*Computers, Materials & Continua* CMC, vol.65, no.2, pp.1203-1220, 2020

# A Direct Data-Cluster Analysis Method Based on Neutrosophic Set Implication

**Sudan Jha[1], Gyanendra Prasad Joshi[2], Lewis Nkenyereya[3], Dae Wan Kim[4, *] and Florentin Smarandache[5]**

**Abstract:** Raw data are classified using clustering techniques in a reasonable manner to create disjoint clusters. A lot of clustering algorithms based on specific parameters have been proposed to access a high volume of datasets. This paper focuses on cluster analysis based on neutrosophic set implication, i.e., a *k*-means algorithm with a threshold-based clustering technique. This algorithm addresses the shortcomings of the *k*-means clustering algorithm by overcoming the limitations of the threshold-based clustering algorithm. To evaluate the validity of the proposed method, several validity measures and validity indices are applied to the Iris dataset (from the University of California, Irvine, Machine Learning Repository) along with *k*-means and threshold-based clustering algorithms. The proposed method results in more segregated datasets with compacted clusters, thus achieving higher validity indices. The method also eliminates the limitations of threshold-based clustering algorithm and validates measures and respective indices along with *k*-means and threshold-based clustering algorithms.

**Keywords:** Data clustering, data mining, neutrosophic set, *k*-means, validity measures, cluster-based classification, hierarchical clustering.

## 1 Introduction

Today, data repositories have become the most favored systems. To name a few, we have relational databases, data mining, and temporal and transactional databases. However, due to the high volume of data in these repositories, the prediction level at the same time has become too complex and tough. Today's scenarios also indicate the diversity of these data (for example, from scientific to medical, geographic to demographic, and financials to marketing). Therefore, the diversity of the data and the extensive volume of those data resulted in the emergence of the field of data mining in recent years [Hautamäki,

[1] School of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, 144411, India.

[2] Department of Computer Science and Engineering, Sejong University, Seoul, 05006, Korea.

[3] Department of Computer and Information Security, Sejong University, Seoul, 05006, Korea.

[4] Department of Business Administration, Yeungnam University, Gyeongsan, 38541, Korea.

[5] University of New Mexico, New Mexico, 87301, USA.

[*] Corresponding Author: Dae Wan Kim. Email: c.kim@ynu.ac.kr.

Received: 19 May 2020; Accepted: 25 June 2020.

Cherednichenko and Kärkkäinen et al. (2005)]. Secondly, grouping data objects and converting them into unknown classes (called clustering) has become a strong tool and a favorite choice in recent years. In clustering, similar data objects are grouped together, and dissimilar data objects are put into other groups. These are called unsupervised classification. In unsupervised classification, analysis is done on dissimilar data objects or raw information, and then, the relationships among them are discovered without any external interference. Several clustering methods exist in the literature, and they are broadly classified into hierarchical-based clustering algorithms and partitioning-based clustering algorithms [Reddy and Vinzamuri (2019); Rodriguez, Comin, Casanova et al. (2019)]. Some other types of clustering (probabilistic clustering, fuzzy-based clustering, density- and grid-based clustering) are also found in the literature [Aggarwal (2019); Nerurkar, Shirke, Chandane et al. (2018); Sánchez-Rebollo, Puente, Palacios et al. (2019); Zhang, He, Jin et al. (2020)].

In this work, we discuss a method geared towards the threshold value concept in a cluster-analysis method based on neutrosophic set implication (NSI). Although the use of this method is still in its infancy, we feature the advantages of the proposed method over a *k*-means algorithm. Neutrosophic systems use confidence, dependency, and falsehood (*c, d, f*) to make uncertainty more certain; in other words, it decreases complexity. A neutrosophic system is a paraconsistency approach because (per the falsehood theory of neutrosophic sets) no event, task, or signal can be perfectly consistent until the job is done [Jha, Son, Kumar et al. (2019)]. We intend to enhance the neutrosophic set in a detailed paraconsistent plan to apply to clustering in various algorithms. Our contribution is to make this approach result-oriented via correlating neutrosophic sets, i.e., confidence and dependency, justifying falsehood.

The rest of the paper is organized as follows. Section 2 presents related work and the advantages of NSI over a *k*-means algorithm. Section 3 discusses basic theory and definitions. Applications of two neutrosophic products (the neutrosophic triangle product and the neutrosophic square product) are described in Section 4. Section 5 discusses the direct neutrosophic cluster analysis method. The performance evaluation of the threshold and *k*-means–based methods are presented in Section 6. Finally, Section 7 concludes the paper.

## 2 Related work

Supervised and unsupervised learning are two fundamental categories of data analysis techniques. A supervised data analysis method includes training in the patterns for inferring a function from labeled training data; an unsupervised data analysis method includes unlabeled data. The unsupervised data analysis method uses an object function to optimize the maximum and minimum similarity among similar and dissimilar objects, respectively. The biggest challenge observed in previous work shows that data clustering is more complicated and challenging than data classification, because it falls under unsupervised learning. The main goal of data clustering is to group similar objects into one group.

Recent works published in data clustering indicates that most of the researchers use *k*-means clustering, hierarchical clustering, and similar techniques. Specially, Hu et al. [Hu, Nurbol, Liu et al. (2010); Sánchez-Rebollo, Puente, Palacios et al. (2019)] have published work in which it can be clearly seen that clustering is difficult because it itself is an

unsupervised learning problem. Most of the times, we use a dataset and are asked to infer structure within it, in this case, the latent clusters or categories in the data. The problem is the classification problems. Though, deep artificial neural networks are very good at classification, but clustering is still a very open problem. For clustering, we lack this critical information. This is why data clustering is more complicated and challenging when unsupervised learning is considered. Authors believe that the best example to illustrate this is to predict whether or not a patient has a common disease based on a list of symptoms.

Many researchers Boley et al. [Boley, Gini, Gross et al. (1999); Arthur and Vassilvitskii (2007); Cheung (2003); Fahim, Salem, Torkey et al. (2006); Khan and Ahmad (2017)] proposed partitioning-based methodologies, such as *k*-means, edge-based strategies and variants. The *k*-means strategy is perhaps the most widely used clustering algorithm, being an iterative process that divides a given dataset into *k* disjoint groups. Jain [Jain (2010)] presented a study that indicated the importance of the widely accepted *k*-means technique. Many researchers have proposed variations of partitioning algorithms to improve the efficiency of clustering algorithms [Celebi, Kingravi and Vela (2013); Erisoglu, Calis and Sakallioglu (2011); Reddy and Jana (2012)]. Finding the optimal solution from a *k*-means algorithm is NP-hard, even when the number of clusters is small [Aloise, Deshpande, Hansen et al. (2009)]. Therefore, a *k*-means algorithm finds the local minimum as approximate optimal solutions.

