Neutrosophic Sets and Systems

Volume 51 Article 36

10-5-2022

Clustering Algorithm Based on Data indeterminacy in Neutrosophic Set

Dan Zhang

Yingcang Ma

Florentin Smarandache

Xuezhen Dai

Yaqin Qiao

Follow this and additional works at: https://digitalrepository.unm.edu/nss_journal

Recommended Citation
Zhang, Dan; Yingcang Ma; Florentin Smarandache; Xuezhen Dai; and Yaqin Qiao. "Clustering Algorithm Based on Data indeterminacy in Neutrosophic Set." *Neutrosophic Sets and Systems* 51, 1 (2022).
https://digitalrepository.unm.edu/nss_journal/vol51/iss1/36

This Article is brought to you for free and open access by UNM Digital Repository. It has been accepted for inclusion in Neutrosophic Sets and Systems by an authorized editor of UNM Digital Repository. For more information, please contact disc@unm.edu.

Neutrosophic Sets and Systems, Vol. 51, 2022
University of New Mexico

Clustering Algorithm Based on Data indeterminacy in Neutrosophic Set

Dan Zhang [1], Yingcang Ma [1,*], Florentin Smarandache [2], Xuezhen Dai [3] and Yaqin Qiao[3]
[1] School of Science, Xi'an Polytechnic University, Xi'an, China;
[2] Mathematics and Science Division, Gallup Campus, University of New Mexico, Gallup, NM, USA;
[3] The Public Sector, Xi'an Traffic Engineering Institute, Xi'an, China;

* Correspondence: mayingcang@xpu.edu.cn;

Abstract: Clustering research is an important field in machine learning, pattern recognition and other fields. The neutrosophic set characterizes the data through true membership functions, indeterminate membership functions and false membership functions. Data clustering using neutrosophic set has become one of the current research hotspots. In this paper, first, a new definition of data uncertainty in a neutrosophic set is proposed in this paper based on the density of data. Next, a clustering model based on the uncertainty value of neutrosophic set data is proposed by considering the main cluster (true membership) and the noise cluster (false membership) in the data set. The model takes into account the distance of the data points to the cluster centers and the indeterminacy value of each data point, and then minimizes the proposed cost function by the method of Lagrangian multipliers. The true membership value and false membership value of each data point can be obtained. Finally, the effectiveness of the method is demonstrated by experiments on the various datasets. Experimental results show that the cost function has more accurate membership degree when dealing with boundary points and outliers, and outperforms existing clustering methods on datasets.

Keywords: neutrosophic set; data indeterminacy; clustering algorithm

1. Introduction

Clustering is to divide data into disjoint groups, each of which satisfies two rules: Objects are similar (or related) to each other within the same group (minimizing intra-cluster distance), and at the same time different (or unrelated) to other groups (maximizing inter-cluster distance). Data clustering is an important field in machine learning and has a wide range of applications in computer vision, image processing, medicine, geology, and pattern recognition [1-6].

In k-type clustering, the clustering method represented by k-means [7] is hard clustering, and *k*-means makes each data point belong to exactly one cluster. It divides the data into *k* clusters by minimizing the intra-cluster squared distance and the main disadvantage is that it cannot ensure a global minimum variance. *K*-medoid is a variant of *k*-means that computes the median of each cluster for its cluster center. One of the strongest assumptions in median-based clustering models is that objects must belong to one (and only one) cluster. However, Krishnapuram proposed the fuzzy *k*-center clustering algorithm (FKM) [8]. The essential difference between FKM and *k*-means is that FKM allows each data point to have membership in all clusters, rather than a single cluster with different memberships. Kannan [9] proposed a robust kernel-based FKM by combining normed

kernel function and center initialization algorithm. Reference[10] introduced adaptive spatial information theory fuzzy clustering into traditional FKM to improve robustness.

Different from the hard clustering, the fuzzy clustering allows each object to be assigned to all clusters with different degrees of membership. FCM [11] is the most typical fuzzy clustering algorithm. But FCM has four major problems: 1) It just minimizes the variance within the class and does not consider the variance between clusters like the *k*-means algorithm does. 2) The result of clustering depends largely on the initialization. 3) It is sensitive to noise, and the membership degree of noise points may be high. 4) It is also sensitive to the type of distance metric and cannot distinguish between equally likely and equally less likely data points. Krishnapuram and Keller proposed a new possibility *c*-means (PCM) [12]. However, it is sensitive to cluster center initialization, requires additional parameters to be tuned, and may generate overlapping clusters. Reference [13] proposed a robust sparse fuzzy *k*-means algorithm (RSFKM), which introduced a robust function to deal with outliers and noise points to enhance the robustness and sparsity of the FCM algorithm. Reference [14] proposed a variant of fuzzy clustering and hard clustering called relational fuzzy c-means. In recent years, many clustering methods have been developed based on different theories [15-17].

The neutrosophic theory [18] was first proposed by Smarandache in 1995. Picture fuzzy set is a standardized form of neutrosophic set. Thong [19] proposed an picture fuzzy clustering algorithm(FC-PFS). This algorithm needs to calculate three matrices of the same scale, and the clustering effect is not good for high-dimensional data. Li [20] proposed a single-valued neutrosophic clustering algorithm based on Tsallis entropy maximization in the framework of picture fuzzy set clustering and single-valued neutropenic set. The algorithm showed good results in image segmentation. Another the algorithms are based on the original neutrosophic set framework. For example, Guo [21] proposed the neutrosophic c-means clustering algorithm (NCM) based on the neutrosophic set and FCM, which can effectively distinguish the sample points, boundary points and outliers in the cluster. The true membership is not affected by noise, which effectively solves the problem that the FCM algorithm cannot detect abnormal data points. Rashno [22] proposed a neutrosophic clustering algorithm based on data indeterminacy, which can effectively separate boundary points and noise points. Ye [23] proposed a single-valued neutrosophic minimum spanning tree clustering algorithm (SVNMST) by defining a generalized single-valued neutrosophic set distance measure, which showed great superiority in the clustering of single-valued neutrosophic observation data. Kandasamy [24] proposed a dual-valued neutrosophic minimum spanning tree clustering algorithm (DVNMST) to cluster data represented by dual-valued neutrosophic information. All previous methods deal with boundaries and outliers directly in the cost function. This paper mainly deals with boundary points and outliers by proposing an indeterminate set (I) in the NS set, and expressing this set as a new clustering cost function. The rest of the paper is organized as follows. Section II reviews the FKM algorithm and the NS set. Section III presents the proposed method(INCA). Section IV presents the experimental results of the method on scatter and real datasets. Finally, Section V concludes the paper.

