Journal of Intelligent & Fuzzy Systems 43 (2022) 5661–5672
DOI:10.3233/JIFS-212812
IOS Press

A label-guided weighted semi-supervised neutrosophic clustering algorithm

Dan Zhang[a], Yingcang Ma[a,*], Hengdong Zhu[b] and Florentin Smarandache[c]
[a]*School of Science, Xi'an Polytechnic University, Xi'an, China*
[b]*Department of Public Basic Courses, Hunan Institute Of Traffic Engineering, Hengyang, Hunan*
[c]*Mathematics and Science Division, Gallup Campus, University of New Mexico, Gallup, NM, USA*

Abstract. The traditional neutrosophic clustering method only performs cluster analysis on the data itself, and often ignores the supervision information of data. In order to solve the above problems, a label-guided weighted semi-supervised neutrosophic clustering algorithm is proposed in the paper. On the one hand, the paired constraint information is used to construct the supervision weight coefficient and the distance measurement learning is combined to re-measure the degree of membership of the data and the cluster center; On the other hand, by minimizing the sum of squares of error between membership matrix and label matrix, the purpose of clustering results guided by label information is realized. Experiments on various data sets and comparisons with other clustering algorithms show that the new clustering algorithm can make full use of supervisory information and improve the accuracy of clustering.

Keywords: Semi-supervised clustering, label information, neutrosophic set, clustering

1. Introduction

Clustering plays a crucial role in image processing [1–3], data mining [4], machine learning [5] and other fields. Traditional clustering algorithms include hard clustering and fuzzy clustering. For example, the k-means [6] algorithm is susceptible to an initial value and cannot process data sets with overlapping regions. The fuzzy c-means(FCM) [7] algorithm uses membership matrix to get the membership degree of data points belonging to each cluster, which can well divide overlapping regions, however the FCM is too sensitive to outliers and noise points. Literature [8] proposed a robust sparse fuzzy k-means algorithm (RSFKM), which introduced robust functions to deal with outliers and noise points, so as to enhance the robustness and sparsity of the FCM algorithm. Wang [9] proposed an extreme clustering method, which overcomes the shortcomings of the peak clustering algorithm on the basis of retaining the advantages of the peak clustering algorithm, and has strong robustness to unevenly distributed data sets. Tong [10] proposed a clustering algorithm based on density peaks, which can automatically determine the number of clusters. Chang [11] proposed a novel compound rank- k projection (CRP) algorithm for bilinear analysis. The CRP algorithm deals with matrices directly, thus preserving the correlations within the matrix and decreasing the computation complexity. Bui [12] proposed the F-Mapper algorithm, based on the foundation of the Mapper algorithm, to solve the problem of automating when dividing cover intervals with an arbitrary percentage of overlap. Bui [13] proposed a new method called the shape fuzzy c-means (SFCM) algorithm, and the SFCM algorithm can not only exhibit the same clustering ability as the FCM but also reveal

[*]Corresponding author. Yingcang Ma, School of Science, Xi'an Polytechnic University, Xi'an, China. E-mail: mayingcang@xpu.edu.cn.

some relationships through visualizing the global shape of data supplied by the Mapper. These algorithms have effectively improved the performance of the FCM, but they have some limitations for fuzzy clustering of uncertain problems.

In order to solve the problems of fuzzy clustering, the neutrosophic set theory proposed by Smarandache in 1993 [14] is welcomed by scholars, and the neutrosophic clustering algorithm based on neutrosophic theory [15–17] is widely used in various fields. Guo [18] proposed the neutrosophic c-means clustering algorithm (NCM) based on the neutrosophic set framework and the FCM, which can be effectively distinguish sample points, boundary points and outliers in clusters. The membership degree of the NCM is not affected by noise, which effectively solves the problem that the FCM cannot detect not typical data points. Ye [19] proposed a single-valued neutrosophic minimum spanning tree clustering algorithm by defining a generalized single-valued neutrosophic set distance measure, which shows great superiority in the clustering of single-valued neutrosophic observation data. Kandasamy [20] proposed a double-valued neutrosophic minimum spanning tree clustering algorithm to cluster data represented by doublel-valued neutrosophic information. Thong [21] proposed an image fuzzy clustering algorithm (FC-PFS) based on image fuzzy sets. The algorithm needs to calculate three matrices, so it has high computational complexity for high-dimensional data. Li [22] proposed a single-value neutrosophic clustering algorithm based on Tsallis entropy maximization based on image fuzzy set clustering and single-value neutrosophic set. The algorithm showed satisfactory results in image segmentation. According to the density characteristics of data points, Rashno [23] proposed a neutrosophic clustering algorithm based on data uncertainty. This method deals with the membership degree of boundary points and outliers more accurately. Akbulut [24] proposed the kernel neutrosophic c-means clustering(KNCM), which uses nonlinear transformation to map the inseparable nonlinear data in the low-dimensional space to the separable high-dimensional linear feature space for clustering. It has good clustering effect.

The above algorithms do not rely on any prior information, and mainly performs cluster analysis on a datasets itself, and ignores a small volume of labeled information on the dataset, which belongs to unsupervised learning. On the contrary, supervised learning requires a large volume of label information. However, in real applications, some datasets often contain only a small volume of label information, which cannot support supervised learning. Semi-supervised learning [25], as a learning method that combines supervised learning and unsupervised learning, can use labeled data and unlabeled data at the same time, and rationally use more or less labeled information in the data.

Semi-supervised clustering is a combination of supervised information and clustering algorithms, which can guide the data clustering process more effectively. Studies have found that high-quality supervision information can not only guide unlabeled samples to correctly cluster to improve the accuracy of the clustering results, but also speed up the convergence of clustering [26, 27]. Scholars have proposed various semi-supervised clustering algorithms. Fariba [28] used the entropy and relative entropy divergence measure to replace the fuzzy measure of the unsupervised segment and the geometric distance measure of the semi-supervised segment respectively, emphasizing the explicit combination of unsupervised and semi-supervised segments. Zhuang [29] explicitly incorporated the labeled information into the graph learning method by forcing the edge weight between different types of labeled samples to zero. Nie [30] proposed using elastic embedding constraints on the predictive label matrix to better explore the manifold structure and effectively learn normal data under Gaussian distribution. Zhu [31] improved the clustering effect by combining the label information matrix and the manifold regular term, and proposed an adaptive semi-supervised neighborhood clustering algorithm. Chen [32] proposed a semi-supervised deep model that not only addresses the problem in multimodal sensor data, but also addresses the imbalanced distribution of labeled data across categories at the same time.