Nayini et al. [Nayini, Geravand and Maroosi (2018)] overcame *k*-means weaknesses by using a threshold-based clustering method. This work also proposed a partitioning-based method to automatically generate clusters by accepting a constant threshold value as an input. Authors used similarity and threshold measures for clustering to help users to identify the number of clusters. They identified outlier data, and decreased the negative impact on clustering. The time complexity of this algorithm is $O(nk)$, which is better than *k*–means [Mittal, Sharma and Singh (2014)]. In this algorithm, instead of providing initial centroids, only one centroid is taken, which is one of the data objects. Afterwards, the formation of a new cluster depends upon the distance between the existing centroid and the next randomly selected data objects.

Even in the same dataset, clustering algorithms' results can differ from one another, particularly the results from the *k*-means and edge-based system techniques. Halkidi et al. [Halkidi, Batistakis and Vazirgiannis (2000)] proposed quality scheme assessment and clustering validation techniques [ Halkidi, Batistakis and Vazirgiannis (2001)]. Clustering algorithms produce different partitions for different values of the input parameters. The scheme selects best clustering schemes to find the best number of clusters for a specific dataset based on the defined quality index. The quality index validates and assures good candidate estimation based on separation and compactness, two components contained in a quality index.

An index called the Davies–Bouldin index (DBI) was proposed [Davies and Bouldin (1979)] for cluster validation. This validity index is, in fact, a ratio of separation to compactness. In this internal evaluation scheme, the validation is done by evaluating quantities and features inherent in the dataset.

Yeoh et al. [Yeoh, Caraffini and Homapour (2019)] proposed a unique optimized stream (OpStream) clustering algorithm using three variants of OpStream. These variants were

taken from different optimization algorithms, and the best variant was chosen to analyze robustness and resiliency. Uluçay et al. [Uluçay and Şahin (2019)] proposed an algebraic structure of neutrosophic multisets that allows membership sequences. These sequences have a set of real values between 0 and 1. Their proposed neutrosophic multigroup works with the neutrosophic multiset theory, set theory, and group theory. Various methods and applications of a *k* means algorithm for clustering have been worked out recently. Wang et al. [Wang, Gittens and Mahoney (2019)] identifies and extracts a varied collection of cluster structures than the linear *k*-means clustering algorithm. However, kernel *k*-means clustering is computationally expensive when the non-linear feature map is high-dimensional and there are many input points. On the other hand, Jha et al. [ Jha, Kumar, Son et al. (2019)] uses a different clustering technique to resolve stock market prediction. They have used a rigorous machine learning approaches in hand to hand with clustering of the high volume of data.

This paper studied the applications of hierarchical (ward, single, average, centroid and complete linkages) and *k*-means clustering techniques in air pollution studies of almost 40 years data.

## 3 Neutrosophic basics and definitions

In this section, we proceed with fundamental definitions of neutrosophic theory that include truth (T), indeterminacy (I) and falsehood (F). The degree of T, I, and F are evaluated with their respective membership functions. The respective derivations are explained below.

### *3.1 Definitions in the neutrosophic set*

Let $S$ be a space for objects with generic elements, $s \in S$ . A neutrosophic set (NS), $N$ in $S$, is characterized by a truth membership function, $Q_N$ , an indeterminacy membership function, $I_N$ , and a falsehood membership function, $F_N$ . Here $Q_N(s)$ , $Q_N(s)$ , and $Q_N(s)$ are real standard or non-standard subsets of $[0^-,1^+]$ such that $Q_N, I_N, F_N : S \rightarrow [0^-,1^+]$. Tab. 1 shows the acronyms and nomenclatures used in the definitions.

**Table 1:** Nomenclatures and acronyms

| Nomenclature/ acronyms | Definition | Nomenclature/ acronyms | Definition |
|---|---|---|---|
| $S$ | Space of objects | OpStream | Optimized Stream clustering |
| $N$ | Neutrosophic set | T | Truth |
| $Q_N$ | Truth membership function | I | Indeterminacy |
| $I_N$ | Indeterminacy membership function | F | Falsehood |
| $F_N$ | Falsehood membership function | NS | Neutrosophic sets |

| Nomenclature/ acronyms | Definition | Nomenclature/ acronyms | Definition |
|---|---|---|---|
| $Q_N(s)$ | Singleton subinterval or subsets of $S$ | □ | Square product |
| $Q_N(s)$ | Singleton subinterval or subsets of $S$ | ◅ | Triangular product |
| $F_N(s)$ | Singleton subinterval or subsets of $S$ | $\Phi$ | Lukasiewicz implication operator |
| NSI | Neutrosophic set implication | CIN | Intuitionistic neutrosophic implication |
| CNR | Complex neutrosophic Relations | CNS | Complex neutrosophic sets |

A singleton set, which is also called as a unit set, contains exactly one element. For example, the set *{null}* is a singleton containing the element null. The term is also used for a 1-tuple, a sequence with one member. A singleton interval is an interval of one such elements. Assume that functions $Q_N(s)$ , $Q_N(s)$ , and $F_N(s)$ are singleton subintervals or subsets of the real standard, such that with $Q_N(s): S \to [0,1], \mathrm{I}_N(s): S \to [0,1], F_N(s): S \to \in [0,1]$ . Then, a simplification of neutrosophic set $N$ is denoted by

$$N = \left\{ \left( s, Q_N(s), I_N(s), F_N(s) \right) : s \in S \right\}$$

with $0 \le Q_N(s) + I_N(s) + F_N(s) \le 3$. It is a simplified neutrosophic set, i.e., a subclass of the neutrosophic set. This subclass of the neutrosophic set covers the notions of the interval neutrosophic set and the single-valued neutrosophic set [Haibin, Florentin, Yanqing et al. (2010); Ye (2014)].

### *3.2 Operations in the neutrosophic set*

Assume that $S_1$ and $S_2$ are two neutrosophic sets, where $N_1 = \{\langle s; Q_1(s); I_1(s); F_1(s) \rangle \mid s \in S\}$ and $\mathrm{N}_2 = \{\langle s; Q_2(s); I_2(s); F_2(s) \rangle \mid s \in S\}$ . Then

a. $N_1 \subseteq N_2$ if and only if $Q_1(s) \le Q_2(s); I_1(s) \ge I_2(s); F_1(s) \ge F_2(s)$,

b. $N_1^c = \{\langle s; F_1(s); I_1(s); Q_1(s) \rangle \mid s \in S\}$ ,

c. $N_1 \cap N_2 = \{\langle x; \min\{T_1(x); Q_2(x)\}; \max\{I_1(s); I_2(s)\}; \max\{F_1(s); F_2(s)\} \rangle \mid s \in S\}$ ,

d. $N_1 \cup N_2 = \{\langle s; \max\{Q_1(s); Q_2(s)\}; \min\{I_1(s); I_2(s)\}; \min\{F_1(s); F_2(s)\} \rangle \mid s \in S\}$

### *3.3 Definition of states of a set*

*Former and latter*

Let us assume that $V_i(i=1,2)$ are two ordinary subsets with an ordinary relation: $R \subseteq V_1 \times V_2$. Then, for any $q, f \in V_2$, $Rf = \{q \mid qRf\}$ is called a former set, and $qR = \{q \mid qRf\}$ is called the latter set.

