Vietnam Journal of Science and Technology ***63*** *(1) (2025) 149-160*

doi:10.15625/2525-2518/19648

A novel distributed semi-supervised fuzzy clustering method applied on dental X-ray images

Le Tuan Anh[1, 2, 3], Tran Manh Truong[4], To Huu Nguyen[2], Nguyen Truong Thang[4], Nguyen Nhu Son[4, *]

[1]*Graduate University of Science and Technology, Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet, Cau Giay, Ha Noi, Vietnam*

[2]*University of Information and Communication Technology, Thai Nguyen University, Quyet Thang, Thai Nguyen City, Thai Nguyen, Vietnam*

[3]*Aritificial Intelligence Research Center, VNU Information Technology Institute, Vietnam National University, 144 Xuân Thuy, Cau Giay, Ha Noi, Vietnam*

[4]*Institute of Information Technology, Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet, Cau Giay, Ha Noi, Vietnam*

[*]Emails: *nnson@ioit.ac.vn*

Received: 13 December 2023; Accepted for publication: 19 September 2024

Abstract. Semi-supervised clustering methods are applied in various fields. However, these methods have the limitations when the size of data is large. Thus, distributed clustering models are proposed. By separating data set into some parts, distributed models overcome the challenges of big size dataset. In this paper, a novel model, based on semi-supervised fuzzy clustering and distributed mechanism, is proposed. In our research, the method to define distributed additional information is introduced. This additional information is used to implement distributed semi-supervised fuzzy clustering model on three datasets. The main contribution in this paper is the proposal of distributed semi-supervised fuzzy clustering algorithm (denoted as DSSFCM) that performs on Master – Slave model on homogeneous datasets.

Keywords: Distributed clustering, semi-supervised clustering, master - slave model, additional information, dental X-ray images.

Classification numbers: 4.7.2, 4.7.4

1. INTRODUCTION

Semi-supervised fuzzy clustering has been extensively utilized in many fields such as those mentioned in [1]. However, the advent of big data has presented both challenges and opportunities, particularly when dealing with distributed data across multiple sites. Traditional centralized processing algorithms, which have long been employed in these domains, struggle to effectively cope with the massive growth of data in such distributed scenarios.

When the size and the number of dimensions in data sets grow up, the storage space and processing power required to handle the data become limited [2]. Moreover, the centralized

architecture faces difficulties in harnessing the full potential of distributed data, often leading to suboptimal clustering results and compromised accuracy. Communication bottlenecks and network congestion further exacerbate the limitations of centralized algorithms.

To address these challenges, an effective algorithm is urgently needed that can solve semi-supervised fuzzy clustering problems with big distributed data. Although the literature on semi-supervised fuzzy clustering is rich, the majority of existing algorithms were originally conceived to process centralized data, overlooking the challenges posed by distributed environments. While there have been some distributed fuzzy clustering algorithms [3, 4], they have primarily focused on unsupervised learning scenarios, leaving a significant gap in the domain of distributed semi-supervised fuzzy clustering.

The advantages offered by distributed clustering are significant. It enables the analysis of large-scale datasets that surpass the capabilities of individual machines, facilitating efficient processing and scalability. Additionally, distributed clustering effectively handles geographically distributed data by enabling local analysis at each site, minimizing the need for extensive data transfer. This approach overcomes challenges related to data privacy, network constraints, and communication latency.

In [5], Lu *et al*. proposed dSimpleGraph model to overcome the issues caused by huge data in clustering. Based on MapReduce structure and micro cluster techniques, the framework and necessary algorithms in this model is presented. The implementations on the Open Cloud UIC cluster are performed. The results showed that time consuming and storage memory of this model are better than other compared models.

The distributed clustering is applied in the application of text collections [6]. In this paper, the authors introduced a novel clustering model named as Distributed Shared Nearest Neighbors clustering (D-SNN), a modification of shared nearest neighbor algorithm [7]. The proposed clustering algorithm with different approaches in parameter tuning are presented in this paper. The comparison among proposed algorithm and other related methods shows the higher performance of D-SNN, especially in handling high-dimension data. This is an ineffective algorithm in dealing with the noise. Moreover, the sensitivity to hyper-parameters is another disadvantage of D-SNN.

Several studies, such as those by Yasunori *et al*. [8], Zhang and Lu [9], and Khang *et al*. [10] have proposed effective semi-supervised fuzzy clustering algorithms for centralized data. However, these approaches may not directly address the complexities associated with big distributed data. The scarcity of efficient algorithms in this area inhibits the effective utilization of distributed data and hampers the accuracy and scalability of clustering results. Besides, while some distributed fuzzy clustering algorithms have been developed for unsupervised learning scenarios [3], limited efforts have been made to extend these algorithms to incorporate supervision and handle semi-supervised learning scenarios.

