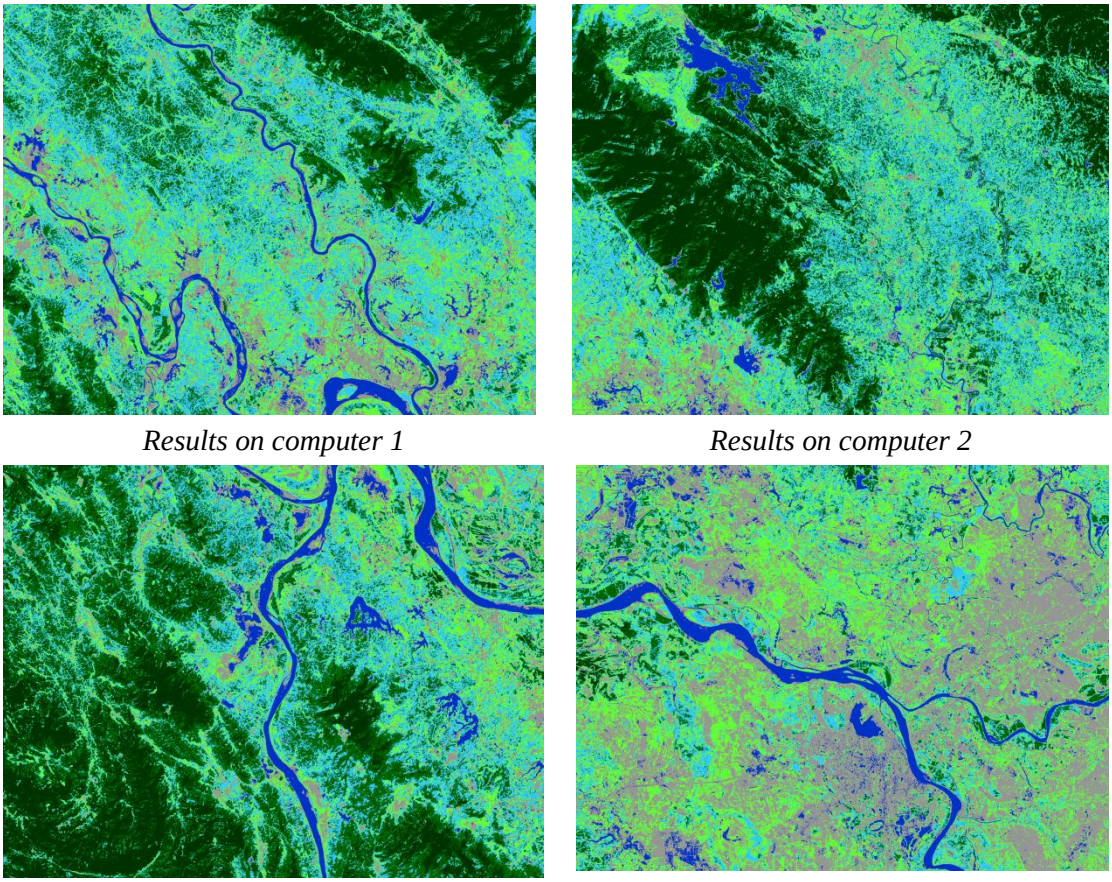
$$Overall\ Accuracy = \frac{n^{*}}{N^{*}}\cdot 100\% \tag{12}$$

in which the total number of labeled samples is N^{*}, and the number of correctly classified labeled samples is n^{*}.

3. Result and discussion

In each sub-dataset, we take 50 sample pixels to label each land cover class. The total number of sampled pixels for the entire dataset is 1,200 pixels. The parameters chosen for the experiment are $m = 2$, stop condition $\varepsilon = 10^{-6}$, maximum number of iterations $T_{max} = 1000$. The sub-datasets have the same role in the whole dataset therefore the cooperation coefficients between datasets $\beta[ii/jj] = 1$. Labeled data is used for the SVM, RF, SFCM, CSFCM algorithms.

Figure 3 is the classification result on four computers using the proposed CSFCM algorithm. It can be seen that the classes are relatively clear, especially class 1 describing rivers, lakes, and ponds.