At present, most of the neutrosophic clustering algorithms are unsupervised neutrosophic clustering algorithms [33–35], which cannot adjust the membership matrix by using labeled information to obtain a membership matrix more in line with the real situation, thus affecting the clustering effect. To solve this problem, a label-guided weighted semi-supervised intermediate-intelligence clustering algorithm (LG-WSSNCM) is proposed in this paper. Semi-supervised learning is introduced into the neutrosophic clustering algorithm, and supervised information is used to optimize the neutrosophic clustering algorithm to improve the clustering performance.

2. Related work

2.1. Neutrosophic clustering method

Neutrosophic c-means (NCM) [10] defines the degree of membership of deterministic and uncertain clusters, so that it can handle the noise or outliers contained in the data set itself, and proposes solving the following convex optimization problems:

$$J_{NCM}(T, I, F, C) = \sum_{i=1}^{N}\sum_{j=1}^{C}\left(\varpi_1 T_{ij}\right)^m \left\|x_i - c_j\right\|_2^2 + \sum_{i=1}^{N}(\varpi_2 I_i)^m \left\|x_i - c_{i\max}\right\|_2^2 + \delta^2 \sum_{i=1}^{N}(\varpi_3 F_i)^m, \tag{1}$$

where m is a constant. T_{ij}, I_i, and F_i are true membership, uncertain membership and false membership respectively. x_i is the i-th data point, c_j is the j-th cluster center. ϖ_1, ϖ_2 and ϖ_3 are constants, which belong to [0,1]. Define $0 < T_{ij}, I_i, F_i < 1$ to satisfy the following constraints:

$$\sum_{j=1}^{C} T_{ij} + I_i + F_i = 1. \tag{2}$$

For each data $\mathrm{x_i}$, use the cluster center with the largest and second largest $\mathrm{T_{ij}}$ values to calculate $c_{i\max}$:

$$c_{i\max} = \frac{c_{pi} + c_{qi}}{2}, \tag{3}$$

$$\begin{cases} p_i = \underset{j=1,2,\ldots C}{\arg\max}(T_{ij}), \\ q_i = \underset{j \neq p_i \cap j=1,2,\ldots C}{\arg\max}(T_{ij}) \ . \end{cases} \tag{4}$$

δ^2 is defined as follows:

$$\delta^2 = \lambda \frac{\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{C}\left\|x_i - c_j\right\|_2^2}{NC}. \tag{5}$$

2.2. Supervision information

Supervision information in the clustering process usually refers to the constraint information of the sample. In semi-supervised clustering, it generally contains in two types:

(1) Sample label information: usually refers to the information with labeled data in the supervised learning algorithm. In most cases, the labeled data set is far less than the unlabeled data set, and its main function is auxiliary clustering.

(2) Paired constraint information: usually refers to the information about whether the data is in the same cluster. Literature [36, 37] proposes two commonly used paired constraint information: $must - link$ (ml) and $cannot - link$ (cl). For points x_i and x_j, if they belong to cluster C_m and C_n, the paired constraint information is defined as: if $m = n$, then x_i and x_j belong to the same cluster, then $\left(x_i, x_j\right) \in must - link$; otherwise, if $m \neq n$, then it means x_i and x_j do not belong to the same cluster, then $\left(x_i, x_j\right) \notin cannot - link$. Paired constraint information has the properties of transitivity and consistency.

3. Label-guided weighted semi-supervised neutrosophic clustering algorithm

3.1. Define supervision information

First, giving the definition of the supervision information used in this paper. Let $X \in R^{N\times d}$ be a set of n objects, the number of clusters is c. Assuming that the label information of l data points is known, the data point x_i belongs to the j-th cluster, where $i = 1, 2, ..., N$ and $j = 1, 2, ..., C$, define $y_{ij} = 1$ in the label vector, and the rest take the value 0, the definition $Y = \begin{cases} Y_l \\ Y_u \end{cases} \in R^{N\times C}$ of the label matrix can be obtained, and$Y_u$is the label matrix corresponding to the unlabeled data, and it is a zero matrix.

According to the label matrix, the paired constraint information can be obtained, that is, whether the data x_i and the center c_j are of the same cluster. It is defined as follows:

$$Pcl = \begin{cases} 1, & x_i \in Y_l \cap y_{ij} = 1; \\ 0, & others; \\ -1, & x_i \in Y_l \cap y_{ij} = 0. \end{cases} \tag{6}$$

where $Pcl \in R^{N\times C}$, the value of Pcl_{ij} is 1 or -1 stands for $\left(x_i, c_j\right) \in must - link$ or $\left(x_i, c_j\right) \notin cannot - link$. Therefore, the definition of weight matrix A is as follows:

$$A = \begin{cases} a, & Pcl_{ij} = 1; \\ 1, & Pcl_{ij} = 0; \\ 1/a, & Pcl_{ij} = -1. \end{cases} \tag{7}$$

where a is a constant and $0 < a < 1$.

3.2. Proposed method

It can be seen from model (1) that the distance reflects the degree of membership. A smaller distance $\left\|x_i - c_j\right\|_2^2$ corresponds to a larger degree of membership, indicating that the data point x_i is easy to be classified into the cluster corresponding to c_j; on the contrary, a larger distance $\left\|x_i - c_j\right\|_2^2$ corresponds to a smaller degree of membership, indicating that the data point x_i is not easy to be divided into the cluster corresponding to c_j. Therefore, if the paired constraint information of x_i and c_j is known, a new distance measurement function is constructed through the weight coefficient A_{ij}. Specifically, if $\left(x_i, c_j\right) \in must - link$ is known, the new distance $A_{ij}\left\|x_i - c_j\right\|_2^2$ is made smaller by applying a small weighting coefficient, thereby enlarging the value of membership; on the contrary, $\left(x_i, c_j\right) \notin cannot - link$, by applying a large weight coefficient, makes the new distance $A_{ij}\left\|x_i - c_j\right\|_2^2$ larger, thereby reducing the value of the degree of membership.

Different from the traditional least squares method $\|T - Y\|_F^2$, in order to ensure the consistency of the labeled data's label matrix and the clustering result, and to avoid the membership degree $T_u \rightarrow 0$ of the unlabeled data, the label matrix is used to construct diagonal matrix U to impose the least squares constraint on the labeled data. The diagonal elements $U_{ii} = 1$ correspond to the l labeled data, and the value of the diagonal elements of unlabeled data is 0.

From the above analysis, let $m = 2$, our optimization model is as follows:

$$
\begin{aligned}
J(T, I, F, C) = & \sum_{i=1}^{N}\sum_{j=1}^{C}\left(\varpi_1 T_{ij}\right)^2 A_{ij}\left\|x_i - c_j\right\|_2^2 \\
& + \alpha tr\left[(T-Y)^T U (T-Y)\right] \\
& + \sum_{i=1}^{N}(\varpi_2 I_i)^2 \left\|x_i - c_{i\max}\right\|_2^2 \\
& + \delta^2 \sum_{i=1}^{N}(\varpi_3 F_i)^2, \\
s.t. \quad & \sum_{j=1}^{C} T_{ij} + I_i + F_i = 1.
\end{aligned}
\tag{8}
$$

Where T_{ij}, I_i, and F_i are the membership value of determined data points, boundary data points and noise data points respectively; Y and U are label matrices; A_{ij} is weight coefficient; ϖ_1, ϖ_2, ϖ_3 and $\alpha > 0$ are constants.

