Clustering Neutrosophic Data Sets and Neutrosophic Valued Metric Spaces

Ferhat Taş [1], Selçuk Topal [2,*] and Florentin Smarandache [3]

[1] Department of Mathematics, Faculty of Science, İstanbul University, İstanbul 34134, Turkey;
    tasf@istanbul.edu.tr
[2] Department of Mathematics, Faculty of Science and Arts, Bitlis Eren University, Bitlis 13000, Turkey
[3] Department of Mathematics, University of New Mexico, Gallup, NM 87301, USA; smarand@unm.edu
*   Correspondence: stopal@beu.edu.tr

Abstract: In this paper, we define the neutrosophic valued (and generalized or *G*) metric spaces for the first time. Besides, we newly determine a mathematical model for clustering the neutrosophic big data sets using *G*-metric. Furthermore, relative weighted neutrosophic-valued distance and weighted cohesion measure, is defined for neutrosophic big data set. We offer a very practical method for data analysis of neutrosophic big data although neutrosophic data type (neutrosophic big data) are in massive and detailed form when compared with other data types.

Keywords: G-metric; neutrosophic G-metric; neutrosophic sets; clustering; neutrosophic big data; neutrosophic logic

1. Introduction and Preliminaries

Neutrosophic Logic is a neonate study area in which each proposition is estimated to have the proportion (percentage) of truth in a subset T, the proportion of indeterminacy in a subset I, and the proportion of falsity in a subset F. We utilize a subset of truth (or indeterminacy, or falsity), instead of a number only, since in many situations we do not have ability to strictly specify the proportions of truth and of falsity but only approximate them; for instance, a proposition is between 25% and 55% true and between 65% and 78% false; even worse: between 33% and 48% or 42 and 53% true (pursuant to several observer), and 58% or between 66% and 73% false. The subsets are not essential intervals, but any sets (open or closed or half open/half-closed intervals, discrete, continuous, intersections or unions of the previous sets, etc.) in keeping with the given proposition. Zadeh initiated the adventure of obtaining meaning and mathematical results from uncertainty situations (fuzzy) [1]. Fuzzy sets brought a new dimension to the concept of classical set theory. Atanassov introduced intuitionistic fuzzy sets including membership and non-membership degrees [2]. Neutrosophy was proposed by Smarandache as a computational approach to the concept of neutrality [3]. Neutrosophic sets consider membership, non-membership and indeterminacy degrees. Intuitionistic fuzzy sets are defined by the degree of membership and non-membership and, uncertainty degrees by the 1-(membership degree plus non-membership degree), while the degree of uncertainty is evaluated independently of the degree of membership and non-membership in neutrosophic sets. Here, membership, non-membership, and degree of uncertainty (uncertainty), such as degrees of accuracy and falsity, can be evaluated according to the interpretation of the places to be used. It depends entirely on the subject area (the universe of discourse). This reveals a difference between neutrosophic set and intuitionistic fuzzy set. In this sense, the concept of neutrosophic is a possible solution and representation of problems in various fields. Two detailed and mathematical fundamental differences between relative truth (IFL) and absolute truth (NL) are:

(i)   NL can discern absolute truth (truth in all possible worlds, according to Leibniz) from the relative truth (truth in at least one world) because NL (absolute truth) = 1^+ while IFL (relative truth) = 1. This has practice in philosophy (see the Neutrosophy). The standard interval [0, 1] used in IFL has been extended to the unitary non-standard interval $]^-0, 1^+[$ in NL. Parallel earmarks for absolute or relative falsehood and absolute or relative indeterminacy are permitted in NL.

(ii)   There is no limit on T, I, F other than they are subsets of $]^-0, 1^+[$, thus: $^-0 \le \inf T + \inf I + \inf F \le \sup T + \sup I + \sup F \le 3^+$ in NL. This permissiveness allows dialetheist, paraconsistent, and incomplete information to be described in NL, while these situations cannot be described in IFL since F (falsehood), T (truth), I (indeterminacy) are restricted either to $t + i + f = 1$ or to $t^2 + f^2 \le 1$, if T, I, F are all reduced to the points t, i, f respectively, or to $\sup T + \sup I + \sup F = 1$ if T, I, F are subsets of [0, 1] in IFL.

Clustering data is one of the most significant problems in data analysis. Useful and efficient algorithms are needed for big data. This is even more challenging for neutrosophic data sets, particularly those involving uncertainty. These sets are elements of some decision-making problems, [4–8]. Several distances and similarities are used for decision-making problems [9,10]. Algorithms for the clustering big data sets use the distances (metrics). There are some metrics used in algorithms to analysis neutrosophic data sets: Hamming, Euclidean, etc. In this paper, we examine clustering of neutrosophic data sets via neutrosophic valued distances.