2. Related Algorithms

2.1 Definition of NS

X is a set of objects, *x* is an element in *X*, and the neutrosophic set *A* on *X* can be expressed as

$$A = \{[x, (T_A(x), I_A(x), F_A(x))] \mid x \in X\}, \tag{1}$$

where $T_A(x)$ is the true value of the object, $I_A(x)$ is the indeterminate value, $F_A(x)$ is the false value. They belongs to the standard and non-standard subsets in $]0^-, 1^+[$, namely

$T_A(x), I_A(x), F_A(x) : X \to]0^-, 1^+[$. The sum of $T_A(x), I_A(x), F_A(x)$ has no limit, so there is $0^- \le \sup T_A(x) + \sup I_A(x) + \sup F_A(x) \le 3^+$.

2.2 FKM

The FKM algorithm is a clustering algorithm based on the median of objects, and its objective function is as follows

$$\min Z_{FKM} = \sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}\left(e_{ij}\right)^h, \tag{2}$$

$$\begin{aligned} s.t.\ & \sum_{j=1}^{n} e_{ij} = 1, \forall i \in \{1,\cdots,n\}; \\ & e_{ij} \le e_{jj}; \quad \sum_{j=1}^{n} e_{jj} = k; \\ & e_{ij} \in [0,1], \forall i, j \in \{1,\cdots,n\}, i \ne j; \\ & e_{jj} \in \{0,1\}, \forall j \in \{1,\cdots,n\}. \end{aligned} \tag{3}$$

where e_{ij} is the representation of the data object o_i to the cluster center o_j (if o_j is not the cluster center, then e_{ij}), and *h* is the fuzzy factor. The fuzzy factor is a hyperparameter that represents the expected degree of overlap between the clusters to be found. When *h* → 1+, data objects are often assigned to a cluster, the clustering is very clear. When *h* → ∞, objects tend to be evenly distributed in each cluster. The final membership value for each non-cluster center and each cluster center is 1/*k*.

Given a set of known cluster centers (selected from sample points), the membership of each object to the selected cluster center can be found by computing the following expression:

$$e_{ij} = \Big/ \sum_{t|e_{it}=1} \left(\frac{d_{ij}}{d_{it}}\right)^{1/(h-1)}, \tag{4}$$

2.3 NCM

The neutrosophic *c*-means (NCM) [10] defines the true membership, false membership and indeterminate membership of the data. NCM can handle boundary points and outliers contained in the dataset itself . Solve the following convex optimization problem:

$$J(T,I,F) = \sum_{i=1}^{N}\sum_{j=1}^{C}\left(\varpi_1 T_{ij}\right)^m \left\|x_i - c_j\right\|^2 + \sum_{i=1}^{N}\left(\varpi_2 I_i\right)^m \left\|x_i - \overline{c}_{i\max}\right\|^2 + \sum_{i=1}^{N}\delta^2\left(\varpi_3 F_i\right)^m, \tag{5}$$

where *m* is a constant. T_{ij}, I_i, F_i are the membership value belongs to the determinate clusters, boundary regions and noise datasets. Define $0 < T_{ij}, I_i, F_i < 1$, satisfying the following constraints:

$$\sum_{j=1}^{C} T_{ij} + I_i + F_i = 1, \tag{6}$$

For each data point *i*, the cluster centeris $c_{i\max}$ calculated using T_{ij} with the largest and second largest value:

$$c_{i\max} = \frac{c_{pi} + c_{qi}}{2}, \tag{7}$$

$$\begin{cases} p_i = \arg\max\limits_{j=1,2,...C}(T_{ij}) \\ q_j = \arg\max\limits_{j\neq pi \cap j=1,2,...C}(T_{ij}) \end{cases} \tag{8}$$

3. INCA

3.1 Characterization of indeterminacy

A new clustering method is proposed in this paper, which can cluster data containing outliers and boundary points. The basic idea is to combine the FCM algorithm with the neutrosophic set. First, we define the indeterminate for each data point through Euclidean distance, and use the uncertainty in the neutrosophic set to describe it.

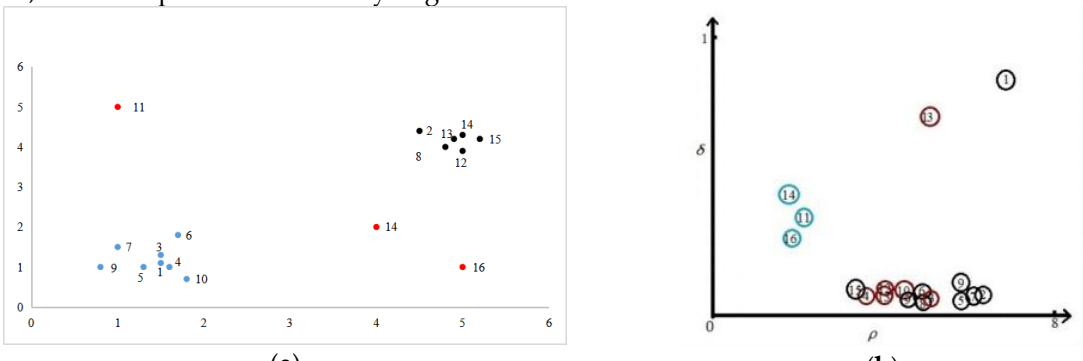
$$\rho_i = \sum_j \chi(d_{ij} - d_c), \quad \chi(x) = \begin{cases} 1, & x < 0 \\ 0, & x > 0 \end{cases}, \tag{9}$$