Equation (8) is solved by alternating optimization method to construct its Lagrangian function:

$$
\begin{aligned}
L(T, I, F, C, \lambda) = & \sum_{i=1}^{N}\sum_{j=1}^{C}\left(\varpi_1 T_{ij}\right)^2 A_{ij}\left\|x_i - c_j\right\|_2^2 \\
& + \alpha tr\left[(T-Y)^T U (T-Y)\right] \\
& + \sum_{i=1}^{N}(\varpi_2 I_i)^2 \left\|x_i - c_{i\max}\right\|_2^2 \\
& + \delta^2 \sum_{i=1}^{N}(\varpi_3 F_i)^2 \\
& - \sum_{i=1}^{N}\lambda_i\left(\sum_{j=1}^{C} T_{ij} + I_i + F_i - 1\right).
\end{aligned}
\tag{9}
$$

Taking the derivative of (9), we get

$$
\frac{\partial L}{\partial T_{ij}} = 2\varpi_1^2 T_{ij} A_{ij}\left\|x_i - c_j\right\|_2^2 + 2\alpha U_{ii}\left(T_{ij} - Y_{ij}\right) - \lambda_i, \tag{10}
$$

$$
\frac{\partial L}{\partial I_i} = 2\varpi_2^2 I_i \left\|x_i - c_{i\max}\right\|_2^2 - \lambda_i, \tag{11}
$$

$$
\frac{\partial L}{\partial F_i} = 2\delta^2 \varpi_3^2 F_i - \lambda_i, \tag{12}
$$

$$
\frac{\partial L}{\partial c_j} = -2\sum_{i=1}^{N}\left(\varpi_1 T_{ij}\right)^2 A_{ij}\left(x_i - c_j\right). \tag{13}
$$

By considering $\frac{\partial L}{\partial T_{ij}} = 0$, $\frac{\partial L}{\partial I_i} = 0$, $\frac{\partial L}{\partial F_i} = 0$ and $\frac{\partial L}{\partial c_j} = 0$, we get:

$$
T_{ij} = \frac{2\alpha U_{ii} Y_{ij} + \lambda_i}{2\varpi_1^2 A_{ij}\left\|x_i - c_j\right\|_2^2 + 2\alpha U_{ii}} \tag{14}
$$

$$
I_i = \frac{\lambda_i}{2\varpi_2^2 \left\|x_i - c_{i\max}\right\|_2^2} \tag{15}
$$

$$
F_i = \frac{\lambda_i}{2\delta^2 \varpi_3^2} \tag{16}
$$

$$
c_j = \frac{\sum_{i=1}^{N}\left(\varpi_1 T_{ij}\right)^2 A_{ij} x_i}{\sum_{i=1}^{N}\left(\varpi_1 T_{ij}\right)^2 A_{ij}} \tag{17}
$$

λ can be computed by replacing T_{ij}, I_i, F_i into the constraints in (8),

$$\sum_{j=1}^{C} \frac{2\alpha U_{ii} Y_{ij} + \lambda_i}{2\varpi_1^2 A_{ij} \left\| x_i - c_j \right\|_2^2 + 2\alpha U_{ii}} + \frac{\lambda_i}{2\varpi_2^2 \left\| x_i - c_{i\max} \right\|_2^2} + \frac{\lambda_i}{2\delta^2 \varpi_3^2} = 1 \tag{18}$$

$$\lambda_i = \frac{\left(1 - \sum\limits_{j=1}^{c} \frac{2\alpha U_{ii} Y_{ij}}{2\varpi_1^2 A_{ij} \left\| x_i - c_j \right\|_2^2 + 2\alpha U_{ii}}\right)}{\left(\sum\limits_{j=1}^{c} \frac{1}{2\varpi_1^2 A_{ij} \left\| x_i - c_j \right\|_2^2 + 2\alpha U_{ii}} + \frac{1}{2\varpi_2^2 \left\| x_i - c_{i\max} \right\|_2^2} + \frac{1}{2\delta^2 \varpi_3^2}\right)} \tag{19}$$

LG-WSSNCM Algorithm:

```
inputs: dataset $X$, label matrix $Y$, $U$, parameter $\varpi_1, \varpi_2, \varpi_3, \alpha$
initialization: $T, I, F, \delta^2$
while not converge do
    1. Calculate $T$ by Equation (14)
    2. Calculate $A$ by Equation (20)
    3. Calculate $I$ by Equation (15)
    4. Calculate $F$ by Equation (16)
    5. Calculate $c_j$ by Equation (17)
    6. Calculate $\lambda_i$ by Equation (19)
    7. Check convergence conditions $\left\| T^{k+1} - T^k \right\|_F^2 \le \varepsilon$
end while
outputs: Optimal solution $T, I, F, c_j$
```

The time complexity of LG-WSSNCM is mainly composed of two parts: (1) To construct the supervision matrix and the weight matrix, the coefficients need to be determined according to the information of each sample point, and the time complexity is O(n^2); (2) LG -WSSNCM flow is mainly the alternating iteration of T, I, F, the time complexity of iterative update for T is O($n^2 cd$), the time complexity of iterative update for I is O($n^2 d$), and the time complexity of iterative update for F is O($n^2 d$), where n is the number of data X, c is the number of clusters, and d is the dimension of X. If the number of iterations of the algorithm is t, the time complexity is O(tn^2cd+tn^2d). Therefore, the overall time complexity of LG-WSSNCM is O(tn^2cd+tn^2d).