Distributed semi-supervised fuzzy clustering faces a significant challenge due to the lack of efficient algorithms tailored specifically for this task. While there have been notable advancements in semi-supervised fuzzy clustering and distributed fuzzy clustering separately, the integration of these two areas remains relatively unexplored. Existing literature predominantly focuses on centralized scenarios, with limited attention given to the unique requirements of distributed data in a semi-supervised setting.

Addressing this research gap requires a development of specialized algorithms that can effectively handle distributed semi-supervised fuzzy clustering. These algorithms should consider the challenges posed by distributed data, such as data privacy, communication latency,

and resource utilization optimization. By bridging this gap and developing efficient algorithms, researchers can unlock the potential of distributed semi-supervised fuzzy clustering and enable more accurate and robust analysis in various domains.

This study aims to propose a distributed approach for the Semi-Supervised Fuzzy C-Means (SSFCM) algorithm that can effectively handle the challenges associated with big distributed datasets. By leveraging the distributed nature of the data and incorporating semi-supervised learning principles, the proposed algorithm will address the limitations of traditional centralized processing algorithms. This research endeavors to contribute to the field by providing an effective solution for distributed semi-supervised fuzzy clustering, ultimately enabling improved clustering accuracy and scalability in the context of distributed data environments.

The structure of this paper is organized by the following sections. Section 2 presents some fundamental concepts and algorithms. The methodology and the details of our proposed model are introduced in Section 3. Section 4 shows the experimental environment and the results of implementing the new model on specific data sets. The last section includes the conclusions and the directions for further works.

2. PRELIMINARIES

2.1. Semi-Supervised Standard Fuzzy Clustering (SSSFC)

In general, a clustering problem is to classify n objects $X = \{X_k = (X_{k1},...,X_{kp})^T \in R^p, k = \overline{1,n}\}$ into c clusters $C = \{C_i, i = \overline{1,c}\}$ whose centers are $V = \{V_i = (V_{i1},...,V_{ip}) \in R^p, i = \overline{1,c}\}$ respectively, so that similarities between objects in the same cluster are higher than those between objects of different clusters. In the form of fuzzy clustering, each object can belong to more than one cluster. Hence $U = \{u_{ki} \mid u_{ki} \in [0,1], \sum_{i=1}^{c} u_{ki} = 1,\ k = \overline{1,n}\}$, in which u_{ki} represents the membership grade that X_k belongs to C_i.

Moreover, in a wide range of practical problems we may have additional information about membership grades of several data points. In that case, we have a semi-supervised fuzzy clustering problem in which $\overline{U} = \{\bar{u}_{ki} \in [0,1],\ i = \overline{1,c},\ k = \overline{1,n}\}$ denoting the supervised membership grades are given in advance. The additional information is not required to include membership grades for every couple of X_k and C_i, instead when information about $\bar{u}_{ki}$ is not given, its value is equal to zero. Thus, the constraints for $\bar{u}_{ki}$ are $\sum_{i=1}^{c} \bar{u}_{ki} \le 1, \forall k = \overline{1,n}$.

To address the aforementioned semi-supervised fuzzy clustering problem, Yasunori *et al*. [8] introduced the Semi-Supervised Standard Fuzzy Clustering (SSSFC) algorithm to find a solution that satisfies constraint for u_{ki} and minimize the following objective function.

$$J(U,V) = \sum_{k=1}^{n} \sum_{i=1}^{c} \left| u_{ki} - \bar{u}_{ki} \right|^m \left\| X_k - V_i \right\|^2 \to min \tag{1}$$

2.2. Pedrycz 's Semi-Supervised Fuzzy C-Mean algorithm (SSFCM)

The essence of SSFCM is its use of labeled data to guide iterative optimization. Below is the objective function proposed by Pedrycz [11] with $m = 2$.

$$J(U,V) = \sum_{k=1}^{N}\sum_{j=1}^{C} u_{kj}^2 \left\|X_k - V_j\right\|^2 + \sum_{k=1}^{N}\sum_{j=1}^{C} \left|u_{kj} - \bar{u}_{kj}\right|^2 \left\|X_k - V_j\right\|^2 \to min \quad (2)$$

The constraints of this optimal problem are: $\sum_{j=1}^{C} u_{kj} = 1, u_{kj} \in [0,1], \forall k = \overline{1,N}$.