Results on computer 1 *Results on computer 2*

Results on computer 3 *Results on computer 4*

Fig. 3. Results of land cover/land use classification at the computers.

Table 1 shows the results of land cover classification using SVM, RF, SFCM and CSFCM algorithms according to the PC, PE, Overall Accuracy and Running time indices. Overall, it can be seen that the proposed CSFCM method gives the best results in all the indices. Figure 4 is the result of land cover classification using the CSFCM algorithm after synthesizing the classification results on 4 computers (4 sub-datasets).

Table 1. The comparison of classification results of different algorithms

No.	**Index**	**Algorithms**			
		SVM	**RF**	**SFCM**	**CSFCM**
1	PC	0.6643	0.7648	0.8183	**0.8209**
2	PE	0.5589	0.5476	0.4928	**0.4876**
3	Overall Accuracy	90.41%	94.12%	95.88%	**96.09%**
4	Running time	8m28s	9m30s	15m32s	**5m46s**

For the PC index, the CSFCM algorithm gives a value of 0.8209, while the other algorithms achieve 0.8183 with SFCM, 0.7648 with RF and 0.6643 with SVM. The PE

index achieves 0.4876 with CSFCM, 0.4928 with SFCM, 0.5476 with RF and 0.5589 with SVM.

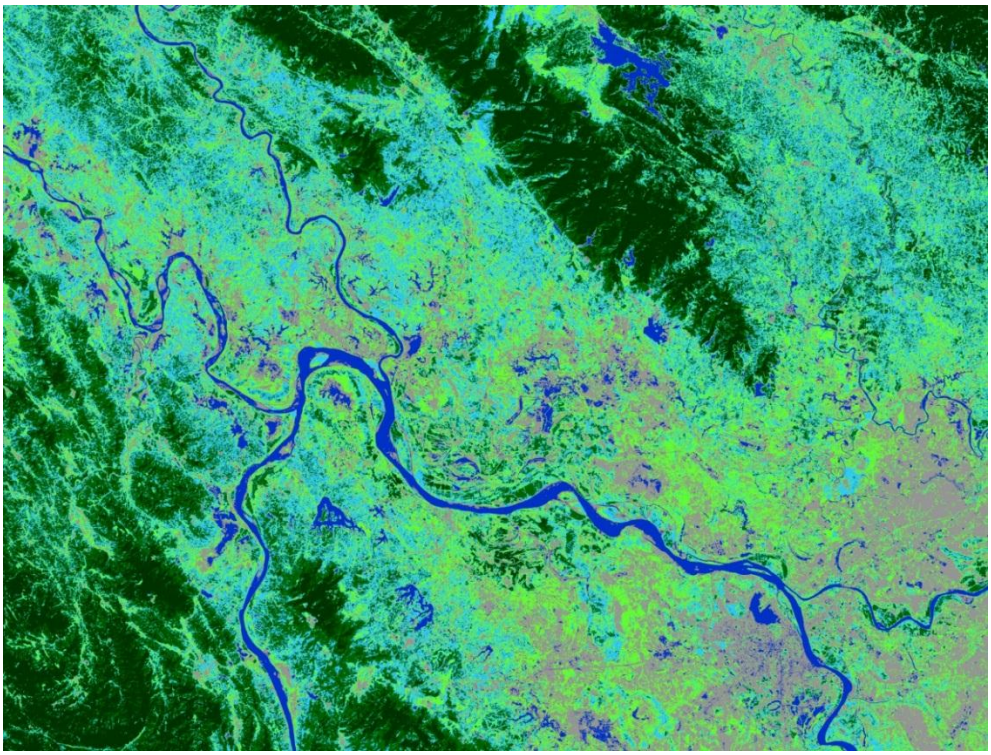


Fig. 4. Results of land cover/land use classification in Hanoi area and surrounding areas.

The Overall Accuracy index calculated on the entire dataset (not calculated separately on each sub-dataset) also shows that the proposed CSFCM algorithm gives much higher accuracy than other algorithms. Specifically, the CSFCM algorithm has an Overall Accuracy value of 96.09%, while the other algorithms are 95.88% with SFCM, 94.12% with RF and 90.41% with SVM, respectively.