3.3. The update strategy of the weight matrix

In the neutrosophic clustering method, the clustering results are obtained by dividing the membership matrix, that is, the most valuable (maximum) T_{ij} is corresponding to the j-th cluster, indicating that x_i belongs to the j-th cluster. Assuming that the number of clusters is c, it is obvious that when $T_{ij} = 1/C$ is the membership matrix is the most difficult to divide. In order to avoid the above situation and make full use of the paired constraint information, in the iterative optimization process, the obtained membership matrix is compared with $1/C$, and the weight coefficients are dynamically adjusted. The definition formula is given as follows:

$$A_{ij} = \begin{cases} a_1, & Pcl_{ij} = 1 \cap T_{ij} > \frac{1}{C}; \\ a_2, & Pcl_{ij} = 1 \cap T_{ij} \le \frac{1}{C}; \\ 1, & others; \\ \frac{1}{a_1}, & Pcl_{ij} = -1 \cap T_{ij} > \frac{1}{C}; \\ \frac{1}{a_2}, & Pcl_{ij} = -1 \cap T_{ij} \le \frac{1}{C}. \end{cases} \tag{20}$$

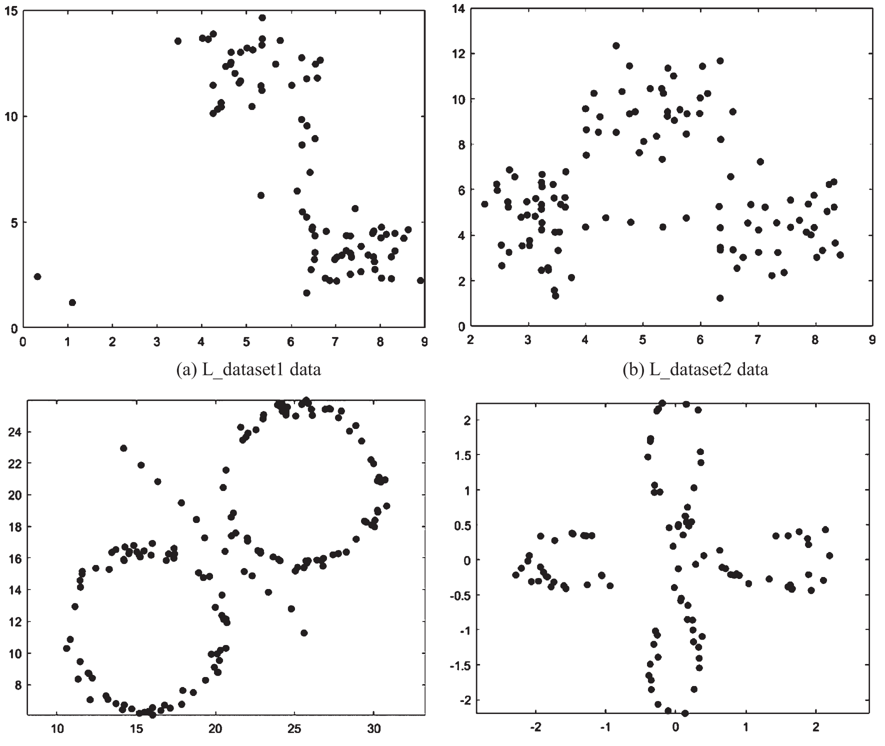
Where $0 < a_2 < a_1 < 1$, the degree of adjustment of weight coefficient is also different for different membership values. Specifically, if $\left(x_i, c_j\right) \in must - link$ is known, the corresponding membership T_{ij} is amplified by giving a smaller weighting coefficient A_{ij}. In order to carry out supervised learning with the help of paired constraint information more reasonably, considering the comparative relationship between membership degree and $1/C$, assuming $T_{ij} > 1/C$, it means that T_{ij} itself has certain distinguishing ability, and the weight coefficient a_1 does not need to be too small. If $T_{ij} \le 1/C$, it indicates that the value of T_{ij} has a large deviation,

Table 1
Parameter selection of the method in this paper

	w_1	w_2	w_3	a_1	a_2	α
Iris	0.1	0.8	0.9	0.6	0.003	5
Zoo	0.3	0.4	0.55	0.6	0.003	5
USPSdata_20	0.3	0.4	0.45	0.6	0.001	25
MnistData_05	0.1	0.8	0.9	0.6	0.001	25
Yale32x32	0.3	0.4	0.45	0.6	0.001	25
ORL_32x32	0.1	0.8	0.9	0.6	0.001	25
COIL20data_25	0.1	0.8	0.9	0.6	0.001	25
caltech101_silhouettes_28	0.2	0.4	0.45	0.6	0.001	25

Table 2
Four artificial data sets

Dataset	No. of instances	No. of feature	No. of classes
L_dataset1	85	2	2
L_dataset2	115	2	3
L_dataset3	152	2	3
Cdata04	100	2	4



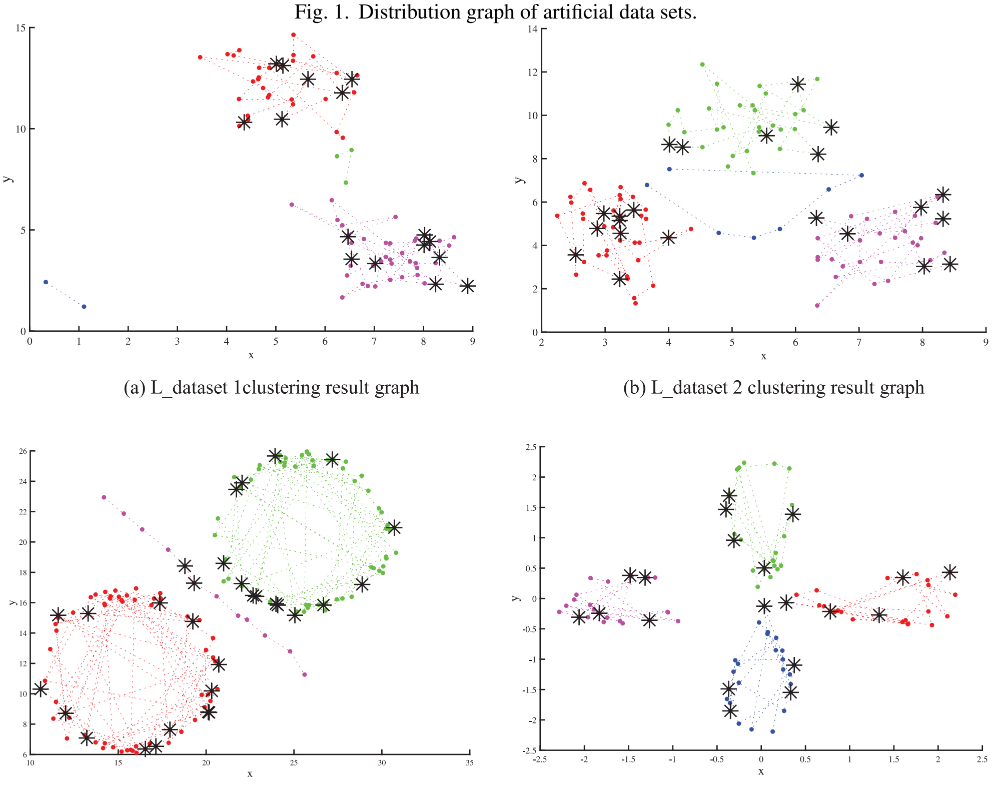
(a) L_dataset1 data

(b) L_dataset2 data

(c) L_dataset3 data

(d) Cdata04 data

Fig. 1. Distribution graph of artificial data sets.



(a) L_dataset 1clustering result graph

(b) L_dataset 2 clustering result graph

(c) L_dataset 3 clustering result graph

(d) Cdata04 clustering result graph

Fig. 2. Experimental results on artificial data sets.

and a small weight coefficient a_2 is needed to minimize the deviation as much as possible.