$$\delta_i = \begin{cases} \min_{j\in I_s^i}\{d_{ij}\}, & I_s^i \neq \varnothing \\ \max_j \{d_{ij}\}, & I_s^i = \varnothing \end{cases} \tag{10}$$

$$I_s^i = \{k : \rho_k > \rho_i\} \tag{11}$$

$$I_i = \frac{1}{\rho_i \delta_i} \tag{12}$$

where ρ_i is the local density of the i-th data sample and δ_i is the distance attribute of the *i*-th data sample. If a point is denser than its neighbors and has a relatively large distance from the more dense point, the point is considered to be within the main cluster and should have less uncertainty. Instead, the point has a larger indeterminate value. This idea makes the uncertainty close to 1 for noise points and close to 0 for the points within the main cluster. Lower uncertainty is assigned to the points in dense regions and not vice versa. As shown in Figure 1, points 1 and 18 in the left figure are the cluster centers of the two clusters. It can be seen from the right figure that the values of δ_i and ρ_i of the two points are large, so the indeterminate value is small. The indeterminate values of 11, 14 and 16 points are relatively large.



(**a**) (**b**)

Figure 1. The distribution of data points, ρ_i and δ_i (**a**) data points; (**b**) ρ_i and δ_i

3.2 Model

In INCA, the determinate and indeterminate membership of the main cluster and noise points is considered. Set A is the union of determinate clusters and indeterminate clusters, $A = C_i \cup R; i = 1,2,\cdots,k$; where C_i and R represent determinate clusters and indeterminate clusters I, and $\cup$ is the union operator. In clustering applications, C_i and R represent the membership degree of the true set and the false set. Therefore, C_i and R are the union of true and false set in the NS set. We hope that a smaller distance $\|x_i - c_j\|^2$ corresponds to a larger true membership T_{ij} and a smaller false membership F_i. It indicates that the data points x_i are easily divided into the corresponding clusters c_j. A larger distance $\|x_i - c_j\|^2$ corresponds to a smaller true membership T_{ij} and a larger false membership F_i. It indicates that the data points are not easily divided into the corresponding clusters c_j. The objective function of the proposed algorithm is:

$$L(T,F) = \sum_{i=1}^{n}\sum_{j=1}^{n}\|x_i - x_j\|_2^2 \left(\omega_1 I_i T_{ij}\right)^m + \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\omega_2 (1-I_i) F_i\right)^m e^{-\|x_i - x_j\|^2}, \tag{13}$$

where T_{ij} and F_i are the membership of the data i to the main cluster j and the membership of the noise cluster. For each data point, the following conditions are simultaneously met:

$$s.t. \ \sum_{j=1}^{n} T_{ij} + F_i = 1, \quad \forall i \in \{1,\cdots,n\}, \tag{14}$$

The decision variable $T_{ij}\left(i, j \in \{1,2,\cdots n\}\right)$ is the membership degree that assigns the data object i to the cluster center j (if the data point j is not a cluster center, $T_{ij} = 0$). To comply with the constraints of NS theory, constraints (14) are defined. As can be seen from the above model, there are two conditions for data point i to have the highest membership degree to the cluster j: a) the distance of data point i to cluster center j is less than the distance to other cluster centers. b) The data point i should have a small indeterminacy. Similarly, there are two conditions for data point i to have the highest membership to a noisy cluster: a) it has the largest sum distance from all main clusters. b) The data point i should have a large indeterminacy.

3.3 Model solution

The Lagrangian function of the model is:

$$L(T,F) = \sum_{i=1}^{n}\sum_{j=1}^{n}\|x_i - x_j\|_2^2 \left(\omega_1 I_i T_{ij}\right)^m + \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\omega_2 (1-I_i) F_i\right)^m e^{-\|x_i - x_j\|^2} - \sum_{i=1}^{n}\lambda_i\left(\sum_{j=1}^{n} T_{ij} + F_i - 1\right), \tag{15}$$

To minimize the Lagrange objective function, we use the following operations:

$$\frac{\partial L}{\partial T_{ij'}} = m\left(\omega_1 I_i\right)^m T_{ij'}^{m-1}\|x_i - c_{j'}\|_2^2 - \lambda_i, \tag{16}$$

$$\frac{\partial L}{\partial F_i} = m\left(\omega_2\left(1-I_i\right)^m\right)F_i^{m-1}\sum_{j'=1}^{c} e^{-\left\|x_i-c_{j'}\right\|_2^2} - \lambda_i, \tag{17}$$

The norm is specified as the Euclidean norm. Let $\frac{\partial L}{\partial T_{ij'}} = 0$ and $\frac{\partial L}{\partial F_i} = 0$, then:

$$T_{ij'} = \left(\frac{\lambda_i}{m}\right)^{\frac{1}{m-1}}\left(\omega_1 I_i\right)^{-\frac{m}{m-1}}\left\|x_i - c_{j'}\right\|_2^{-\frac{2}{m-1}}, \tag{18}$$

$$F_i = \left(\frac{\lambda_i}{m}\right)^{\frac{1}{m-1}}\left(\omega_2\left(1-I_i\right)\right)^{-\frac{m}{m-1}}\left(\sum_{j'=1}^{c} e^{-\left\|x_i-c_{j'}\right\|_2^2}\right)^{-\frac{1}{m-1}}, \tag{19}$$

Let $\left(\frac{\lambda_i}{m}\right)^{\frac{1}{m-1}} = Ktemp$,

$$Ktemp = \left(\left(\omega_1 I_i\right)^{-\frac{m}{m-1}}\sum_{j'=1}^{c}\left\|x_i - c_{j'}\right\|_2^{-\frac{2}{m-1}} + \left(\omega_2\left(1-I_i\right)\right)^{-\frac{m}{m-1}}\left(\sum_{j'=1}^{c} e^{-\left\|x_i-c_{j'}\right\|_2^2}\right)^{-\frac{1}{m-1}}\right)^{-1}, \tag{20}$$