SSFCM solves the above optimal problem by using Lagrange multiplier method to calculating *U* and *V*.

To find out final *U*, *V*, the process of re-calculating *U*, *V* will be repeated until the objective function satisfies a certain predetermined stopping condition.

3. PROPOSED DISTRIBUTED FUZZY SEMI-SUPERVISED CLUSTERING METHOD

In this part, the main idea of Distributed fuzzy semi-supervised clustering model is introduced in Section 3.1., Section 3.2 presents the method for defining the additional information used in fuzzy semi-supervised clustering. Lastly, model details and algorithms are given in Section 3.3.

3.1. The main idea of Distributed Semi-Supervised Fuzzy C-Mean clustering model (DSSFCM)

The Master-Slave framework of DFSSC is shown as in Figure 1. This model consists of two main components, including Master and Slave.

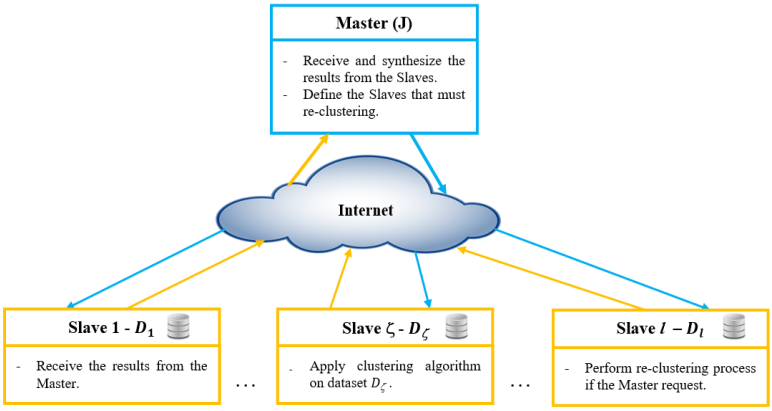

Figure 1. The framework of Master-Slave model.

In this model, Master is a computer using a simple algorithm to control the re-clustering progress of the Slaves. The Slaves are the computers in different locations where the data is stored. The Slaves apply SSFMC to determine distributed additional information in order to re-clustering in these Slaves.

In Figure 1, $D = \{D_i, i = \overline{1,l}\}$ is the distributed data set located in the Slaves. Each D_i has N_i data elements, $X^\zeta = \{X_1^\zeta, X_2^\zeta, ..., X_{N_\zeta}^\zeta\}$. The model aims to clustering the data D_i at the Slaves to enhance the clustering speed in big data problem.

At first, the Slave performs the clustering process using SSFCM to calculate *U, V* and cluster quality. The results are sent to Master. After getting the information from all Slaves, Master will calculate and make the decision of re-clustering if necessary. In other words, the algorithm used in Master is performed only when getting all the essential information from the Slaves. To solve this problem, the combination of *Pub/Sub* model and *REQ/REP* model to construct the communication model of this framework.

The briefly description of the main steps in distributed fuzzy semi-supervised clustering on Master-Slave model can be stated as below:

Step 1: Each Slave uses SSFCM [11] to calculate *U, V* and cluster quality. The clustering results in all Slaves have been sent to Master.

Step 2: After Master receives the results from all Slaves, it will find out the best quality clustering Slave to get the centers of clusters ${}_{best}V$. The chosen *V* is sent to all Slaves.

Step 3: Each Slave receives the ${}_{best}V$ from Master. They will use ${}_{best}V$ to define the new additional information (named as distributed additional information). This distributed additional information is used in SSFCM algorithm to perform re-clustering process on each Slave. Re-clustering progress will re-compute the centers of clusters *V* and new quality of clustering on Slaves. The results of re-clustering are also sent back to Master.

Step 4: At the Master, the result of the best quality clustering on Slaves is defined (new *Vbest*). Then, the Master make a comparison:

- If the quality of the clustering of the t^{th} iteration is better than that of the $(t-1)^{th}$ iteration following the threshold epsilon (ε) or the number of iteration is equal to or greater than pre-defined maximum steps, the algorithm will be stopped.
- Otherwise, if the quality of the clustering of the t^{th} iteration is not good as that of the $(t-1)^{th}$ iteration following the threshold epsilon (ε) and the number of iterations is smaller than pre-defined maximum steps, Master will send the new *Vbest* to all Slaves. The model will go back to *Step 3*.