### *3.4 Definition of neutrosophic algebraic products*

*Triangle product and square product*

Let us assume that $V_j\ (j=1,2,3)$ are ordinary subsets $R_1 \subseteq V_1 \times V_2$ and $R_2 \subseteq V_2 \times V_3$, such that triangle product $R_1 \triangleleft R_2 \subset V_1 \times V_3$ of $V_1$ and $V_3$ can be defined as follows:

$$eV_1 \triangleleft V_2 g \Leftrightarrow eV_1 \subset V_2 g, \text{ for any } (e,g) \in V_1 \times V_2 \tag{1}$$

Correspondingly, $R_1 \square R_2$, a square product, can be defined as follows:

$$eV_1 \square V_2 g \Leftrightarrow eV_1 = V_2 g, \text{for any } (e,g) \in V_1 \times V_2 \tag{2}$$

where $eV_1 \subset V_2 g$ if and only if $eV_1 \subset V_2 g$ and $eV_1 \supset V_2 g$.

### *3.5 Definition of neutrosophic implication operators*

If $\alpha$ is a binary operation on $[0, 1]$, and if $\alpha(0, 0, 0) = \alpha(0, 0, 1) = \alpha(0, 1, 1) = \alpha(1, 1, 0) = \alpha(1, 0, 1) = \alpha(1, 1, 1) = 1$ and $\alpha(1, 0, 0) = 0$

In this case, $L$ is called a neutrosophic implication operator.

For any $a,b,c \in [0, 1]$, $\alpha(a, b, c)$ is a neutrosophic implication operator. If we extend the Lukasiewicz implication operator to the neutrosophic implication operator, then $\Phi(a, b, c) = \min(1-a+b+c, 1, 1)$.

### *3.6 Definition of generalized neutrosophic products*

Let us extend the Lukasiewicz implication operator to a neutrosophic valued environment. If we consider membership degrees $Q_\mu$ and $Q_\nu$ of $\mu$ and $\nu$ only, for any two neutrosophic valued environments, $\mu = (Q_\mu, \mathrm{I}_\mu, \mathrm{F}_\mu)$ and $\nu = (Q_\nu, \mathrm{I}_\nu, \mathrm{F}_\nu)$, then $\min\{1-Q_\mu+Q_\nu, 1, 1\}$ is unable to reflect the dominance of the neutrosophic environment, and therefore, we consider the indeterminacy and non-membership $I_\mu, I_\nu$ and $F_\mu, F_\nu$ as well. Now, we define neutrosophic Lukasiewicz implication operator $\Phi(\mu, \nu)$ based on the neutrosophic valued environmental components and the Lukasiewicz implication operator. The membership degree, the degree of indeterminacy, and the non-membership degree of $\Phi(\mu, \nu)$ are expressed as follows:

$$\begin{aligned}&\min\left\{1,\ \min\left\{1-Q_\mu+Q_\nu,\ 1-I_\mu+I_\nu,\ 1-F_\mu+F_\nu\right\}\right\}\\&\quad=\min\left\{1,\ 1-Q_\mu+Q_\nu,\ 1-I_\nu+I_\mu,\ 1-F_\nu+F_\mu\right\}\end{aligned}$$

and

$$\begin{aligned}&\max\left\{0,\ \min\left\{1-\left(1-Q_\mu+Q_\nu\right),1-\left(1-I_\nu+I_\mu\right),1-\left(1-F_\nu+F_\mu\right)\right\}\right\}\\&\quad=\max\left\{0,\ \min\left\{Q_\mu-Q_\nu,I_\nu-I_\mu,F_\nu-F_\mu\right\}\right\}\end{aligned}\tag{3}$$

respectively; i.e.,

$$\Phi(\mu,\ \nu)=\begin{pmatrix}\min\left\{1,\ 1-T_\mu+T_\nu,\ 1-I_\nu+I_\mu,\ 1-F_\nu+F_\mu\right\},\\ \max\left\{0,\ \min\left\{Q_\mu-Q_\nu,I_\nu-I_\mu,F_\nu-F_\mu\right\}\right\}\end{pmatrix}\tag{4}$$

Let us prove that the value of $\Phi(\mu,\ \nu)$ satisfies the conditions of the neutrosophic valued environment. In fact, from Eq. (4), we have

$$\begin{aligned}&\min\left\{1,\ 1-Q_\mu+Q_\nu,\ 1-I_\nu+I_\mu,\ 1-F_\nu+F_\mu\right\}\geq 0\\&\max\left\{0,\ \min\left\{Q_\mu-Q_\nu,I_\nu-I_\mu,F_\nu-F_\mu\right\}\right\}\geq 0\end{aligned}\tag{5}$$

and since

$$\begin{aligned}&\max\left\{0,\ \min\left\{Q_\mu-Q_\nu,I_\nu-I_\mu,F_\nu-F_\mu\right\}\right\}=\\&\qquad 1-\min\left\{1,\ \max\left\{1-Q_\mu+Q_\nu,1-I_\nu+I_\mu,1-F_\nu+F_\mu\right\}\right\}\end{aligned}\tag{6}$$

and

$$\begin{aligned}&\min\left\{1,\ \max\left\{1-Q_\mu+Q_\nu,1-I_\nu+I_\mu,1-F_\nu+F_\mu\right\}\right\}\\&\qquad\geq\min\left\{1,\ 1-Q_\mu+Q_\nu,\ 1-I_\nu+I_\mu,\ 1-F_\nu+F_\mu\right\}\end{aligned}\tag{7}$$

then

$$\begin{aligned}&1-\min\left\{1,\ \max\left\{1-Q_\mu+Q_\nu,1-I_\nu+I_\mu,1-F_\nu+F_\mu\right\}\right\}\\&\qquad+\min\left\{1,1-Q_\mu+Q_\nu,1-I_\nu+I_\mu,1-F_\nu+F_\mu\right\}\leq 3\end{aligned}\tag{8}$$

This shows that the value of $\Phi(\mu,\ \nu)$ derived through Eq. (6) is a neutrosophic environment.

Along with the neutrosophic Lukasiewicz implication, the square product, and the traditional triangle product, we introduce the neutrosophic triangle product and the neutrosophic square product as follows.