4. Experiments and results

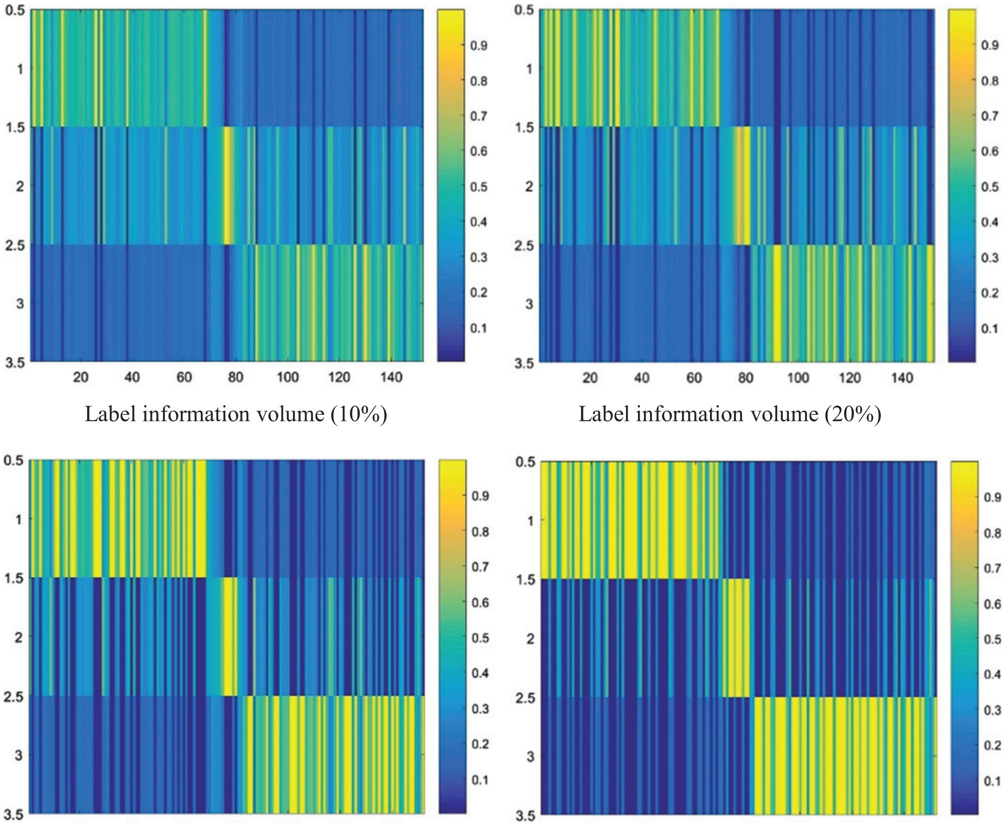
4.1. Parameters tuning

In the following experiments, the LG-WSSNCM algorithm will be compared with the SSLRR [29], SEE [30] and CSFS [38] respectively. Accuracy(ACC) index and Rand Index(RI) are used to evaluate the clustering performance. Experimental data sets contain artificial data sets and real data sets. The artificial data sets select Cdata04 and other artificial data sets to illustrate the effectiveness of LG-WSSNCM algorithm. Iris, ZOO, USPS, Mnist, Yale, Coil20, ORL and caltech101_silhouettes_28 are used for comparison experiments.

In the parameter setting, SSLRR: $\sigma = 0.1$, $\gamma = 1$; SEE: $\lambda = 0.1$; CSFS: $k = 10$, $\alpha = 0.1$. The parameters of the LG-WSSNCM algorithm are shown in Table 1.

4.2. Experiment and analysis of artificial data sets

The detailed information of the datasets is shown in Table 2. Distribution graph of the four data sets



Label information volume (10%)

Label information volume (20%)

Label information volume (40%)

Label information volume (60%)

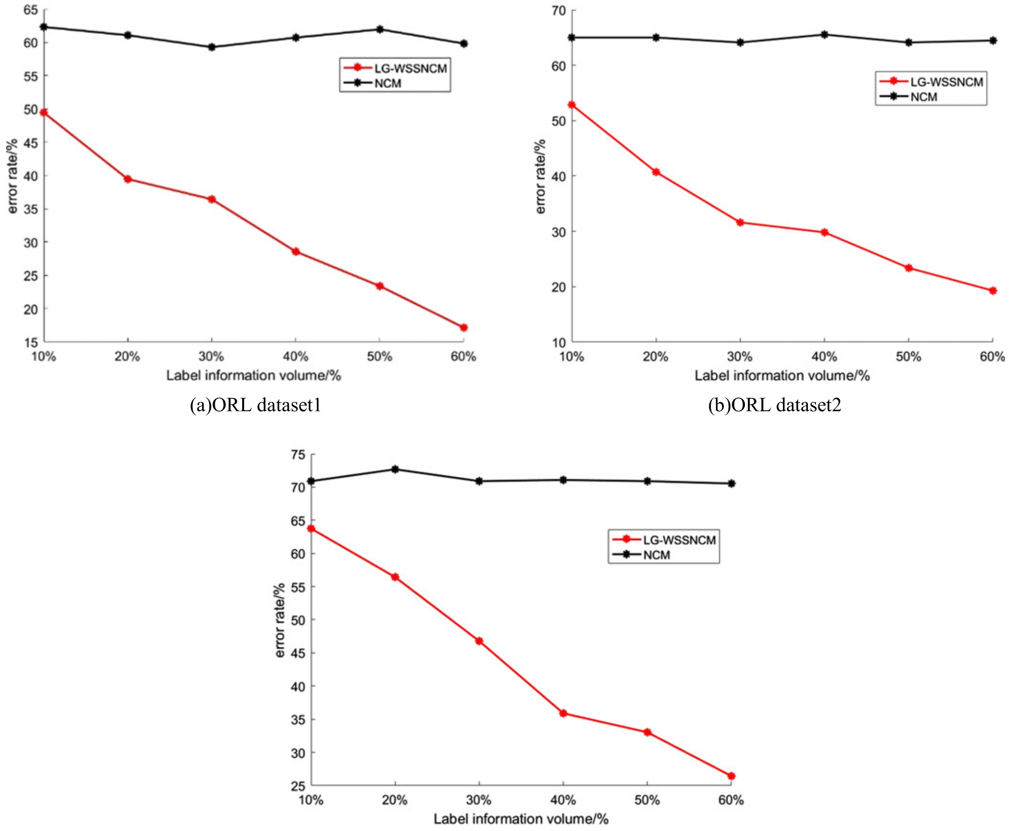
Fig. 3. Influence of the volume of labeled information on the membership matrix based on L_dataset3.

ORL dataset1

ORL dataset2

ORL dataset3

Fig. 4. The ORL face data sets.