Therefore:

$$T_{ij'} = Ktemp\left(\omega_1 I_i\right)^{-\frac{m}{m-1}}\left\|x_i - c_{j'}\right\|_2^{-\frac{2}{m-1}}, \tag{21}$$

$$F_i = Ktemp\left(\omega_2\left(1-I_i\right)\right)^{-\frac{m}{m-1}}\left(\sum_{j'=1}^{c} e^{-\left\|x_i-c_{j'}\right\|_2^2}\right)^{-\frac{1}{m-1}}, \tag{22}$$

The above equations allow the formulation of INCA algorithm. It can be summarized in the following steps:

INCA algorithm:

input：X、n、k、D

output：T, F;

1：randomly select k centers;

2：Calculate T using Equation (21);

3：Calculate F using Equation(22);

4：Calculate the value of the objective function Z_1;

5：Select k centers by exhaustive method;

6：Calculate T_2;

7：Calculate F_2;

8：Calculate the value of the objective function Z_2;

9：Compare the values of Z_1 and Z_2, if $Z_2 < Z_1$, go back to step 5. If $Z_2 > Z_1$, assign the center of Z_1 to Z_2, T_2 to T, F_2 to F and the end.

The time complexity of INCA is divided into two parts. The first part is the calculation of the memberships T and F. It is related to the sample dimension, the number of samples and the number

of categories, and needs to traverse all the data points in the data. If the dimension of the given dataset is m, the number of sample points is n, and the number of clusters is c, the algorithm complexity is $O(n^2mc+n^2m)$. The second part is the exhaustive optimization process, which needs to iteratively calculate the memberships T and F, so the complexity of this part of the algorithm is $O[n!(n^2mc+n^2m)]$. The overall algorithm complexity of this paper is $O[n!(n^2mc+n^2m)]$. We can see that the computational complexity is very high when m and n are large.

4. Results

4.1. Datasets

The performance of INCA is evaluated on artificial datasets and real datasets. The proposed method is compared with INCM [22], FC-PFS [19], RFKM [13], NCM [21], and FKM [8] methods. In the experiment of the exhaustive clustering center, we only extract the same proportion of sample points from each class, and appropriately reduce the running time of the algorithm.

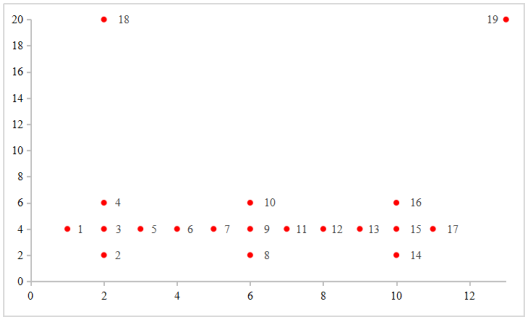
The parameter dc of the uncertainty calculation part is set by the method in the article [25]. In the cost function of INCA, the parameters are configured as $m=1.3, \omega_1=1, \omega_2=2$.

In this section, three types of datasets are used to evaluate the performance of INCA. The first is the diamond dataset, including the X19 and X24 scatter datasets proposed by Guo [25], and a scatter dataset we designed. In these datasets, border points between the main clusters and outliers far from the main clusters are considered. It is easy to see how the clustering method is affected by the main points in each dataset. The second is the UCI dataset, which includes higher-dimensional and larger-scale datasets. There are mainly dermatology, pima, TOX-171, votel, ecoli, iris, ionosphere and vote.

4.2. Results

4.2.1. Artificial datasets

The X19 dataset has three clusters in Figure 2, points 1-5, 7-11 and 13-17 are points in the main cluster, points 6 and 12 are boundary points, points 18 and 19 are noise points. Figure 3 shows the clustering results of INCA. The memberships calculated by INCA and the FKM are counted in Table 1. Although INCA and FKM assign the same cluster label to all points, INCA assigns the points(e.g. 5, 7, 11, 12) with higher indeterminate membership in their corresponding clusters. Data point 5 has the same distance between the main and border clusters, but it belongs to the main cluster. FKM cannot distinguish point 5 as a boundary or a main cluster. INCA solves this problem, and the membership of point 5 assigned to the main cluster is 0.67, while the FKM is 0.36. Figure 4 visually depicts the membership of INCA and the FKM algorithm.



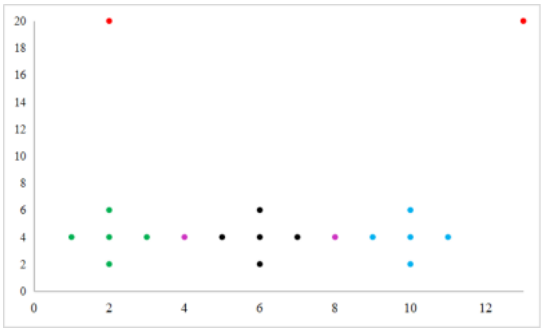
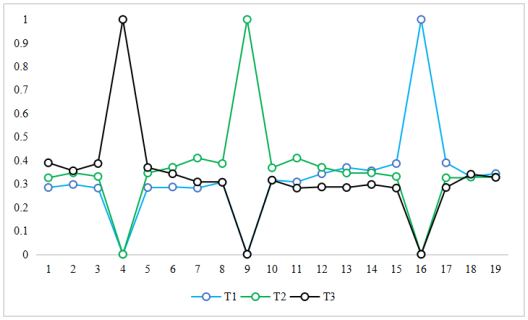
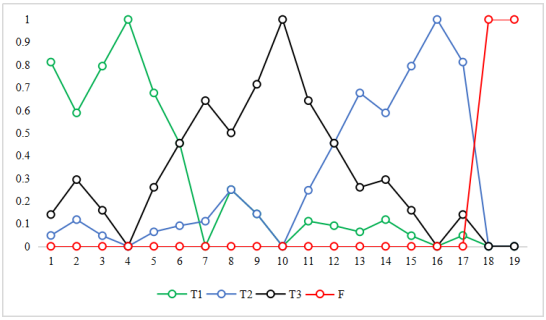


Figure 2. X19



(**a**)