### *3.7 Definitions of neutrosophic relations*

Neutrosophic relations are based on the conventional arithmetic, algebraic and geometric theories which are used in dealing various real time engineering problems. Neutrosophic relations also relate various neutrosophic sets.

*Triangle product*

Let $\mu=\{\mu_1,\mu_2,...,\mu_p\}$, $\nu=\{\nu_1,\nu_2,...,\nu_q\}$, and $\omega=\{\omega_1,\omega_2,...,\omega_r\}$ be three neutrosophic valued sets. $S_1\in N(\mu\times\nu)$ and $S_2\in N(\nu\times\omega)$ are two neutrosophic relations, and then, a neutrosophic triangle product, $S_1\triangleleft S_2\in N(\mu\times\nu)$ of $S_1$ and $S_2$, can be expressed as follows:

$$(S_1\triangleleft S_2)(\mu_i,\ \nu_j)=\begin{pmatrix}\frac{1}{q}\sum_{k=1}^{q}Q_{X_1(\mu_i,\ \omega_k)\to x_2(\omega_k,\ \nu_j)},\\ \frac{1}{q}\sum_{k=1}^{q}I_{X_1(\mu_i,\ \omega_k)\to x_2(\omega_k,\ \nu_j)},\\ \frac{1}{q}\sum_{k=1}^{q}Q_{X_1(\mu_i,\ \omega_k)\to x_2(\omega_k,\ \nu_j)}\end{pmatrix} \tag{9}$$

for any $(\mu_i,\ \nu_j)\in(\mu,\ \nu), i=1,2,...,p,\ j=1,2,...,r$, where $\to$ represents the neutrosophic Lukasiewicz implication.

*Square product*

Similarly, we define the neutrosophic square product, $(S_1\square S_2)\in N(\mu\times\nu)$ of $S_1$ and $X_2$, as follows:

$$(S_1\square S_2)(\mu_i,\ \nu_j)=\min_{1\le k\le q}\begin{pmatrix}Q_{\min(S_1(\mu_i,\ \omega_k)\to S_2(\omega_k,\ \nu_j),S_2(\omega_k,\ \nu_j)\to S_1(\mu_i,\ \omega_k))},\\ \mathrm{I}_{\min(S_1(\mu_i,\ \omega_k)\to S_2(\omega_k,\ \nu_j),S_2(\omega_k,\ \nu_j)\to S_1(\mu_i,\ \omega_k))},\\ F_{\min(S_1(\mu_i,\ \omega_k)\to S_2(\omega_k,\ \nu_j),S_2(\omega_k,\ \nu_j)\to S_1(\mu_i,\ \omega_k))},\end{pmatrix} \tag{10}$$

for any $(\mu_i,\ \nu_j)\in(\mu,\ \nu), i=1,2,...,p,\ j=1,2,...,r$.

Denote $\mathrm{x}_{ik}$ as $S(\mu_i,\ \omega_k)$ for short, similar to the others, for convenience. Subsequently, we can simplify Eq. (9) and Eq. (10) as follows:

$$(S_1\triangleleft S_2)(\mu_i,\ \nu_j)=\begin{pmatrix}\frac{1}{q}\sum_{j=1}^{q}Q_{S_{ik}\to S_{kj}},\\ \frac{1}{q}\sum_{k=1}^{q}I_{S_{ik}\to S_{kj}},\\ \frac{1}{q}\sum_{k=1}^{q}F_{S_{ik}\to S_{kj}}\end{pmatrix} \tag{11}$$

$$(S_1 \square S_2)(\mu_i,\ \nu_j) = \min_{1\le k\le q} \begin{pmatrix} Q_{\min(S_{ik}\to S_{kj}, S_{kj}\to S_{ik}),} \\ \mathrm{I}_{\min(S_{ik}\to S_{kj}, S_{kj}\to S_{ik}),} \\ F_{\min(S_{ik}\to S_{kj}, S_{kj}\to S_{ik})} \end{pmatrix} \tag{12}$$

Indeed, the neutrosophic triangle product and the neutrosophic square product are firmly related to each other. That is, the neutrosophic triangle product is the basis of the neutrosophic square product, and because of that, $(S_1 \square S_2)(\mu_i,\ \nu_j)$ is directly derived from $(S_1 \triangleleft S_2)(\mu_i,\ \nu_j)$ and $(S_2 \triangleleft S_1)(\mu_i,\ \nu_j)$.

**4 Applications of the two neutrosophic products**

In this subsection, we use the neutrosophic triangle product to compare multi-attribute decision making with neutrosophic information. Subsequently, we use the neutrosophic square product for constructing an anneutrosophic similarity matrix. This anneutrosophic similarity matrix is used for analyzing the neutrosophic clustering method.

Assume a multiple attribute decision making issue. Let $W = \{w_1, w_2, ..., w_p\}$ and $N = \{n_1, n_2, ....., n_q\}$ define sets of *p* alternatives and *q* attributes, respectively. The attribute values (also called a characteristic) of each alternative $w_i$ under all the attributes $N_j\,(j = 1,2,...,m)$ represent the neutrosophic set. We make a decision based on the multiple attributes:

$$w_i = \left\{ \left\langle N_j,\ Q_{w_i}(N_j),\ I_{w_i}(N_j), F_{w_i}(N_j) \right\rangle \middle| N_j \in N \right\},\ i = 1,2,..,n \text{ and } j = 1,2,...,m \tag{13}$$

where $Q_{w_i}(N_j)$ denotes the degree of membership, $I_{w_i}(N_j)$ denotes the degree of indeterminacy, and $F_{w_i}(N_j)$ denotes the degree of non-membership of $w_i$ to $N_j$. Apparently, the degree of uncertainty of $w_i$ to $N_j$ is $\Psi_{w_i}(N_j) = 3 - Q_{w_i}(N_j) - I_{w_i}(N_j) - F_{w_i}(N_j)$.

Let $s_{ij} = (Q_{ij},\ I_{ij},\ F_{ij}) = (Q_{wi}(N_j),\ I_{wi}(N_j),\ F_{wi}(N_j))$ be a neutrosophic value. An $n \times m$ neutrosophic decision matrix, $S = (s_{ij})_{n\times m}$, can be constructed based on the neutrosophic valued set $s_{ij}\,(i = 1,2,...,n;\ j = 1,2,...,\mathrm{m})$.