3.2. The detail of formulation for distributed additional information based on SSFCM algorithm

In [1, 11], the formula of objective function is given by equation (3).

$$J(U,V)=\sum_{k=1}^{N}\sum_{j=1}^{C}u_{kj}^{m}\left\|X_k-V_j\right\|^2+\sum_{k=1}^{N}\sum_{j=1}^{C}\left|u_{kj}-\bar{u}_{kj}\right|^m\left\|X_k-V_j\right\|^2\rightarrow min \quad (3)$$

The constraints of this optimal problem are:

$$\sum_{j=1}^{C}u_{kj}=1, u_{kj}\in[0,1], \forall k=\overline{1,N} \quad (4)$$

To solve the problems (3) and (4) by Lagrange multiplier method with the value of *m* as *2* and this value is applied in all following equations. The centers of clusters and the membership degrees are defined by:

$$V_j=\frac{\sum_{k=1}^{N}u_{kj}^{m}X_k+\sum_{k=1}^{N}\left|u_{kj}-\bar{u}_{kj}\right|^m X_k}{\sum_{k=1}^{N}u_{kj}^{m}+\sum_{k=1}^{N}\left|u_{kj}-\bar{u}_{kj}\right|^m} \quad (5)$$

$$u_{kj} = \frac{1}{2}\left(\frac{2 - \sum_{i=1}^{C} \bar{u}_{ki}}{\sum_{i=1}^{C} \frac{\left\| X_k - V_j \right\|^2}{\left\| X_k - V_i \right\|^2}} + \bar{u}_{kj} \right) \tag{6}$$

The equations for V and u on the ζ^{th} Slave are:

$${}_{\zeta}V_j = \frac{\sum_{k=1}^{N_\zeta} {}_{\zeta}u_{kj}^m \, {}_{\zeta}X_k + \sum_{k=1}^{N_\zeta} \left| {}_{\zeta}u_{kj} - {}_{dis\zeta}\bar{u}_{kj} \right|^m \, {}_{\zeta}X_k}{\sum_{k=1}^{N_\zeta} {}_{\zeta}u_{kj}^m + \sum_{k=1}^{N_\zeta} \left| {}_{\zeta}u_{kj} - {}_{dis\zeta}\bar{u}_{kj} \right|^m} \tag{7}$$

$${}_{\zeta}u_{kj} = \frac{1}{2}\left(\frac{2 - \sum_{i=1}^{C} {}_{dis\zeta}\bar{u}_{ki}}{\sum_{i=1}^{C} \frac{\left\| X_k^{\zeta} - {}_{\zeta}V_j \right\|^2}{\left\| X_k^{\zeta} - {}_{\zeta}V_i \right\|^2}} + {}_{dis\zeta}\bar{u}_{kj} \right) \tag{8}$$

where: m is fuzzifier; C is the number of clusters; ζ is ζ^{th} Slave, $\zeta = \overline{1,l}$; l is the number of Slaves; N_ζ is the number of data elements on the ζ^{th} Slave; ${}_{\zeta}u_{kj}$ is membership degree of X_k data element to the j^{th} cluster on ζ^{th} Slave; ${}_{dis\zeta}\bar{u}_{kj}$ is additional membership function of X_k data element to the j^{th} cluster on ζ^{th} Slave; $X_k^{\zeta} \in R^r$ is k^{th} data element of $X^{\zeta} = \{X_1^{\zeta}, X_2^{\zeta}, ..., X_{N_\zeta}^{\zeta}\}$ on ζ^{th} Slave; r is the number of data dimensions; ${}_{\zeta}V_j$ is the cluster center j on the ζ^{th} Slave; ${}_{\zeta}q$ is the clustering quality on the ζ^{th} Slave

The constraints include, $\forall k = \overline{1, N_\zeta}$:

$$\sum_{j=1}^{C} {}_{\zeta}u_{kj} = 1; {}_{\zeta}u_{kj} \in [0,1]; \sum_{j=1}^{C} {}_{dis\zeta}\bar{u}_{kj} \le 1 \tag{9}$$

To measure the quality of clustering on Slaves, the *DB* (Davies – Bouldin) measurement is used. The value of *DB* is defined by:

$${}_{\zeta}q = DB = \frac{1}{C}\sum_{i=1}^{C} \max_{i \ne j}\left(\frac{{}_{\zeta}\sigma_i - {}_{\zeta}\sigma_j}{\left\| {}_{\zeta}V_i - {}_{\zeta}V_j \right\|} \right) \tag{10}$$

To build a distributed additional membership function for DSSFCM model, we perform the following tasks:

1) After the Slaves receive all essential information from the Master including the data type *X*, the number of cluster centers *C* and fuzzifier *m* = 2. We run the FCM [12] algorithm to calculate the membership degree according to the equation:

$${}_{\zeta}u_{kj} = \frac{1}{\sum_{i=1}^{C} \frac{\left\| X_k^{\zeta} - {}_{\zeta}V_j \right\|^2}{\left\| X_k^{\zeta} - {}_{\zeta}V_i \right\|^2}} \tag{11}$$

Using equation (11), additional membership degrees in equation (6) on ζ^{th} Slave is calculated as below:

$$\overset{-0}{{}_{\zeta}u_{kj}} = \begin{cases} \max({}_{\zeta}u_{kj}), j = \overline{1,C} \\ 0, otherwise \end{cases} \quad (12)$$

Calculating *u*, *V* and *DB* for each Slave according to equations (6), (7) and (10), we get ${}_{\zeta}u_{kj}$, ${}_{\zeta}V_j$ and ${}_{\zeta}q$ at the the ζ^{th} Slave. The Slave will then send its ${}_{\zeta}V_j$ and ${}_{\zeta}q$ to the Master.

2) At Master having ${}_{\zeta}q, \zeta = \overline{1,l}$, as in equations(7) and (8) received from Slaves, we get:

$$q_{\min} = \min\{{}_{\zeta}q, \zeta = \overline{1,l}\} \Rightarrow {}_{best}V_j, \forall j = \overline{1,C} \quad (13)$$

From equation (13), we identify the Slave with the best *DB* along with its ${}_{\zeta}V_j$ as ${}_{best}V_j$. The ${}_{best}V_j$ will be sent to the Slaves to perform to recalculate the distributed additional membership function ${}_{\text{dis}\zeta}\bar{u}_{kj}$ for each Slave based on the dataset at that Slave.

3) At Slave, after receiving ${}_{best}V_j$ from Master, ${}_{\text{dis}\zeta}\bar{u}_{kj}$ is calculated as below:

$$\overset{-t}{{}_{dis\zeta}u_{kj}} = \frac{1}{2}\left(\frac{2 - \sum_{i=1}^{C} {}_{\zeta}\bar{u}_{ki}}{\sum_{i=1}^{C} \frac{\left\| X_k^{\zeta} - {}_{dis}V_j \right\|^2}{\left\| X_k^{\zeta} - {}_{\zeta}V_i \right\|^2}} + {}_{dis\zeta}\bar{u}_{kj} \right) \quad (14)$$

where ${}_{dis\zeta}\bar{u}_{kj}^{t}$ is the distributed additional membership degree of X_k data element to the j^{th} cluster on ζ^{th} Slave at t^{th} iteration ($t \geq 1$).

Recalculate ${}_{\zeta}V_j$, ${}_{\zeta}u_{kj}$ and ${}_{\zeta}q$ according to equations(7), (8), (10) and(14), then send ${}_{\zeta}V_j$ and ${}_{\zeta}q$ return to Master.

3.3. Model details and algorithms

The framework of the algorithm is given as in Figure 2.

There are two algorithms to solve the optimal problem. In which, one algorithm is performed on the Slaves and one algorithm performed on the Master.

Algorithm on the Slaves:

Algorithm 1 (*SSFCM – Table 1*): to calculate ${}_{dis\zeta}\bar{u}_{kj}^{t}$, ${}_{\zeta}V_j$ and ${}_{\zeta}q$ at ζ^{th} Slave ($\zeta = \overline{1,l}$).

The input: The data set ${}_{\zeta}X$ including ${}_{\zeta}N$ elements, number of clusters *C*, fuzzifier *m* = 2, additional information function $\bar{U}$, threshold ε_1, maximum iterations *maxStep1* > 0.

The output: Cluster center ${}_{\zeta}V$, cluster quality measurement ${}_{\zeta}q$.

Table 1. The steps of Algorithm 1 (SSFCM).

Steps	Tasks
1.	$t_1 = 0$
2.	Random initialization ${}_{\zeta}V_j^{(t_1)}; (j = \overline{1,C})$
3.	Calculate ${}_{\zeta}\bar{u}_{kj}$ using equation (10) or calculate ${}_{dis\zeta}\bar{u}_{kj}^{t}$ using equation (14)
4.	Repeat
5.	$t_1 = t_1 + 1$
6.	Calculate ${}_{\zeta}u_{kj} (k = \overline{1, N_{\zeta}}, j = \overline{1,C})$ using equations (6) or (8)
7:	Calculate ${}_{\zeta}V_j^{(t_1)}; (j = \overline{1,C})$ using equations (5) or (7)
8.	Until $\left\| {}_{\zeta}V_j^{(t_1)} - {}_{\zeta}V_j^{(t_1-1)} \right\| \le \varepsilon_1$ or $t_1 > maxStep1$
9.	Calculate ${}_{\zeta}q$ using equation (10)

Algorithm on Master:

Algorithm 2 *(DSSFCM* – Table 2*)*: an proposed algorithm launched by Master when it received ${}_{\zeta}V_j$ and ${}_{\zeta}q$ (where $\zeta = \overline{1,l}$) from Slaves. Then, calculate whether the Slaves need to perform the re-clustering process or not.