Figure 3. Clustering results of INCA on X19



(**b**)

Figure 4. Membership calculated by FKM and INCA on X19

Table 1. Clustering results of X19

	FKM			INCA				
	U_1	U_2	U_3	T_1	T_2	T_3	F	
1	0.2844	0.3259	0.3897	0.8122	0.0478	0.1400	0	
2	0.2974	0.3469	0.3556	0.5882	0.1176	0.2941	0	
3	0.2821	0.3313	0.3865	0.7944	0.0467	0.1589	0	
4	0	0	1	1	0	0	0	
5	0.2843	0.3462	0.3695	0.6762	0.0638	0.2601	0	
6	0.2868	0.3704	0.3429	0.4545	0.0910	0.4545	0	boundary
7	0.2819	0.4099	0.3082	0.2470	0.1107	0.6422	0	
8	0.3068	0.3865	0.3068	0.25	0.25	0.5	0	
9	0	1	0	0.1429	0.1429	0.7143	0	
10	0.3158	0.3684	0.3158	0	0	0	1	
11	0.3082	0.4099	0.2819	0.1107	0.2470	0.6422	0	
12	0.3429	0.3704	0.2868	0.0910	0.4545	0.4545	0	boundary
13	0.3695	0.3462	0.2843	0.0638	0.6762	0.2601	0	
14	0.3556	0.3469	0.2974	0.1176	0.5882	0.2941	0	
15	0.3865	0.3313	0.2821	0.0467	0.7944	0.1589	0	
16	1	0	0	0	1	0	0	
17	0.3897	0.3259	0.2844	0.0478	0.8122	0.1400	0	
18	0.3304	0.3287	0.3409	0	0	0	1	
19	0.3437	0.3289	0.3274	0	0	0	1	

We also conduct more experiments, using the four-class X24 shown in Fig. 5 to compare INCA and FKM. Data points 6, 12 and 18 are boundaries and 24 is an outlier. Fig. 6 presents the clustering results of INCA. Table 2 lists the results of INCA and FKM. The first five data points belong to a main cluster because their T_4 values are higher for the other clusters (T_2, T_3 and T_4). It can also be inferred that similar observations data points 6, 12 and 18 are ambiguous because there are two highest T values. The last data point 24 was inferred as an outlier. Fig. 7 visually depicts the degree of membership.

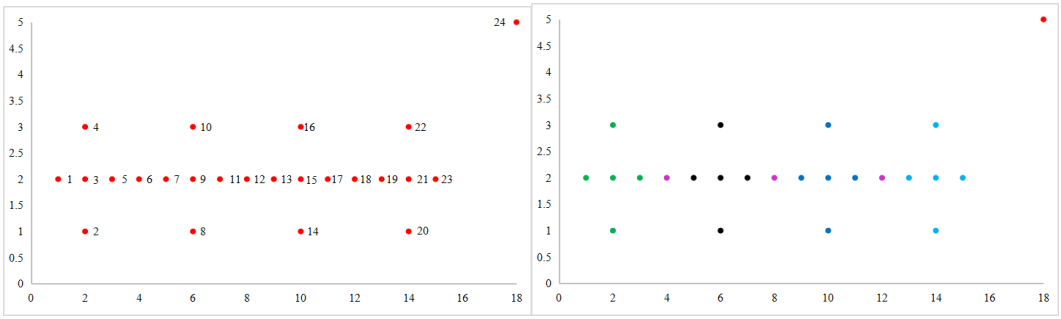
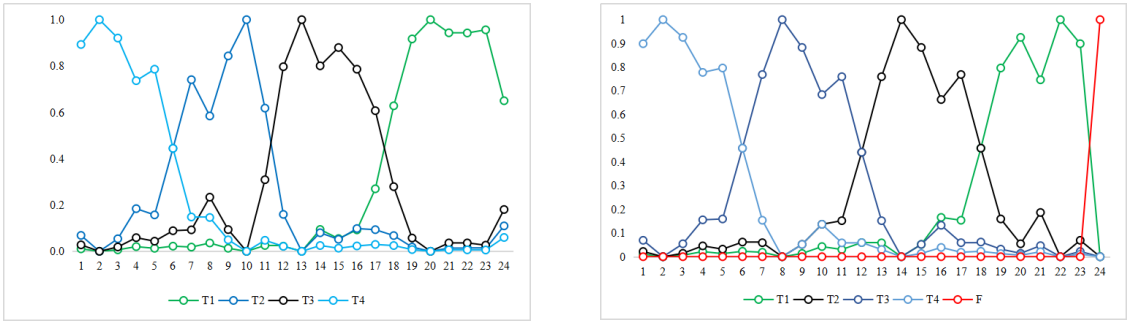


Figure 5. X24Figure 6. Clustering results of INCA on X24



(**a**) (**b**)

Figure 7. Membership calculated by FKM and INCA on X24

Table 2. Clustering results of X24

	FKM				INCA					
	U_1	U_2	U_3	U_4	T_1	T_2	T_3	T_4	F	
1	0.0106	0.0687	0.0279	0.8929	0.0106	0.0219	0.0691	0.8984	0	
2	0	0	0	1	0	0	0	1	0	
3	0.0064	0.0542	0.0188	0.9207	0.0064	0.0142	0.0544	0.9250	0	
4	0.0203	0.1842	0.0589	0.7366	0.0216	0.0457	0.1554	0.7772	0	
5	0.0130	0.1572	0.0437	0.7861	0.0130	0.0318	0.1592	0.7959	0	
6	0.0222	0.4444	0.0889	0.4444	0.0227	0.0619	0.4577	0.4577	0	boundary
7	0.0183	0.7409	0.0926	0.1482	0.0187	0.0591	0.7685	0.1537	0	
8	0.0360	0.5843	0.2337	0.1461	0	0	1	0	0	
9	0.0132	0.8435	0.0937	0.0496	0.0136	0.0519	0.8826	0.0519	0	
10	0	1	0	0	0.0427	0.1368	0.6838	0.1368	0	
11	0.0252	0.6181	0.3091	0.0476	0.0304	0.1519	0.7593	0.0584	0	
12	0.0221	0.1594	0.7969	0.0215	0.0595	0.4405	0.4405	0.0595	0	boundary
13	0	0	1	0	0.0584	0.7593	0.1519	0.0304	0	
14	0.0942	0.0801	0.8007	0.0250	0	1	0	0	0	
15	0.0550	0.0517	0.8797	0.0135	0.0519	0.8826	0.0519	0.0136	0	
16	0.0925	0.0983	0.7861	0.0231	0.1657	0.6628	0.1325	0.0340	0	
17	0.2698	0.0934	0.6071	0.0296	0.1537	0.7685	0.0591	0.0187	0	
18	0.6281	0.0679	0.2791	0.0249	0.4577	0.4577	0.0619	0.0227	0	boundary
19	0.9168	0.0183	0.0573	0.0075	0.7959	0.1592	0.0318	0.0130	0	