(a)ORL dataset1

(b)ORL dataset2

(3)ORL dataset3

Fig. 5. The ORL face data of 1, 2, 3 recognition error rate of LG-WSSNCM and NCM algorithms.

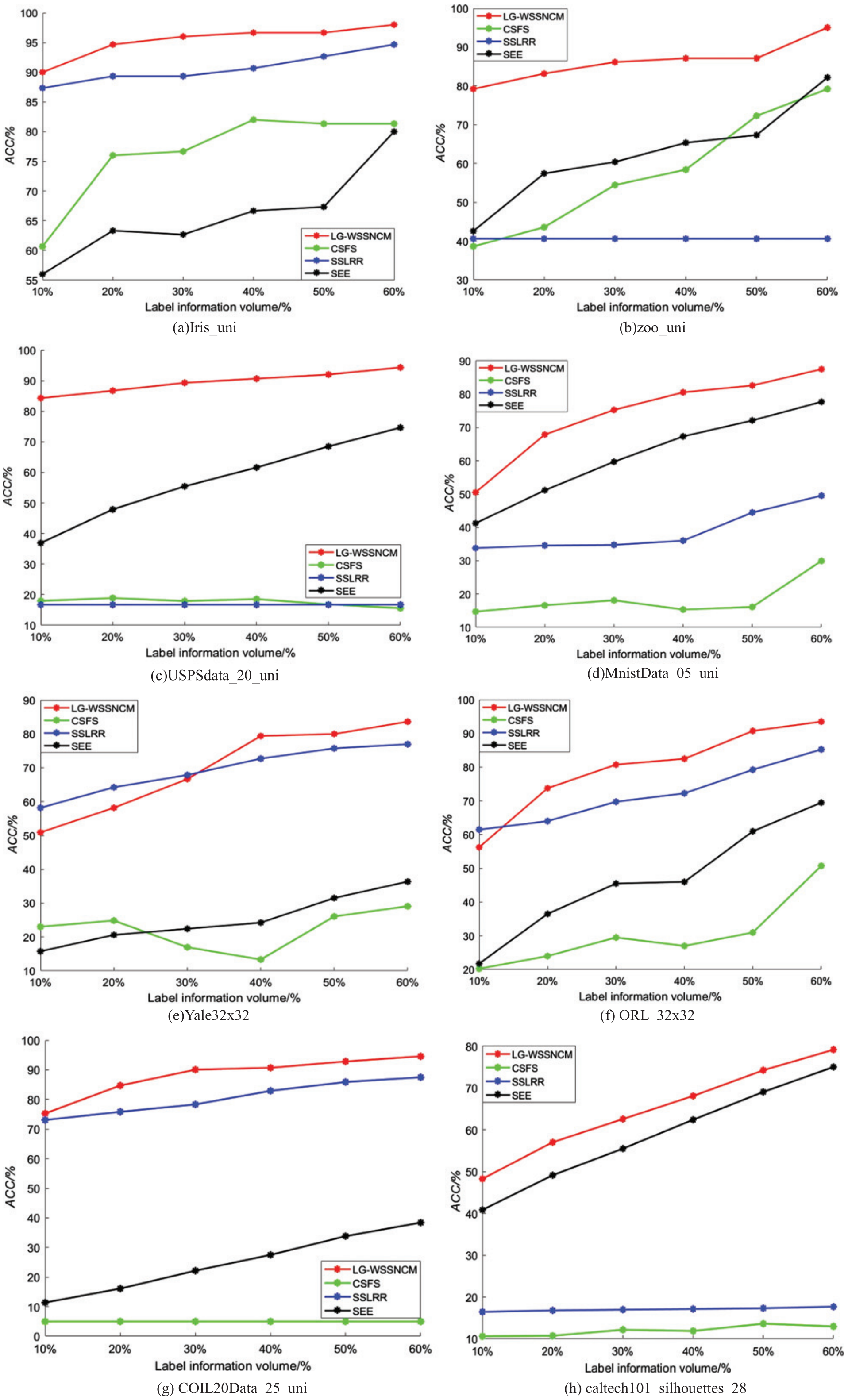
are shown in Fig. 1. In the experiment, the value of label information is 20%. Random extraction of label information will result in some data sets not being extracted. Therefore, for fairness, we extract the same proportion of label information from each class of each data set to ensure that the label information contains the information of each class. Label matrix Y is constructed according to the data point category information of each selected data set. Finally, the paired constraint information of each data set can be obtained through Equation (6).

Experimental results are shown in Fig. 2 on four datasets, where Fig. 2(a), (b), (c) and (d) represent the original data graph, and (e), (f), (g) and (h) are clustering result graph, the black "*" in the figure indicates the known label information of the points. It can be seen clearly that the clustering results satisfy the $tr(T\text{-}Y)$ in the model. At the same time, it can be seen that LG-WSSNCM has obtained accurate clustering results, and found boundary points and noise points. On the one hand, LG-WSSNCM uses paired constraint information to construct supervision weight coefficients to adjust the membership matrix. On the other hand, LG-WSSNCM uses label information to correctly guide the clustering process to obtain correct clustering results.

From the experiments, it can be seen that LG-WSSNCM can correctly guide the clustering process, to improve the clustering effect.

Table 3
UCI real data set

Datasets	No. of instances	No. of feature	No. of class
Iris	150	4	3
zoo	101	16	7
USPSdata_20	1854	265	10
MnistData_05	3495	784	10
Yale32x32	165	1024	15
ORL_32x32	400	1024	40
COIL20data_25	1440	1024	20
caltech101_silhouettes_28	8671	784	101