The big data notion is a new label for the giant size of data–both structured and unstructured—that overflows several sectors on a time-to-time basis. It does not mean overall data are significant and the significant aspect is to obtain desired specific data interpretation. Big data can be analyzed for pre-cognition that make possible more consistent decisions and strategic having positions. Doug Laney [11] sort to make the definition of big data the three Vs and Veracity widespread: (1) Velocity: This refers to dynamic data and captures data streams in near real-time. Data streams in at an exceptional speed and must be dealt with in a well-timed mode. (2) Variety: Data comes in all types of formats—from structured, numeric data in traditional databases to formless materials. On the one hand, variety denotes to the various sources and types of organized and formless data. Storing data is made from sources like worksheets and databases. (3) Volume: Organizations gather data from a range of sources, including social media, business operations, and data from the sensor or machine to machine. (4) Veracity: It mentions to the biases, noise, and anomaly in data. That corresponds with the question "Is the data that is being put in storage and extracted meaningful to the problem being examined?".

In this paper, we also focus on K-sets cluster algorithm which is a process of analyzing data with the aim of evaluating neutrosophic big data sets. The K-sets cluster is an unrestrained type of learning that is used when one wants to utilize unlabeled data, [12]. The goal of the algorithm is to find groups of data with the number of groups represented by variable K. The algorithm works iteratively to set-aside each data point obtained to one of the K groups based on the properties obtained. The data points are clustered according to feature similarity. Instead of identifying groups before examining patterns, clustering helps to find and analyze naturally occurring groups. "Choosing K" has the goal of "how the number of groups can be determined". Each center of a congregation is a collection of property values describe the groups that emerged. Analysis of centroid feature weights can be used to qualitatively interpret what kind of group is represented by each cluster. The algorithm finds the clusters and data set labels for a particular pre-chosen K. To have the number of clusters in the data, the user must run the K-means clustering algorithm for a range of K values and compare the results. In general, there is no technique to determine a specific K value, but a precise estimate can be obtained using the following methods. In general, one of the metrics used to compare the results between the different K values as the average distance between the data points and their cluster synthesis. As the number of sets increases, it will always reduce the distance to the data points, while the K increment will always lower this metric as other criteria, and when K is the same as the number of data points, reaching zero will be excessive. Thus, this metric cannot be used as a single purpose. Rather, the

average distance to the center as a function of K is plotted where the shear rate falls sharply, it can be used to determine K approximately.

A number of other techniques are available for verification of K, including cross-validation, information criteria, information theoretical jump method, and G-tools algorithm. In addition, monitoring the distribution of data points between groups provides information about how the algorithm splits data for each K. K-sets algorithms base on the measurement of distances of sets. A distance is a measurement of how far apart each pair of elements of a given set is. Distance functions in mathematics and many other computational sciences are important concepts. They have wide usage areas, for example, the goal of quantifying a dissimilarity (or equivalently similarity) between two objects, sets or set of sets in some sense. However, due to the massive, complicated and different type data sets today, definitions of distance functions are required to be more generalized and detailed. For this purpose, we define a novel metric for similarity and distance to give Neutrosophic Valued-Metric Spaces (NVGMS). We present relative weighted measure definition and finally K-sets algorithm after given the definition of NVGMS.

Some readers who are unfamiliar with the topic in this paper need to have a natural example to understand the topic well. There is a need for earlier data in everyday life to give a natural example for the subject first described in this paper. There is no this type of data (we mean neutrosophic big data) in any source, but we will give an example of how to obtain and cluster such a data in Section 6 of the paper. If we encounter a sample of neutrosophic big data in the future, we will present the results with a visual sample as a technical report. In this paper, we have developed a mathematically powerful method for the notion of concepts that are still in its infancy.

1.1. *G -Metric Spaces*

Metric space is a pair of (*A*, *d*), where *A* is a non-empty set and *d* is a metric which is defined by a certain distance and the elements of the set *A*. Some metrics may have different values such as a complex-valued metric [13,14]. Mustafa and Sims defined *G*-metric by generalizing this definition [15]. Specifically, fixed point theorems on analysis have been used in *G*-metric spaces [16,17].