The input: Number of Slaves l, cluster center ${}_{\zeta}V$, cluster quality q, threshold ε_2, maximum iterations $maxStep2 > 0$.

The output: Cluster center ${}_{best}V_j$ of Slave which has the highest cluster quality.

Table 2. The steps of Algorithm 2 (DSSFCM).

Steps	Tasks
1.	$t = 0$
2.	Repeat
3.	Call Algorithm1
4.	Calculate $q_{\min}^{(t)}$ using equation (13) to determine ${}_{best}V_j$
5.	$t = t + 1$
6.	Until $\left\| q_{\min}^{(t)} - q_{\min}^{(t-1)} \right\| \le \varepsilon_2$ or $t > maxStep2$

4. RESULTS AND DISCUSSION

4.1. Data description

To evaluate the proposed model, two UCI datasets are used, including IRIS and WINE. Apart from that, a set of 680 dental X-ray images is also implemented.

The dental X-ray image dataset is collected from 2019 to 2021 at Hanoi Medical University only for research purpose. These images are periapical and panorama images. In image pre-

processing step, the features of these images are extracted [13] including Entropy, edge-value and intensity features, Local Binary Patterns, Gradient and Patch level feature.

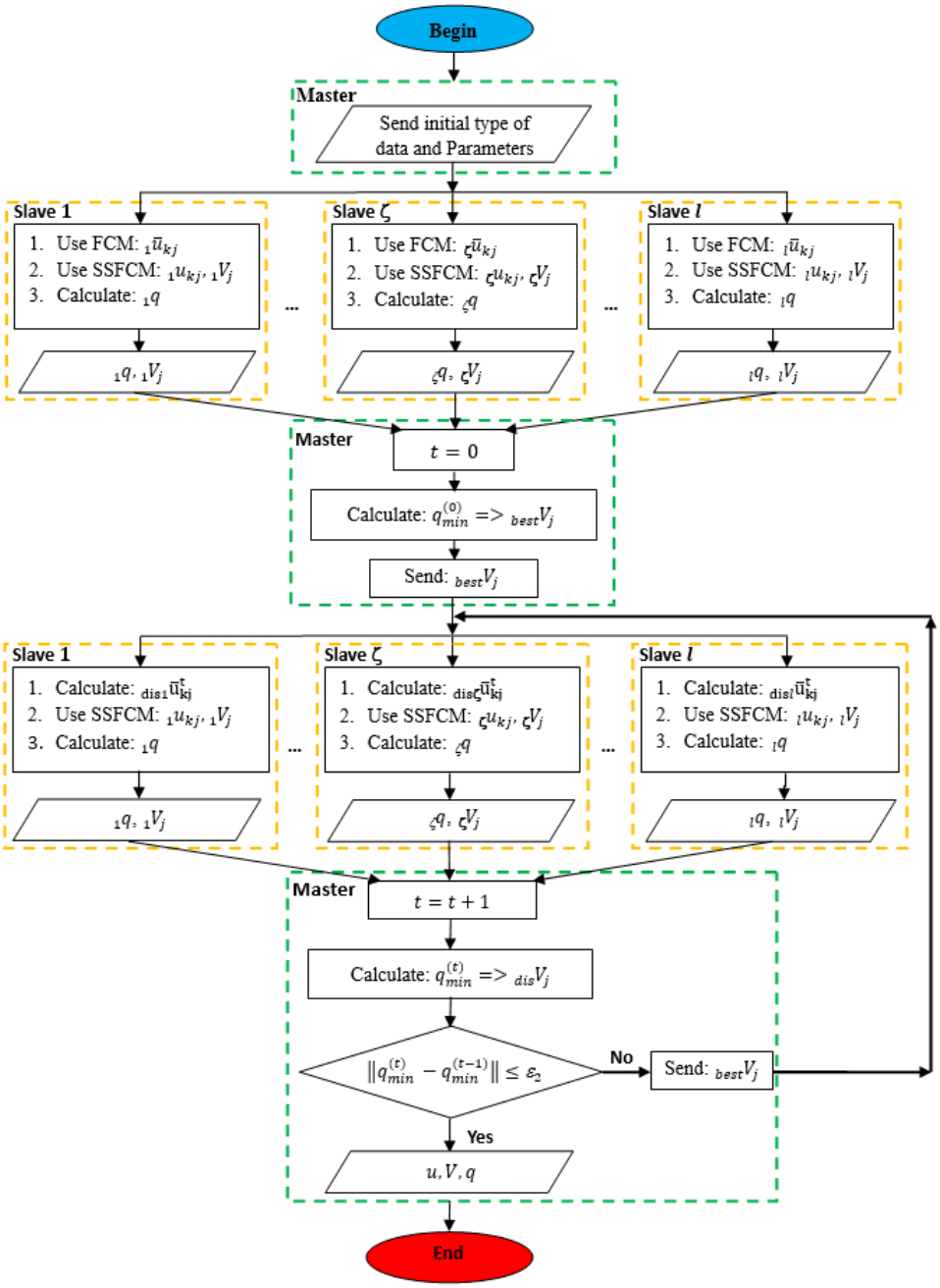

Figure 2. Framework of DSSFCM on Master – Slave model.

Experiment environment: The implementation of proposed model and two other related models, FCM and SSFCM, are performed by Python on the computers with the configuration shown in Table 3.

To get the results of clustering, FCM and SSFCM need one computer to run the experiments while DSSFCM needs 6 computers, including 1 Master and 5 Slaves. The combination of *Pub/Sub* and *REQ/REP* models in ZMQ library is applied in order to establish the communication mechanism between Master and Slaves.

Table 3. The configuration of computers used in experiments.

Specification	Value
Cores	4
Memory(GB)	12
Operating system	Windows 10

Validity Indices: The comparison among FCM, SSFCM and DSSFCM is performed by using *DB* measurement and time consuming.

4.2. Experimental results and discussions

There are two scenarios in our experiments as below.

Scenario 1: Two UCI datasets are implemented by using FCM and SSFCM. Two these datasets are divided into 5 parts to run on 5 Slaves.