### *4.1 Neutrosophic triangle product's application*

The characteristic vectors of two alternatives for the issues described above, say $S_i$ and $S_j$, can be expressed as $S_i = (s_{i1}, s_{i2}, ..., s_{im})$ and $S_j = (s_{j1}, s_{j2}, ..., s_{jm})$, respectively. The neutrosophic triangle product can be calculated as follows:

$$\left(S_i \triangleleft S_j^{-1}\right)_{ij} = \begin{pmatrix} \frac{1}{m}\sum_{k=1}^{m} Q_{s_{ik}\to s_{jk}}, \\ \frac{1}{m}\sum_{k=1}^{m} I_{s_{ik}\to s_{jk}}, \\ \frac{1}{m}\sum_{k=1}^{m} F_{x_{ik}\to x_{jk}} \end{pmatrix} \tag{14}$$

This shows the degree of the alternative, $w_j$, for preferred alternative $w_i$, where $S_j^{-1}$ is the inverse of $S_j$ and can be defined as $\left(S_j^{-1}\right)_{kj} = \left(S_j\right)_{kj} = s_{jk}$, $Q_{s_{ik}\to s_{jk}}$, $I_{s_{ik}\to s_{jk}}$ and $F_{s_{ik}\to s_{jk}}$.

Similarly, we can calculate

$$\left(S_j \triangleleft S_i^{-1}\right)_{ji} = \begin{pmatrix} \frac{1}{m}\sum_{k=1}^{m} Q_{s_{jk}\to s_{ik}}, \\ \frac{1}{m}\sum_{k=1}^{m} I_{s_{jk}\to s_{ik}}, \\ \frac{1}{m}\sum_{k=1}^{m} F_{s_{jk}\to s_{ik}} \end{pmatrix} \tag{15}$$

This shows that degree alternative $w_i$ is preferred to alternative $w_j$. The alternatives ordering $w_i$ and $w_j$ can be obtained from Eqs. (14) and (15). In fact,

a. if $\left(S_i \triangleleft S_j^{-1}\right)_{ij} > \left(S_j \triangleleft S_i^{-1}\right)_{ji}$, alternative $w_j$ is preferred to $w_i$;

b. if $\left(S_i \triangleleft S_j^{-1}\right)_{ij} = \left(S_j \triangleleft S_i^{-1}\right)_{ji}$, there is similarity between $w_i$ and $w_j$;

c. if $\left(S_i \triangleleft S_j^{-1}\right)_{ij} < \left(S_j \triangleleft S_i^{-1}\right)_{ji}$, then $w_i$ is preferred to $w_j$.

### *4.2 Neutrosophic square product's application*

As we know from Eq. (10), mathematically, neutrosophic square product $(S_1 \times S_2)_{ij}$ can be deciphered as follows: $(S_1 \times S_2)_{ij}$ measures the degree of similarities of the $i^{th}$ row of

neutrosophic matrix $S_1$ and the $j^{th}$ row of neutrosophic matrix $S_2$. Therefore, considering the issue expressed at the start of Section 4, $(S_i \times S_j^{-1})_{ij}$ expresses the similarity of alternatives $w_i$ and $w_j$. The following formula can be used for constructing a neutrosophic similarity matrix for $w_i = (i = 1, 2, ..., n)$.

$$sim(w_i, w_j) = \left(S_i \square S_j^{-1}\right)_{ij} = \min_{1 \le k \le n} \begin{pmatrix} Q_{\min(s_{ik} \to s_{jk}, s_{jk} \to s_{ik})}, \\ I_{\min(s_{ik} \to s_{jk}, s_{jk} \to s_{ik})}, \\ F_{\min(s_{ik} \to s_{jk}, s_{jk} \to s_{ik})} \end{pmatrix} \tag{16}$$

Eq. (16) has the following desirable properties:

1. $sim(w_i, w_j)$ is the neutrosophic value.
2. $sim(w_i, w_i) = (1, 0)$ $(i = 1, 2, ..., n)$.
3. $sim(w_i, w_j) = sim(w_j, w_i)$ $(i = 1, 2, ..., n)$.

*Proof for property 1*

We can prove that $sim(w_i, w_j)$ is the neutrosophic value.

Since the results $s_{ik} \longrightarrow s_{jk}$ and $s_{jk} \longrightarrow s_{ik}$ are all neutrosophic valued sets as proven previously, then $\begin{pmatrix} Q_{\min(s_{ik} \to s_{jk}, s_{jk} \to s_{ik})}, \\ I_{\min(s_{ik} \to s_{jk}, s_{jk} \to s_{ik})}, \\ F_{\min(s_{ik} \to s_{jk}, s_{jk} \to s_{ik})} \end{pmatrix}$ is the neutrosophic value for any $k$.

*Proof for property 2*

Since

$$sim(w_i, w_i) = \left(S_i \square S_i^{-1}\right)_{ii} = \min_{1 \le k \le n} \begin{pmatrix} Q_{\min(s_{ik} \to s_{ik}, s_{ik} \to s_{ik})}, \\ I_{\min(s_{ik} \to s_{ik}, s_{ik} \to s_{ik})}, \\ F_{\min(s_{ik} \to s_{ik}, s_{ik} \to s_{ik})} \end{pmatrix}$$

we know from definition (10) that $sim(w_i, w_i) = (1, 0)$.

*Proof for property 3*

Since

$$sim\left(w_i,\ w_j\right)=\left(S_i \square S_j^{-1}\right)_{ij}=\min_{1\le k\le n}\begin{pmatrix} Q_{\min\left(s_{ik\to}s_{jk},s_{jk}\to s_{ik}\right)},\\ I_{\min\left(s_{ik\to}s_{jk},s_{jk}\to s_{ik}\right)},\\ F_{\min\left(s_{ik\to}s_{jk},s_{jk}\to s_{ik}\right)} \end{pmatrix}$$

$$=\min_{1\le k\le m}\begin{pmatrix} Q_{\min\left(s_{jk}\to s_{ik},s_{ik\to}s_{jk}\right)},\\ I_{\min\left(s_{jk}\to s_{ik},s_{ik\to}s_{jk}\right)},\\ F_{\min\left(s_{jk}\to s_{ik},s_{ik\to}s_{jk}\right)} \end{pmatrix}$$

$$=\left(X_j \square X_i\right)=sim\left(w_j,\ w_i\right)$$

then, $sim\left(w_i,\ w_j\right)=sim\left(w_j,\ w_i\right)\ \left(i,j=1,2,...,n\right)$.

At that point, from the above analyses, we can determine that Eq. (16) satisfies the neutrosophic similarity relation conditions. Thus, this can be used to construct a neutrosophic similarity matrix.

## 5 Direct neutrosophic cluster analysis method

After constructing a neutrosophic similarity matrix with the abovementioned method, the equivalent matrix is not required before cluster analysis. The required cluster analysis results can be obtained with the neutrosophic equivalent matrix, starting with the neutrosophic similarity matrix. In fact, Luo [Luo (1989)] proposed a direct method for clustering fuzzy sets. This method considers only membership degrees of fuzzy sets. Our proposed direct neutrosophic cluster analysis technique considers the enrollment degrees, indeterminacy degrees, and non-participation degrees of the neutrosophic esteemed set under the neutrosophic conditions presented below. The proposed method is based on Luo's method, which includes following stages.