(a)Iris_uni

(b)zoo_uni

(c)USPSdata_20_uni

(d)MnistData_05_uni

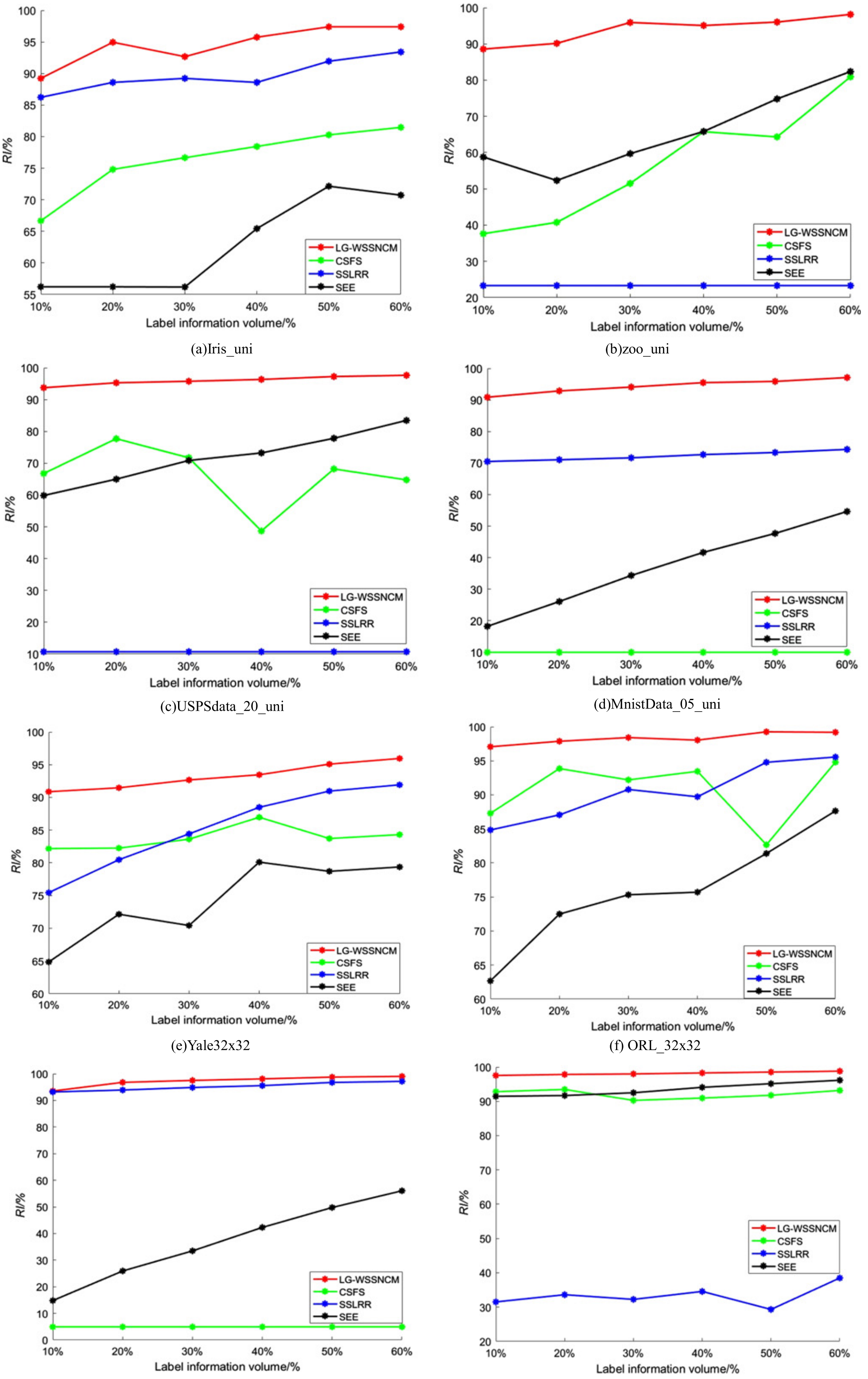
(e)Yale32x32

(f) ORL_32x32

(g) COIL20Data_25_uni

(h) caltech101_silhouettes_28

Fig. 6. Evaluation results of UCI real data set on clustering algorithm performance based on ACC.



(a)Iris_uni

(b)zoo_uni

(c)USPSdata_20_uni

(d)MnistData_05_uni

(e)Yale32x32

(f) ORL_32x32

(g) COIL20Data_25_

(h) caltech101_silhouettes_28

Fig. 7. Evaluation results of UCI real data set on clustering algorithm performance based on RI.

4.3. Exploring the influence of the volume of labeled information on the distinguishability of the membership matrix T

In this part, we use a visualized artificial dataset to explore the influence of label information on the clustering performance of LG-WSSNCM. The distinguishability of T under different volumes of label information is showed in Fig. 3. It can be found that as the volume of label information increases, the degree of membership of T becomes more distinguishable, that is, the value of the corresponding cluster of T increases.

4.4. Face data set

The ORL face database contains 400 face pictures of 40 people. The resolution size of each face image is 112×92. In order to reduce the storage space of the algorithm, the pixel value is standardized to [0, 1]. In this experiment, five images are randomly selected from the ORL face database. LG-WSSNCM and NCM on the ORL face were compared, as shown in Fig. 4.

In Fig. 5, the error rate is shown on ORL face. It can be found that the results of the NCM algorithm fluctuate slightly, however, it is limited by the inability to use the known label information to improve the clustering effect, so the error rate is high. While the LG-WSSNCM can improve the clustering accuracy with the help of label information, and its error rate continues to decrease as the volume of known information increases.

4.5. Experiment and analysis on UCI data set

In order to verify the effect of the algorithm on real data sets and determine whether the algorithm has practical significance, the eight data sets in the UCI database are used for comparison experiments, and their detailed information are shown in Table 3. The label information volume of data used in the experiment ranges from [10%, 60%].

According to the semi-supervised learning theory, as the volume of label information increases, the clustering accuracy should also improve. In Fig. 6, when the volume of label information gradually increases, the ACC of LG-WSSNCM gradually increases. However, the CSFS algorithm is descending, and the ACC of SSLRR is not changed. The NMI also has the same curve trend in Fig. 7.

In conclusion, the LG-WSSNCM can reasonably use label information to improve the clustering effect, and the clustering performance will improve with the increase of the volume of label information of the data, and finally get good clustering results. The experimental results show that the semi-supervised neutrosophic clustering algorithm proposed in this paper is reasonable and effective.

5. Conclusion

The LG-WSSNCM proposed in this paper can make full use of supervised information. On the one hand, supervised weight coefficient is constructed to adjust the membership degree of data points and clustering centers through paired constraint information, and on the other hand, label information is used to guide clustering results. Experimental results on a variety of data sets verify that the LG-WSSNCM is reasonable and effective. In the future, we will consider applying the semi-supervised neutrosophic clustering algorithm to image processing, and reduce the influence of parameters through reasonable methods.

Acknowledgment

This work is supported by the Natural Science Foundation of China (61976130), Natural Science Foundation of Shaanxi Province (2020JQ-923), key research and development projects of Shaanxi Province (2018KW-021).

References

[1] L.H. Son, DPFCM, a novel distributed picture fuzzy clustering method on picture fuzzy sets, *Expert Systems with Applications* **42**(1) (2015), 51–66.

[2] J. Shan, H.D. Cheng and Y. Wang, A novel segmentation method for breast ultrasound images based on neutrosophic l-means clustering, *Medical Physics* **39**(9) (2012), 5669.

[3] Y. Guo, R. Xia, A. Şengür and K. Polat, A novel image segmentation approach based on neutrosophic c-means clustering and indeterminacy filtering, *Neural Comput & Applic* **28**(10) (1997), 3009–3019.

[4] M.C. Massi, F. Ieva and E. Lettieri, Data mining application to healthcare fraud detection: a two-step unsupervised clustering method for outlier detection with administrative databases, *BMC Medical Informatics and Decision Making* **20**(1) (2020), 160.