20	1	0	0	0	0.9250	0.0544	0.0142	0.0064	0
21	0.9433	0.0139	0.0363	0.0065	0.7461	0.1865	0.0467	0.0207	0
22	0.9426	0.0147	0.0363	0.0064	1	0	0	0	0
23	0.9562	0.0117	0.0266	0.0056	0.8984	0.0691	0.0219	0.0106	0
24	0.6499	0.1098	0.1805	0.0597	0	0	0	0	1

In this paper, a dataset is constructed as shown in Fig. 8. The dataset contains 83 data points, including 2 outliers and 3 boundary points. INCA can accurately distinguish main cluster points, boundary points and outlier points, as shown in Fig. 9. Data points 41 and 42 are outliers (blue circles in Figure 8), data points 61, 69 and 70 are boundary points (magenta circles in Figure 8), and the rest belong to the main cluster. Figure 10 visually depicts membership.

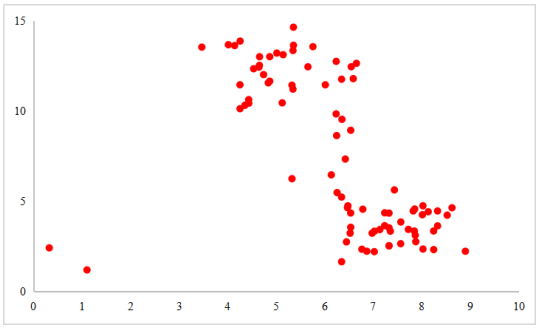


Figure 8. dataset 1

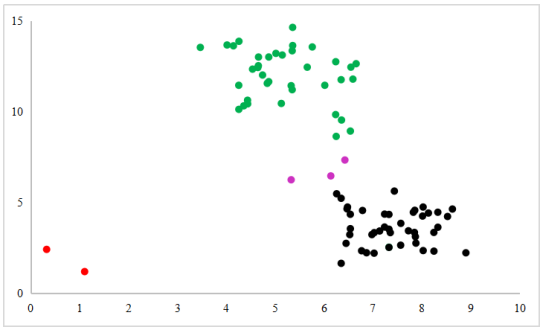


Figure 9. Clustering results of INCA on dataset 1

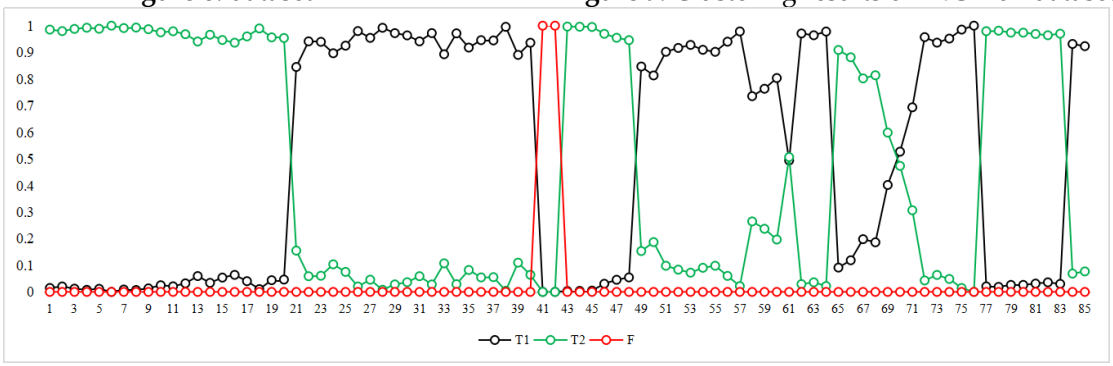


Figure 10. Membership calculated by FKM and INCA on dataset 1

4.2.2. Real dataset

To further evaluate the proposed clustering method, the UCI dataset is considered a standard dataset in the field of machine learning. In this study, the "dermatology", "pima", "TOX-171", "vowel", "ecoli", "iris" and "vote" datasets were selected among other UCI datasets. Table 3 summarizes the number of features, classes, and samples in each data. These datasets are used for traditional clustering methods such as FKM, RSFKM, FC-PFS, NCM and INCM.

Table 3. Datasets

Datasets	No. of instance	No. of feature	No. of class
dermatology	366	34	6
pima	768	8	2
TOX-171	171	5748	4

vowel	528	10	11
ecoli	336	343	8
iris	150	4	3
vote	435	16	2

Table 4 summarizes the accuracy of the proposed method and the FKM, RSFKM, FC-PFS, INCM and NCM methods. The accuracy rates of the proposed method on the "dermatology", "pima", "TOX-171", "vowel", "ecoli", "iris" and "vote" datasets were 82.24%, 74.35%, 51.46%, 40.15%, 76.79%, 98.00% and 84.60%. The accuracy of INCA is higher or second than other comparison algorithms. Table 5 summarizes the mutual information of INCA and FKM, RSFKM, FC-PFS, INCM and NCM methods. The mutual information of INCA is higher or second than other comparison algorithms.