**Stage A.** Let $S=\left(s_{ij}\right)_{n\times n}$ be the neutrosophic similarity matrix, where $s_{ij}=\left(Q_{ij},\ I_{ij},\ F_{ij}\right)\ \left(i,j=1,2,...,n\right)$ is a neutrosophic valued set for determining the confidence level, $\lambda_1$. Select one of the elements, $S$, which obeys the following principles.

a. Rank the degrees of membership of $s_{ij}\ \left(i,j=1,2,...,n\right)$ in descending order. Take $\lambda_1=\left(Q_{\lambda_1},\ I_{\lambda_1},\ F_{\lambda_1}\right)=\left(Q_{i_1j_1},\ I_{i_1j_1},\ F_{i_1j_1}\right)$, where $Q_{i_1j_1}=\max_{i,j}\left\{Q_{ij}\right\}$.

b. If there exist two neutrosophic valued sets, $\left(Q_{i_1j_1},\ I_{i_1j_1},\ F_{i_1j_1}\right)$ and $\left(\overline{Q}_{i_1j_1},\ \overline{I}_{i_1j_1},\ \overline{F}_{i_1j_1}\right)$ in (1), such that $I_{i_1j_1} \neq \overline{I}_{i_1j_1}$ and $F_{i_1j_1} \neq \overline{F}_{i_1j_1}$ (without loss of generality, let $I_{i_1j_1} < \overline{I}_{i_1j_1}$ and $F_{i_1j_1} < \overline{F}_{i_1j_1}$ ), then we choose the first one as $\lambda_1$ , i.e., $\lambda_1 = \left(Q_{i_1j_1},\ I_{i_1j_1},\ F_{i_1j_1}\right)$.

Then, for each alternative $w_i$ , let

$$\left[w_i\right]_S^{(1)} = \left\{w_j \,\middle|\, s_{ij} = \lambda_1\right\} \tag{17}$$

Here, $w_i$ and all alternatives in $\left[w_i\right]_S^{(1)}$ are clustered into one category, and other alternatives are clustered into another category.

**Stage B.** Select the confidence level, $\lambda_2 = \left(Q_{\lambda_2},\ I_{\lambda_2},\ F_{\lambda_2}\right) = \left(Q_{i_2j_2},\ I_{i_2j_2},\ F_{i_2j_2}\right)$, with $Q_{i_2j_2} = \max\limits_{(i,j)\neq(i_1,j_1)}\left\{Q_{ij}\right\}$, specifically if there exist at least two neutrosophic esteemed sets where the membership degrees have the same value as $Q_{i_2j_2}$. At that point, we can follow strategy (b) in Stage A. Now, let alternatives $\left[w_i\right]_S^{(2)}$ be $\left\{w_j \,\middle|\, s_{ij} = \lambda_2\right\}$, and then, $w_i$ and all the alternatives are clustered into one type. Let the merger of $\left[w_i\right]_S^{(1)}$ and $\left[w_i\right]_S^{(2)}$ be $\left[w_i\right]_S^{(1,2)}$. Then, the merged alternatives $\left[w_i\right]_S^{(1,2)} = \left\{w_j \,\middle|\, s_{ij} \in \left\{\lambda_1, \lambda_2\right\}\right\}$, and therefore, $w_i$ and all alternatives in $\left[w_i\right]_S^{(1,2)}$ are clustered into one set. The other alternatives remain unaltered.

**Stage C.** In this stage, we take other confidence levels and analyze clusters according to the procedure in Stage B. The procedure is carried out until all alternatives are clustered into one category. One of the significant advantages of the proposed direct neutrosophic cluster analysis method is that cluster analysis can be acknowledged by simply depending on the subscripts of the alternatives. We observed from the process described above that, in this method, getting even an λ-cutting matrix is not necessary.

In real-world application scenarios, we simply need to affirm their areas in the neutrosophic similarity matrix after choosing some appropriate confidence levels, and afterward, we can get the kinds of considered objects on the basis of their area subscripts.

## 6 Performance evaluation

For the performance evaluation, a *k*-means algorithm and a threshold-based algorithm were used on the Iris dataset from the University of California, Irvine (UCI) Machine Learning Repository. A variable number of clusters (from 2 to 10) were generated for the

experiments. For the *k*-means and threshold-based algorithms, cluster number is the input parameter. The *k* data objects were selected randomly in the *k*-means algorithm (*k* was also taken as an initial centroid of the clusters). On the other hand, only one object was selected randomly in the threshold-based method. The selected object was assigned as the initial centroid of the cluster, and was a member of the first cluster. We observed that this method generates more segregated and compact clusters. Finally, we observed that there was significant enhancement in the indices of validity. The following mathematical analysis proves the above statements.

For any cluster-based intuitionistic neutrosophic implication, let $X(T_i, F_a) \rightarrow Y(T_j, F_b)$, where T and F depict truthfulness and falsehood.

Then, we can define various classes of cluster-based neutrosophic set (CNSS) implications, as expressed below:

$$CNSS=([(1-T_i)\, f \vee T_j]\, F \wedge [(1-f_b)f \vee fX\,],\, fY_f \wedge (1-T_i)) \quad (18)$$

The proposed new cluster-based intuitionistic neutrosophic (CIN) implication is now extended with $X(T_i, ia, fX) \mathcal{N} \rightarrow Y(T_j, iY, fY)$, as follows:

$CIN1\ (T_i f/If \rightarrow T_j,\ iXf\ iY \wedge,\ fXf\ fY \wedge)$

where $T_i f/If \rightarrow T_j$ is any cluster of intuitionistic neutrosophic implications, while *f* is any $\wedge$ neutrosophic conjunction:

$CIN2\ (T_i f/If \rightarrow T_j,\ iXf\ iY\ \vee,\ fXf\ fY \wedge),$ where f is any $\vee$ fuzzy disjunction:

CI*N*3 ($T_i$ f /*I*f $\rightarrow T_j$, $i$X+$i$Y 2 , $f$Xf $f$Y $\wedge$ )

CI*N*4 ($T_i$ f/*I*f $\rightarrow T_j$, $i$X+$i$Y 2 , $f$X+$f$Y 2 )

Referring to the definition proposed by Broumi et al. [Broumi, Smarandache and Dhar (2014)], the classical logical equivalence and predicate relationship now becomes

$(X \rightarrow Y) \leftrightarrow (\neg X \vee Y)$, where, $(X\ N \rightarrow Y)\ N \leftrightarrow (NX \neg N\ Y \vee)$

The above class of neutrosophic implications can now be depicted with the operators ($NX \neg N\ Y \vee$ ). Let us have two cluster-based neutrosophic propositions: X⟨0.3, 0.4, 0.2⟩ and Y⟨0.7, 0.1, 0.4⟩.