The results of Scenario 1 are presented in Table 4. The sign '-' means that the lower values are better. Thus, bold values are the best values.

Table 4. The experiment results of FCM, SSFCM and DSSFCM on UCI datasets.

Method / Data	DB-			Time (second)-		
	FCM	***SSFCM***	***DSSFCM (5 Slaves)***	***FCM***	***SSFCM***	***DSSFCM (5 Slaves)***
IRIS	0.7623	0.7561	**0.6015**	**0.14**	0.24	0.35
Wine	0.9209	0.9158	**0.9053**	47.82	56.80	**44.07**

From the results in Table 4, on UCI datasets, DSSFCM model achieves better performance in term of *DB* measurement. However, by using *DB*, DSSFCM is not good on small size dataset. The value of *DB* on IRIS is much smaller than that on WINE. The running time on WINE of DSSFCM is better than two other related models. Thus, DSSFCM is suitable for big datasets in both *DB* and time consuming.

Scenario 2: There are three smaller cases in this scenario.

- Dental dataset is divided into 3 parts. These parts are implemented separately on 3 Slaves.
- Dental dataset is divided into 4 parts. These parts are implemented separately on 4 Slaves.
- Dental dataset is divided into 5 parts. These parts are implemented separately on 5 Slaves.

The results of Scenario 2 are presented in Table 5.

Table 5. The experiment results of DSSFCM on dental X-ray images
(Bold values are the best values).

Validity indices / Number of Slaves	DB -	Time (second)-
3	0.8416	1.40
4	0.5883	0.96
5	**0.5353**	**0.84**

As shown in Table 5, on Dental dataset, DSSFCM gets the best values of *DB* and time consuming in the case of 5 Slaves. However, in this case, it needs a lot of resources for completing the implementations.

Besides, based on the proposed algorithm, the total amount of data exchanged between the Slaves and the Master is very small including three parts. The first part, initialized data, is the fixed amount of data that Master needs to send to Slaves. Slaves use this data as input information for executing the program. The second part is the amount of data that Slaves send to Master in form of the pairs ($_{\zeta}q$, $_{\zeta}V_j$). The third part, data sent to Slaves from Master, is the value $_{best}V_j$. For example, the total amount of exchanged data between Master and three Slaves on IRIS dataset is about 2160 bytes.