[5] K. Jain, M.N. Mutry and P.J. Flynn, Data clustering in a review, *Acm Computing Surveys* **31**(3) (1999), 264–323.

[6] S.J. Nanda, I. Gulati and R. Chauhan, A k-means galactic swarm optimization based clustering algorithm with otsu's entropy for brain tumor detection, *Applied Artificial Intelligence* **33**(2) (2019), 152–170.

[7] J.C. Bezdek, R. Ehrlich and W. Full, FCM.: the fuzzy c-means clustering algorithm, *Computers and Geoences* **10**(2–3) (1984), 191–203.

[8] J.L. Xu, J.W. Han and K. Xiong, Robust and sparse fuzzy k-means clustering, *Proceedings of the 25th International Joint Conference on Artificial Intelligence* (2016), 2224–2230.

[9] S. Wang, Q. Li, C. Zhao, et al., Extreme clustering – a clustering method via density extreme points, *Information Sciences* **542** (2021), 24–39.

[10] W. Tong, S. Liu, X.Z. Gao. A density-peak-based clustering algorithm of automatically determining the number of clusters, *Neurocomputing* **458**(8) (2020), 655–666.

[11] X. Chang, F. Nie, S. Yang, et al., Compound rank-k projections for bilinear analysis, *IEEE Transactions on Neural Networks & Learning Systems* **27**(7) (2016), 1502–1513.

[12] Q.T. Bui, B. Vo, H.A. Do, et al., F-mapper: a fuzzy mapper clustering algorithm, *Knowledge-Based Systems* **189** (2019), 105107.

[13] Q.T. Bui, B. Vo, V. Snasel, et al., SFCM: a fuzzy clustering algorithm of extracting the shape information of data, *IEEE Transactions on Fuzzy Systems* **29**(1) (2020),75–89.

[14] F. Smarandache, Neutrosophy, a new branch of pilosophy, *Multiple Valued Logic* **8**(3) (2002), 297.

[15] J. Ye, Single-valued neutrosophic clustering algorithms based on similarity measures, *Journal of Classification* **34**(1) (2017), 148–162.

[16] Y. Guo, A. Şengür, NECM: neutrosophic evidential c-means clustering algorithm, *Neural Computing and Applications* **26**(3) (2014), 561–571.

[17] H.V. Long, M. Ali, L.H. Son, et al., A novel approach for fuzzy clustering based on neutrosophic association matrix, *Computers & Industrial Engineering* **127** (2019), 687–697.

[18] Y. Guo and A. Sengur, NCM: neutrosophic c-means clustering algorithm, *Pattern Recognition* **48**(8) (2015), 2710–2724.

[19] J. Ye, Single-valued neutrosophic minimum spanning tree and its clustering method, *Journal of Intelligent Systems* **23**(3) (2014), 311–324.

[20] I. Kandasamy, Double-valued neutrosophic sets, their minimum spanning trees, and clustering algorithm, *Journal of Intelligent Systems* **27**(2) (2018), 163–182.

[21] P.H. Thong and L.H. Son, Picture fuzzy clustering: a new computational intelligence method, *Soft Computing* **20**(9) (2016), 3549–3562.

[22] Q. Li, Y. Ma and F. Smarandache, et al., Single-valued neutrosophic clustering algorithm based on tsallis entropy maximization, *Axioms* **7**(3) (2018), 57.

[23] E. Rashno, B. Minaei-Bidgoli and Y Guo, An effective clustering method based on data indeterminacy in neutrosophic set domain, *Engineering Applications of Artificial Intelligence* **89** (2020), 103411.

[24] Y. Akbulut, A. Şengür, Y.Guo, et al., KNCM: Kernel neutrosophic c-means clustering, *Applied Soft Computing* **52** (2017), 714–724.

[25] C.J. Merz, D.C.S. Clair and W.E. Bond, Semi-supervised adaptive resonance theory, *International Joint Conference on Neural Networks* (1992), 851–856.

[26] Y. Liu and X. Zhang, Semi-supervised spectral clustering based on density adaptive neighborhood similarity graphs, *Computer Application Research* **37**(9) (2020), 2604–2609.

[27] S. Qiu, F. Nie, X. Xu and C. Qing, Accelerating flexible mani-fold embedding foer scalable semi-supervised learning, *IEEE Transactions on Circuits & Systems for Video Technology* **29**(9) (2018), 2786–2795.

[28] S. Fariba, R.K. Mohammad and S. Arash, SMKFC-ER: Semi-supervised multiple kernel fuzzy clustering based on entropy and relative entropy, *Information Sciences* **547** (2021), 667–688.

[29] L. Zhuang, Z. Zhou, S. Gao, J. Yin, Z. Lin and Y. Ma, Label information guided graph construction for semi-supervised learning, *IEEE Transactions on Image Processing* **26**(9) (2017), 4182–4192.

[30] F. Nie, H. Wang, H. Huang and C. Ding, Adaptive loss minimization for semi-supervised elastic embedding, *Proceedings of the 23th International Joint Conference on Artificial Intelligence* (2013), 1565–1571.

[31] H. Zhu, Y. Ma and X. Dai, Adaptive semi-supervised neighborhood clustering algorithm, *Journal of Shandong University(Engineering Science)* **51**(4) (2021), 24–34.

[32] K. Chen, L. Yao, D. Zhang, et al., A semisupervised recurrent convolutional attention model for human activity recognition, *IEEE Transactions on Neural Networks and Learning Systems* **31**(5) (2020), 1747–1756.

[33] B. Prasanth and N. Nagamalleswararao, Optimal kernel based neutrosophic soft sets clustering for image segmentation based on pareto optimal algorithm, *International Journal of Recent Technology and Engineering* **8**(4) (2019), 582–591.

[34] H.V. Long, M. Ali, L.H. Son, M. Khan and D.N. Tu, A novel approach for fuzzy clustering based on neutrosophic association matrix, *Computers & Industrial Engineering* **127**(2018), 687–697.

[35] L. Madan, K. Lakhwinder and G. Savita, Automatic segmentation of tumors in B-Mode breast ultrasound images using information gain based neutrosophic clustering, *Journal of X-ray Science and Technology* **26**(2) (2018), 209–225.

[36] K. Wagstaff and C. Cardie, Clustering with instance level constraints, *Proceedings of 17th International Conference on Machine Learning* (2000), 1097–1103.

[37] K. Wagstaff, C. Cardie, S. Roger and S. Schroedl, Constrained k-means clustering with background knowledge, *Proceedings of 18th International Conference on Machine Learning* (2001), 577–584.

[38] X. Chang, F. Nie, Y. Yang and H. Huang, A convex formulation for semi-supervised multi-label feature selection, *AAAI'14 Proceedings of the 28th AAAI Conference on Artificial Intelligence* (2014), 1171–1177.