Then, X $N \rightarrow$ Y has the neutrosophic truth value of X Y$N \vee N \neg$ , i.e., ⟨0.2, 0.4, 0.3⟩ ⟨ $N$ 0.7, 0.1, 0.4⟩ $\vee$ , or ⟨max{0.2, 0.7}, min{0.4, 0.1}, min{0.3, 0.4}⟩, or ⟨0.7, 0.1, 0.3⟩.

Therefore,

$N\langle t, i, f\rangle = \langle f, i, t\rangle \neg$ for neutrosophic negation

and

$\langle t_1, i_1, f_1\rangle\ \langle t_2, i_2, f_2\ N\rangle \vee = \langle \max\{t_1, t_2\}, \min\{i_1, i_2\}, \min\{f_1, f_2\}\rangle$ for the neutrosophic disjunction.

The dataset that we referred to from Stappers et al. [Stappers, Cooper, Brooke et al. (2016)] and [Systems (2020)] contains 16,259 spurious examples caused by radio frequency interference (RFI)/noise, and 1,639 real pulsar examples with each candidate having eight continuous variables. The first four variables are obtained from the integrated pulse profile. This is an array of continuous variables that describe a longitude-resolved version of the

signal. The remaining four variables were similarly obtained from the dispersion measure (DM)-SNR curve. These are summarized in Tab. 2.

Tab. 2 shows a dataset describing a sample of pulsar candidates collected during the high time–resolution universe survey. The first column is the mean of the integrated profile. Mean1 is the mean of the DM-SNR curve, and SD1 is the standard deviation of the DM-SNR curve. Finally, ET1 is the excess kurtosis of the DM-SNR curve, and Skewness1 is the skewness of the DM-SNR curve.

**Table 2:** Pulsar candidate samples collected during the high time–resolution universe survey

| Mean | SD | ET | Skewness | Mean1 | SD1 | ET1 | Skewness1 | T/F |
|---|---|---|---|---|---|---|---|---|
| 140.5625 | 55.68378 | -0.23457 | -0.69965 | 3.199833 | 19.11043 | 7.975532 | 74.24222 | 0 |
| 102.5078 | 58.88243 | 0.465318 | -0.51509 | 1.677258 | 14.86015 | 10.57649 | 127.3936 | 0 |
| 103.0156 | 39.34165 | 0.323328 | 1.051164 | 3.121237 | 21.74467 | 7.735822 | 63.17191 | 0 |
| 136.7500 | 57.17845 | -0.06841 | -0.63624 | 3.642977 | 20.95928 | 6.896499 | 53.59366 | 0 |
| 88.72656 | 40.67223 | 0.600866 | 1.123492 | 1.17893 | 11.46872 | 14.26957 | 252.5673 | 0 |
| 93.57031 | 46.69811 | 0.531905 | 0.416721 | 1.636288 | 14.54507 | 10.62175 | 131.3940 | 0 |
| 119.4844 | 48.76506 | 0.03146 | -0.11217 | 0.999164 | 9.279612 | 19.20623 | 479.7566 | 0 |
| 130.3828 | 39.84406 | -0.15832 | 0.38954 | 1.220736 | 14.37894 | 13.53946 | 198.2365 | 0 |
| 107.2500 | 52.62708 | 0.452688 | 0.170347 | 2.33194 | 14.48685 | 9.001004 | 107.9725 | 0 |
| 107.2578 | 39.49649 | 0.465882 | 1.162877 | 4.079431 | 24.98042 | 7.397080 | 57.78474 | 0 |
| 142.0781 | 45.28807 | -0.32033 | 0.283953 | 5.376254 | 29.0099 | 6.076266 | 37.83139 | 0 |
| 133.2578 | 44.05824 | -0.08106 | 0.115362 | 1.632107 | 12.00781 | 11.97207 | 195.5434 | 0 |
| 134.9609 | 49.55433 | -0.1353 | -0.08047 | 10.69649 | 41.34204 | 3.893934 | 14.13121 | 0 |
| 117.9453 | 45.50658 | 0.325438 | 0.661459 | 2.83612 | 23.11835 | 8.943212 | 82.47559 | 0 |
| 138.1797 | 51.52448 | -0.03185 | 0.046797 | 6.330268 | 31.57635 | 5.155940 | 26.14331 | 0 |
| 114.3672 | 51.94572 | -0.0945 | -0.28798 | 2.738294 | 17.19189 | 9.050612 | 96.6119 | 0 |
| 109.6406 | 49.01765 | 0.137636 | -0.25670 | 1.508361 | 12.0729 | 13.36793 | 223.4384 | 0 |
| 100.8516 | 51.74352 | 0.393837 | -0.01124 | 2.841137 | 21.63578 | 8.302242 | 71.58437 | 0 |
| 136.0938 | 51.691 | -0.04591 | -0.27182 | 9.342809 | 38.0964 | 4.345438 | 18.67365 | 0 |
| 99.36719 | 41.5722 | 1.547197 | 4.154106 | 27.55518 | 61.71902 | 2.208808 | 3.66268 | 1 |
| 100.8906 | 51.89039 | 0.627487 | -0.02650 | 3.883779 | 23.04527 | 6.953168 | 52.27944 | 0 |
| 105.4453 | 41.13997 | 0.142654 | 0.320420 | 3.551839 | 20.75502 | 7.739552 | 68.51977 | 0 |
| 95.86719 | 42.05992 | 0.326387 | 0.803502 | 1.832776 | 12.24897 | 11.24933 | 177.2308 | 0 |

In Tab. 2, the mean of the integrated profile is compared with pulsar candidates that vary significantly with Mean1. Here, Mean1 is the mean of popular candidates at high time resolution. The dataset that we have referred from Stappers et al. [Stappers, Cooper, Brooke et al. (2016)] and [Systems (2020)] that contains 16,259 spurious examples caused by radio-frequency interference (RFI) or noise, and 1,639 real pulsar examples with each candidate having 8 continuous variables. The first four variables are obtained from the integrated pulse profile. This is an array of continuous variables that describe a longitude-resolved version of the signal. The remaining 4 variables are similarly obtained from the dispersion measure (DM)-SNR curve. These are summarized in Tab. 2.

## 7 Conclusions and future work

One of the major issues in data clustering is the selection of the right candidates. In addition, the appropriate algorithm to choose the right candidates has been a challenging issue in cluster analysis, especially for an efficient approach that best fits the right sets of data. In this paper, a cluster analysis method based on neutrosophic set implication generates the clusters automatically and overcomes the limitation of the *k*-means algorithm. Our proposed method generates more segregated and compact clusters and achieves higher validity indices, in comparison to the mentioned algorithms. The experimentation carried out in this work focused on cluster analysis based on NSI through a *k*-means algorithm along with a threshold-based clustering technique. We found that the proposed algorithm eliminates the limitations of the threshold-based clustering algorithm. The validity measures and respective indices applied to the Iris dataset along with *k*-means and threshold-based clustering algorithms prove the effectiveness of our method.