5. CONCLUSIONS AND FUTURE WORK

In this research, we focus on constructing and describing a novel distributed clustering model based on fuzzy semi-supervised learning. The way to define distributed additional information is presented in this paper.

The main contributions can be stated by:

1) To propose a distributed fuzzy semi-supervised clustering model, named as DSSFCM;
2) To present the method for determining distributed additional information in order to apply DSSFCM on UCI and Dental datasets.
3) To evaluate proposed model by *DB* index and time consuming. The performance of DSSFCM is assessed by comparing this model to other related models.

In further works, other scenarios will be considered to show the efficiency of DSSFCM. The applications of DSSFCM on different datasets are also performed.

Acknowledgements.We are grateful for the support from the Grant CS23.03 funded by the Institute of Information Technology, Vietnam Academy of Science and Technology.

CRediT authorship contribution statement.Le Tuan Anh: Methodology, Model design, Code; Tran Manh Truong: Model design,Funding acquisition, Supervision; To Huu Nguyen: Code; Nguyen Truong Thang: Supervision; Nguyen Nhu Son: Methodology. All authors contributed to the manuscript revision, and read andapproved the submitted version.

Declaration of competing interest.The authors proclaim that they have no conflicts of interest to report concerning the present study.

REFERENCES

1. Lai D. T. C., Miyakawa M., and Sato Y. - Semi-supervised data clustering using particle swarm optimisation, Soft Comput **24** (2020) 3499-3510. https://doi.org/10.1007/s00500-019-04114-z.
2. Ganesan D., Estrin D., and Heidemann J. - DIMENSIONS: Why do we need a new data handling architecture for sensor networks? ACM SIGCOMM Computer Communication Review **33** (3) (2003) 143-148. https://dl.acm.org/doi/abs/10.1145/774763.774786.
3. Karthikeyani Visalakshi N. , Thangavel K., and Parvathi R.- An intuitionistic fuzzy approach to distributed fuzzy clustering, International Journal of Computer Theory and Engineering **2** (4) (2010) 295-302.
4. Son L. H. - DPFCM: A novel distributed picture fuzzy clustering method on picture fuzzy sets, Expert Systems with Applications **42** (3) (2015) 51-66. https://doi.org/10.1016/j.eswa.2014.07.026.
5. Lu L., Gu Y., and Grossman R. - dSimpleGraph: a novel distributed clustering algorithm for exploring very large scale unknown data sets, In: 2010 IEEE International Conference on Data Mining Workshops, 2010, pp. 162-169. IEEE. https://doi.org/10.1109/ICDMW.2010.12.
6. Zamora J., Héctor A. C., and Marcelo M. - Distributed clustering of text collections, IEEE Access 7, 2019, pp. 155671-155685. https://doi.org/10.1109/ ACCESS.2019. 2949455.
7. Ravichandran M., Subramanian K. M., Ganesan P., and Jothikumar R. - A modified method for high dimensional data clustering based on the combined approach of shared nearest neighbor clustering and unscented transform, J. Comput. Theor. Nanosci. **15** (8) (2018),pp. 2050-2054. https://doi.org/10.1166/jctn.2018.7405.
8. Yasunori E., Yukihiro H., Makito Y., and Sadaaki M. - On semi-supervised fuzzy c-means clustering, In 2009 IEEE International Conference on Fuzzy Systems, 2009, pp. 1119-1124. https://doi.org/10.1109/FUZZY.2009.5277177.
9. Zhang H., Lu J. - Semi-supervised fuzzy clustering: A kernel-based approach, Knowledge-Based Systems **22** (8) (2009) 477-481. https://doi.org/10.1016/j.knosys.2009.06.009.
10. Khang T. D., Tran M. K., and Fowler M. - A novel semi-supervised fuzzy c-means clustering algorithm using multiple fuzzification coefficients, Algorithms **14** (11) (2021) 258-269. https://doi.org/10.3390/a14090258.
11. Pedrycz W. - Algorithms of fuzzy clustering with partial supervision, Pattern recognition letters **3** (3) (1985) 13-20. https://doi.org/10.1016/0167-8655(85)90037-6.
12. Bezdek J. C. - Pattern recognition with fuzzy objective function algorithms, Springer Science & Business Media, 2013.
13. Tuan T. M., Minh N. H., Van Tao N., Ngan T. T., and Huu N. T. - Medical diagnosis from dental X-ray images: A novel approach using Clustering combined with Fuzzy Rule-based systems, In: 2016 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS), 2016, pp. 1-6. IEEE. https://doi.org/10.1109/NAFIPS.2016.7851622.