Table 4. ACC of the different datasets

	dermatology	**pima**	**TOX-171**	**vowel**	**ecoli**	**iris**	**vote**
INCM	0.5314	0.6510	0.3918	0.2708	0.6875	0.9466	0.8000
FC-PFS	0.5027	0.6589	0.3977	0.2321	0.6250	0.8933	0.8138
RSFKM	0.8689	0.6602	0.2632	0.2746	0.6518	0.9267	0.8253
NCM	0.5000	0.6302	0.2865	0.2348	0.6280	0.9000	0.8138
FKM	0.6995	0.6563	0.4912	0.3655	0.6280	0.8933	0.8230
INCA	0.8224	0.7435	0.5146	0.4015	0.7679	0.9800	0.8460

Table 5. NMI of the different datasets

	dermatology	**pima**	**TOX-171**	**vowel**	**ecoli**	**iris**	**vote**
INCM	0.0117	0.0022	0.0685	0.2341	0.4867	0.8081	0.2918
FC-PFS	0.3193	0.0317	0.0722	0.2063	0.2614	0.7501	0.3333
RSFKM	0.8477	0.0267	0.0000	0.3027	0.3247	0.7933	0.3644
NCM	0.1998	0.0521	0.0231	0.2168	0.2711	0.7540	0.3297
FKM	0.6070	0.0294	0.2178	0.3915	0.4625	0.7515	0.3359
INCA	0.7240	0.0092	0.2248	0.3933	0.5895	0.9187	0.3636

Figure 11 shows the average accuracy of different algorithms. It can be seen that the average accuracy of INCA is higher than that of other comparison algorithms.

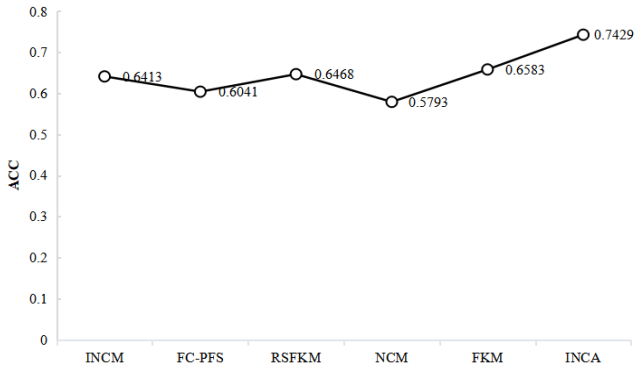


Figure 11. Average accuracy of different algorithms

4.3. Parameter analysis

In this section, the influence of parameters on the clustering results is analyzed. For this task, the Iris was selected for parameter evaluation. In each step, one parameter is changed and the others are fixed. Table 6 reports the results of the clustering methods for different parameter values. In each column, one parameter is considered to have 7 different quantities, while the other parameters are considered to be fixed and the quantities are in the fourth row. Each row in the table is a combination of parameters, and the fourth row is the best combination we chose in our experiments. The reasons for this choice will be discussed in detail in the following chapters.

Based on (13) each data point depends on two factors, namely the distance from the data to the cluster center and the uncertainty of the data, both of which influence each other. The parameter m determines the weighting effect of these factors. If m increases, $\omega_1 I_i T_{ij}$ and $\omega_2 (1-I_i) F_i$ are used more for membership assignments for main clusters and boundary points, respectively, and vice versa. By reducing m, the distance to the cluster center is a more important factor for membership assignment, which is almost the same as FKM. This parameter is 2 in this paper.

The parameters ω_1 and ω_2 are based on equation (18), on the one hand an increase in ω_1 leads to a decrease in T_{ij} and an increase in F_i, which means that the cost function pays more attention to the F set (boundary points) and reduces the accuracy. On the other hand, a smaller number of ω_1 has positive and negative effects on the main and border clusters, respectively. $\omega_1 = 1$ is configured, which is the best balance between the main cluster and the border cluster. The parameter ω_2 has the same effect in equation (19). Figure 12 shows the effect of different parameter combinations on the clustering results.

Table 6. Parameter sensitivity analysis

m	ω_1	ω_2
m=1.3	$\omega_1 = 0.3$	$\omega_2 = 0.5$
ACC=0.9667	ACC=0.9533	ACC=0.9667
m=1.5	$\omega_1 = 0.6$	$\omega_2 = 1.1$
ACC=0.9533	ACC=0.9267	ACC=0.9400
m=1.8	$\omega_1 = 0.7$	$\omega_2 = 1.5$
ACC=0.8800	ACC=0.9533	ACC=0.9400
m=2	$\omega_1 = 1$	$\omega_2 = 2$
ACC=0.9800	ACC=0.9667	ACC=0.9800
m=2.5	$\omega_1 = 1.5$	$\omega_2 = 3$
ACC=0.9300	ACC=0.9200	ACC=0.9567
m=3	$\omega_1 = 2$	$\omega_2 = 4$
ACC=0.9600	ACC=0.9600	ACC=0.9600
m=4	$\omega_1 = 3$	$\omega_2 = 5$
ACC=0.9400	ACC=0.9600	ACC=0.9400

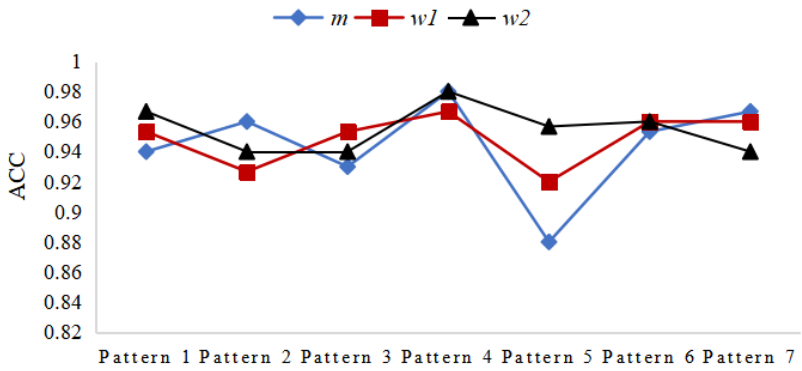


Figure 12. Parameter sensitivity analysis

In this section, the advantages and disadvantages of the proposed method are discussed. Border points and outliers are not considered in methods such as FKM. For example X19: 5, 7, 11, 13 and X24: 5, 7, 11, 13, 17, 19 are not assigned to the main cluster with a high degree of certainty. The reason is that such points are located at the same distance from the center of the main cluster and the center of the boundary cluster.For boundary points, such as X19: 6, 12 and X24: 6, 12, 18, the distances between the two main clusters are equal, but they are forcibly divided into one of the main clusters, which does not meet the actual situation and requirements.