Future work will handle data clustering in various dynamic domains using neutrosophic theory. We also intend to apply a periodic search routine by using propagations between datasets of various domains. The data clustering used by our proposed algorithms was found to be workable in a low computational configuration. In the future, we will also use more datasets.

**Funding Statement:** The work of Gyanendra Prasad Joshi is supported by Sejong University new faculty research grand.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

**Aloise, D.; Deshpande, A.; Hansen, P.; Popat, P.** (2009): NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, vol. 75, no. 2, pp. 245-248.

**Arthur, D.; Vassilvitskii, S.** (2007): k-means++: the advantages of careful seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027-1035, Philadelphia, PA, USA.

**Boley, D.; Gini, M.; Gross, R.; Han, E. H.; Hastings, K. et al.** (1999): Partitioning-based clustering for web document categorization. *Decision Support Systems*, vol. 27, no. 3, pp. 329-341.

**Broumi, S.; Smarandache, F.; Dhar, M.** (2014): Rough neutrosophic sets. *Italian Journal of Pure and Applied Mathematics*, vol. 32, pp. 493-502.

**Celebi, M. E.; Kingravi, H. A.; Vela, P. A.** (2013): A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, vol. 40, no. 1, pp. 200-210.

**Cheung, Y. M.** (2003): k*-means: a new generalized k-means clustering algorithm. *Pattern Recognition Letters*, vol. 24, no. 15, pp. 2883-2893.

**Davies, D. L.; Bouldin, D. W.** (1979): A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224-227.

**Erisoglu, M.; Calis, N.; Sakallioglu, S.** (2011): A new algorithm for initial cluster centers in k-means algorithm. *Pattern Recognition Letters*, vol. 32, no. 14, pp. 1701-1705.

**Fahim, A. M.; Salem, A. M.; Torkey, F. A.; Ramadan, M. A.** (2006): Efficient enhanced k-means clustering algorithm. *Journal of Zhejiang University: Science*, vol. 7, no. 10, pp. 1626-1633.

**Haibin, W.; Florentin, S.; Yanqing, Z.; Rajshekhar, S.** (2010): Single valued neutrosophic sets. *Multispace and Multistructure*, vol. 4, no. 2010, pp. 410-413.

**Halkidi, M.; Batistakis, Y.; Vazirgiannis, M.** (2001): On clustering validation techniques. *Journal of Intelligent Information Systems*, vol. 17, no. 2, pp. 107-145.

**Halkidi, M.; Vazirgiannis, M.; Batistakis, Y.** (2000): Quality scheme assessment in the clustering process. *Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery*, *LNCS*, vol. 1910, no.1, pp. 265-276.

**Hautamäki, V.; Cherednichenko, S.; Kärkkäinen, I.; Kinnunen, T.; Fränti, P.** (2005): Improving k-means by outlier removal. *Proceedings of the Scandinavian Conference on Image Analysis,* LNCS, Joensuu, Finland, vol. 3540, no. 1, pp. 978-87.

**Hu, L.; Nurbol, Liu, Z.; He, J.; Zhao, K.** (2010): A time-stamp frequent pattern-based clustering method for anomaly detection. *IETE Technical Review*, vol. 27, no. 3, pp. 220-227.

**Jain, A. K.** (2010): Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651-666.

**Jha, S.; Kumar, R.; Son, L. H.; Chatterjee, J. M.; Khari, M. et al.** (2019): Neutrosophic soft set decision making for stock trending analysis. *Evolving Systems*, vol. 10, no. 4, pp. 621-627.

**Jha, S.; Son, L. H.; Kumar, R.; Priyadarshini, I.; Smarandache, F. et al.** (2019): Neutrosophic image segmentation with dice coefficients. *Measurement: Journal of the International Measurement Confederation*, vol. 134, pp. 762-772.

**Khan, S. S.; Ahmad, A.** (2017): Cluster center initialization algorithm for k-means clustering-cluster center initialization algorithm for k-means clustering. *Pattern Recognition Letters*, vol. 25, no. 11, pp. 1293-1302.

**Luo, C. Z.** (1989): *Introduction to fuzzy sets*. Beijing Normal University Publishing House, China.

**Mittal, M.; Sharma, R. K.; Singh, V. P.** (2014): Validation of k-means and threshold-based clustering method. *International Journal of Advancements in Technology*, vol. 5, no. 2, pp. 153-160.

**Nayini, S. E. Y.; Geravand, S.; Maroosi, A.** (2018): A novel threshold-based clustering method to solve k-means weaknesses. *International Conference on Energy, Communication, Data Analytics and Soft Computing,* pp. 47-52.

**Nerurkar, P.; Shirke, A.; Chandane, M.; Bhirud, S.** (2018): Empirical analysis of data clustering algorithms. *Procedia Computer Science*, vol. 125, no. 2018, pp. 770-779.

**Reddy, D.; Jana, P. K.** (2012): Initialization for k-means clustering using Voronoi diagram. *Procedia Technology*, vol. 4, pp. 395-400.

**Sánchez-Rebollo, C.; Puente, C.; Palacios, R.; Piriz, C.; Fuentes, J. P. et al.** (2019): Detection of jihadism in social networks using big data techniques supported by graphs and fuzzy clustering. *Complexity*, vol. 2019, no. 1238780, pp. 1-13.

**Stappers, B. W.; Cooper, S.; Brooke, J. M.; Knowles, J. D.** (2016): Fifty years of pulsar candidate selection: From simple filters to a new principled real-time classification approach. *Monthly Notices of the Royal Astronomical Society*, vol. 459, no. 1, pp. 1104-1123

**Systems, C. M. L. I.** (2020): Machine learning repository. *Center for Machine Learning and Intelligent Systems*, https://archive.ics.uci.edu/ml/index.php.

**Uluçay, V.; Şahin, M.** (2019): Neutrosophic multigroups and applications. *Mathematics*, vol. 7, no. 95, pp. 1-17.

**Wang, S.; Gittens, A.; Mahoney, M. W.** (2019): Scalable kernel k-means clustering with Nyström approximation: relative-error bounds. *The Journal of Machine Learning Research*, vol. 20, no. 1, pp. 431-479.

**Yeoh, J. M.; Caraffini, F.; Homapour, E.; Santucci, V.; Milani, A.** (2019): A clustering system for dynamic data streams based on metaheuristic optimisation. *Mathematics*, vol. 7, no. 1229, pp. 1-24.

**Zhang, X.; He, Y.; Jin, Y.; Qin, H.; Azhar, M. et al.** (2020): A robust k-means clustering algorithm based on observation point mechanism. *Complexity*, vol. 2020, no. 3650926, pp. 1-11.