The two semi-supervised clustering algorithms, CSFCM and SFCM, outperform the two conventional machine learning algorithms, SVM and RF. The proposed CSFCM algorithm performs best on all four metrics, followed by SFCM and RF. The SVM algorithm performs worst among the four experimental algorithms.

For the Running time index, the proposed algorithm gives the smallest execution time among the four experimental algorithms with a time of 5m46s. It is almost 3 times faster than the SFCM algorithm. The execution time of the SVM algorithm is 8m28s, the RF algorithm is 9m30s, while the SFCM algorithm has the highest execution time at 15m32s.

From the above experiments, the application of data analysis problems based on the collaborative model of distributed systems not only reduces the calculation time but also improves the accuracy. This also shows that the research in the article has a lot of potential for development, expansion, and application in practice.

4. Conclusion

This article presents a collaborative semi-supervised fuzzy clustering (CSFCM) algorithm on distributed computing systems for land cover classification from satellite image data. The proposed algorithm allows the clustering of data placed on multiple computers, sharing clustering results to improve accuracy without sharing raw data. Moreover, the algorithm also helps reduce computation time since the computation is performed on computers in the network instead of being performed centrally on a single computer. Experiments on land cover classification from Landsat-7 TM remote sensing data using SVM, RF, SFCM, and CSFCM algorithms show that the proposed CSFCM method gives faster classification results and achieves higher accuracy than other algorithms. The results also show the potential of the research direction of developing collaborative clustering/classification models for decentralized or distributed data analysis problems. In the future, we will develop collaborative data analysis algorithms applicable to large, distributed remote sensing datasets on cloud computing platforms.

Acknowledgements

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

References

[1] M. Chi, A. Plaza, J. A. Benediktsson *et al.*, "Big data for remote sensing: Challenges and opportunities", in *Proceedings of the IEEE*, Vol. 104, No. 11, pp. 2207-2219, Nov. 2016. DOI: 10.1109/JPROC.2016.2598228

[2] H. A. Selmy, H. K. Mohamed, and W. Medhat, "Big data analytics deep learning techniques and applications: A survey", *Information Systems*, Vol. 120, 102318, 2024. DOI: 10.1016/j.is.2023.102318

[3] F. Berloco, V. Bevilacqua, and S. Colucci, "Distributed analytics for big data: A survey", *Neurocomputing*, Vol. 574, 127258, 2024. DOI: 10.1016/j.neucom.2024.127258

[4] W. Pedrycz, "Collaborative fuzzy clustering", *Pattern Recognition Letters*, Vol. 23(14), pp. 1675-1686, 2002. DOI: 10.1016/S0167-8655(02)00130-7

[5] W. Pedrycz and P. Rai, "Collaborative clustering with the use of fuzzy c-means and its quantification", *Fuzzy Sets and Systems*, Vol. 159, Iss. 18, pp. 2399-2427, 2008. DOI: 10.1016/j.fss.2007.12.030

[6] Q. Du, B. Tang, W. Xie, and W. Li, "Parallel and distributed computing for anomaly detection from hyperspectral remote sensing imagery", in *Proceedings of the IEEE*, Vol. 109, No. 8, pp. 1306-1319, Aug. 2021. DOI: 10.1109/JPROC.2021.3076455

[7] Z. Wu, J. Sun, Y. Zhang, Z. Wei, and J. Chanussot, "Recent developments in parallel and distributed computing for remotely sensed big data processing", in *Proceedings of the IEEE*, Vol. 109, No. 8, pp. 1282-1305, 2021. DOI: 10.1109/JPROC.2021.3087029

[8] K. Feng and Y. Wu, "Distributed cloud computing architecture in hyperspectral remote sensing image classification under big data", *2022 IEEE 5th International Conference on Information Systems and Computer Aided Education (ICISCAE)*, 2022, pp. 456-460. DOI: 10.1109/ICISCAE55891.2022.9927624

[9] C. O'Reilly, A. Gluhak, and M. A. Imran, "Distributed anomaly detection using minimum volume elliptical principal component analysis", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28, Iss. 9, pp. 2320-2333, 2016. DOI: 10.1109/TKDE.2016.2555804