From the above experiments, it can be seen that INCA is robust and the main cluster centers are not forced away from the boundary points. The experimental results show that INCA is more suitable for partitioning data, especially fuzzy and unclear data. Traditional methods only describe the degree of each cluster. For some samples in the boundary between different clusters, it is difficult to determine which group it belongs to. The method proposed in this paper aims to deal with these shortcomings of traditional partitioning methods.

5. Conclusions

The cost function in the neutrosophic set is proposed. Two types of clusters are considered in the proposed cost function, including main clusters and noise clusters. Experiments on different datasets show that INCA can not only deal with outliers and boundary points, but also outperform the comparative methods in both scatter data clustering and real datasets with these shortcomings of traditional partitioning methods.

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

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nanda, S.J.; Gulati, I.; Chauhan R. A k-means galactic swarm optimization based clustering algorithm with otsu's entropy for brain tumor detection [J]. Applied Artificial Intelligence, 2019, 33(2), 152-170.
2. Guo, Y.; Cheng, H.D. New neutrosophic approach to image segmentation[J]. Pattern Recognit, 2009, 42(5), 587-595.
3. Salafian, B.; Kafieh, R.; Rashno, A.; et al. Automatic Segmentation of Choroid Layer in EDI OCT Images Using Graph Theory in Neutrosophic Space[J]. arXiv preprint, 2018.

4. Guo, Y.; Akbulut, Y.; Şengür, A.; et al. An efficient image segmentation algorithm using neutrosophic graph cut[J]. Symmetry, 2017, 9(9), 185.
5. Guo, Y,; Sengür, A. A novel image edge detection algorithm based on neutrosophic set[J]. Computers & Electrical Engineering, 2014, 40(8): 3-25.
6. Akbulut, Y.; Sengu, A.; Guo, Y.; Smarandache, F. Ns-k-nn: Neutrosophicset-based k-nearest neighbors classifier[J]. Symmetry, 2017, 9 (9), 179.
7. Dhar, S.; Kundu, M. K. Accurate segmentation of complex document image using digitalshearlet transform with Neutrosophic Set as Uncertainty Handling Tool[J]. Applied Soft Computing, 2017, 412-426.
8. Nanda, S.J.; Gulati, I.; Chauhan, R.; et al. A k-means-galactic swarm optimization- based clustering algorithm with otsu's entropy for brain tumor detection[J]. Applied Artificial Intelligence, 2018, 33(2), 152–170.
9. Kannan, S. R.; Ramathilagam, S.; Devi, R.; Hines, E. Strong fuzzy c-means in medical image data analysis[J]. Journal of Systems and Software, 2012, 85(11), 2425–2438.
10. Wang, Z.; Song, Q.; Soh, Y.C.; et al. An adaptive spatial information theoretic fuzzy clustering algorithm for image segmentation[J]. Computer Vision and Image Understanding, 2013, 117(10), 1412–1420.
11. Bezdek, J.C.; Ehrlich, R.; Full, W. FCM: The fuzzy c-means clustering algorithm[J]. Computers and Geoences, 1984, 10(2-3), 191–203.
12. Krishnapuram, R.; Keller, J.M. A possibilistic approach to clustering[J]. IEEE Transactions on Fuzzy Systems, 2002, 1(2), 98-110.
13. Xu, J.L.; Han, J.W.; Xiong, K. Robust and sparse fuzzy k-means clustering, Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence [C]. New York: IJCAI, 2016, 2224-2230.
14. Hathaway, R.J.; Davenport, J.W.; Bezdek, J.C. Relational duals of the c-means clustering algorithms[J]. Pattern Recognition, 1989, 22(2), 205-212.
15. Li, X.; Han, Q.;, Qiu, B. A clustering algorithm using skewness-based boundary detection[J]. Neurocomputing, 2017, 275(JAN.31), 618-626.
16. Cui, G.S.; Li, X.L.; Dong, Y.S. Subspace clustering guided convex nonnegative matrix factorization[J]. Neurocomputing, 2018, 38-48.
17. Saxena, A.; Prasad, M.; Gupta, A.; et al. A review of clustering techniques and developments[J]. Neurocomputing, 2017, 267(6), 664-681.
18. Smarandache, F. Neutrosophy, A new branch of pilosophy[J]. Multiple Valued Logic, 2002, 8(3), 297.
19. Thong, P.H.; Son, L.H. Picture Fuzzy Clustering: a new computational intelligence method[J]. Soft Computing, 2016, 20(9), 3549–3562.
20. Li, Q.; Ma, Y.; Smarandache, F. Single-valued neutrosophic clustering algorithm based on tsallis entropy maximization[J]. Axioms, 2018, 7(3).
21. Guo, Y.; Sengur, A. NCM: Neutrosophic c-means clustering algorithm[J]. Pattern Recognition, 2015, 48(8), 2710–2724.
22. Rashno, E.; Minaei-Bidgoli B, Guo, Y. An effective clustering method based on data indeterminacy in neutrosophic set domain[J]. Engineering Applications of Artificial Intelligence, 2020, 89(3), 1-14.
23. Ye, J. Single-valued Neutrosophic minimum spanning tree and its clustering method[J]. Journal of Intelligent Systems, 2014, 23(3), 311–324.
24. Kandasamy, I. Double-valued neutrosophic sets, their minimum spanning trees, and clustering algorithm[J]. Journal of Intelligent Systems, 2018, 27(2), 163-182.
25. Rodriguez, A.; Laio, A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6), 1492-1496.

Received: June 10, 2022. Accepted: September 25, 2022.