[10] Z. Li, X. Li, T. Liu, and J. Xie, "Towards building a distributed file system for remote sensing image process", *2016 IEEE International Conference on Big Data Analysis (ICBDA)*, 2016, pp. 1-4. DOI: 10.1109/ICBDA.2016.7509846

[11] Y. Wang, Z. Wang, P. Cheng *et al*., "DCM: A distributed collaborative training method for the remote sensing image classification", in *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 61, pp. 1-17, 2023, 5605018. DOI: 10.1109/TGRS.2023.3252544

[12] J. Xie, W. Yu, and G. Li, "An inter-agency collaborative computing framework for fast flood mapping using distributed remote sensing data", *Fifth International Conference on Agro-Geoinformatics*, 2016, pp. 1-5. DOI: 10.1109/Agro-Geoinformatics.2016.7577603

[13] H. Li and P. Tang, "Dps-MuSyQ: A distributed parallel processing system for multi-source data synergized quantitative remote sensing products producing", in *IEEE Access*, Vol. 8, pp. 79510-79520, 2020. DOI: 10.1109/ACCESS.2020.2989138

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

[15] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm", *Computers & Geosciences*, Vol. 10, Iss. 2, pp. 191-203, 1984. DOI: 10.1016/0098-3004(84)90020-7

[16] https://glovis.usgs.gov

[17] 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

ÁP DỤNG THUẬT TOÁN PHÂN CỤM C-MEANS MỜ BÁN GIÁM SÁT DỰA TRÊN MÔ HÌNH PHÂN CỤM CỘNG TÁC ĐỂ PHÂN LOẠI LỚP PHỦ ĐẤT TỪ ẢNH LANDSAT-7

Mai Đình Sinh[1], Nguyễn Tuấn Kiệt[2], Lê Chí Hiểu[2], Trịnh Lê Hùng[1]

[1]*Viện Kỹ thuật công trình đặc biệt, Trường Đại học Kỹ thuật Lê Quý Đôn*

[2]*Lớp Địa hình quân sự Khóa 56, Trường Đại học Kỹ thuật Lê Quý Đôn*

Tóm tắt: Sự phát triển nhanh chóng của các vệ tinh quan sát trái đất đã dẫn đến sự bùng nổ về các nguồn dữ liệu viễn thám. Việc lưu trữ tập trung các nguồn dữ liệu lớn ngày càng trở nên khó khăn, các giải pháp lưu trữ phi tập trung trên các hệ thống phân tán ngày càng được chú ý nhiều hơn. Các kỹ thuật 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, đa chiều và phân tán. Các tập dữ liệu này vì một vài lý do như bảo mật, đường truyền dữ liệu, tính riêng tư... nên không thể chia sẻ trực tiếp giữa các máy tính mà chỉ có thể chia sẻ thông tin về cấu trúc cụm. Bài báo trình bày một thuật toán phân cụm *c*-means mờ bán giám sát dựa trên mô hình phân cụm cộng tác trên các hệ thống phân tán áp dụng cho bài toán phân loại lớp phủ đất từ dữ liệu ảnh viễn thám. Mô hình đề xuất nhằm giải quyết vấn đề phân loại lớp phủ đất mà dữ liệu ảnh viễn thám phi tập trung được lưu trữ trên một hệ thống phân tán các máy tính được kết nối qua mạng. Các thử nghiệm trên bốn tập dữ liệu ảnh vệ tinh quang học cho thấy phương pháp đề xuất mang lại kết quả tốt hơn đáng kể cả về chất lượng phân loại và thời gian phân loại so với việc phân cụm cục bộ trên các tập dữ liệu riêng lẻ. Kết quả này cho thấy việc phát triển các thuật toán phân tích dữ liệu dựa trên mô hình cộng tác có thể giúp giải quyết tốt vấn đề phân tích dữ liệu ảnh viễn thám từ xa hoặc phân tán.

Từ khóa: *Phân loại lớp phủ; ảnh viễn thám; hệ thống phân tán; phân cụm cộng tác.*

Received: 01/10/2024; Revised: 21/12/2024; Accepted for publication: 27/12/2024

❑