Definition 1. *Let A be a non-empty set and d be a metric on A, then if the following conditions hold, the pair (A, d) is called a metric space. Let* $x, y, z \in A$

(1) $d(x, y) \geq 0$, *(non-negativity)*
(2) $d(x, y) = 0 \Leftrightarrow x = y$, *(identity)*
(3) $d(x, y) = d(y, x)$, *(symmetry)*
(4) $d(x, z) \leq d(x, y) + d(y, z)$ *(triangle inequality).*

where $d : A \times A \to R^+ \cup \{0\}$.

Definition 2. *[15] Let A be a non-empty set. A function* $G : A \times A \times A \to [0, +\infty)$ *is called G-distance if it satisfies the following properties:*

(1) $G(x, y, z) = 0$ *if and only if* $x = y = z$,
(2) $G(x, x, y) \neq 0$ *whenever* $x \neq y$,
(3) $G(x, x, y) \leq G(x, y, z)$ *for any* $x, y, z \in A$, *with* $z \neq y$,
(4) $G(x, y, z) = G(x, z, y) = ...$ *(symmetric for all elements),*
(5) $G(x, y, z) \leq G(x, a, a) + G(a, y, z)$ *for all* $a, x, y, z \in A$ *(Rectangular inequality).*

The pair (*A*, *G*) is called a *G*-metric space. Moreover, if *G*-metric has the following property then it is called symmetric: $G(x, x, y) = G(x, y, y), \forall x, y \in A$.

Example 1. *In 3-dimensional Euclidean metric space, one can assume the G-metric space* $\left(E^3, G \right)$ *as the following:*

$$G(x, y, z) = 2\left(\|x \times y\| + \|z \times y\| + \|x \times z\|\right)$$

where $x, y, z \in E^3$ *and* $\|.\times.\|$ *represent the norm of the vector product of two vectors in* E^3. *It is obvious that it satisfies all conditions in the Definition 2 because of the norm has the metric properties, and it is symmetric.*

Example 2. *Let (A, d) is a metric space. Then*

$$G(x, y, z) = d(x, y) + d(y, z) - d(x, z)$$

is a G-metric, where $x, y, z \in A$. *The fact that d is a metric indicates that it has triangle inequality. Thus, G is always positive definite.*

Proposition 1. *[17] Let (A, G) be a G-metric space then a metric on A can be defined from a G-metric:*

$$d_G(x, y) = G(x, x, y) + G(x, y, y)$$

1.2. Neutrosophic Sets

　　Neutrosophy is a generalized form of the philosophy of intuitionistic fuzzy logic. In neutrosophic logic, there is no restriction for truth, indeterminacy, and falsity and they have a unit real interval value for each element neutrosophic set. These values are independent of each other. Sometimes, intuitionistic fuzzy logic is not enough for solving some real-life problems, i.e., engineering problems. So, mathematically, considering neutrosophic elements are becoming important for modelling these problems. Studies have been conducted in many areas of mathematics and other related sciences especially computer science since Smarandache made this philosophical definition, [18,19].

Definition 3. *Let E be a universe of discourse and* $A \subseteq E$. $A = \left\{\left(x, T(x), I(x), F(x)\right) : x \in E\right\}$ *is a neutrosophic set or single valued neutrosophic set (SVNS), where* $T_A, I_A, F_A : A \rightarrow\ ]^-0,1^+[$ *are the truth-membership function, the indeterminacy-membership function and the falsity-membership function, respectively. Here,* $^-0 \le T_A(x) + I_A(x) + F_A(x) \le 3^+$.

Definition 4. *For the SVNS A in E, the triple* $\langle T_A, I_A, F_A \rangle$ *is called the single valued neutrosophic number (SVNN).*

Definition 5. *Let* $n = \langle T_n, I_n, F_n \rangle$ *be an SVNN, then the score function of* n *can be given as follow:*

$$s_n = \frac{1 + T_n - 2I_n - F_n}{2} \tag{1}$$

where $s_n \in [-1, 1]$.

Definition 6. *Let* $n = \langle T_n, I_n, F_n \rangle$ *be an SVNN, then the accuracy function of n can be given as follow:*

$$h_n = \frac{2 + T_n - I_n - F_n}{3} \tag{2}$$

where $h_n \in [0, 1]$.

Definition 7. *Let* n_1 *and* n_2 *be two SVNNs. Then, the ranking of two SVNNs can be defined as follows:*

(I)　*If* $s_{n_1} > s_{n_2}$, *then* $n_1 > n_2$;

(II) If $s_{n_1} = s_{n_2}$ and $h_{n_1} \geq h_{n_2}$, then $n_1 \geq n_2$.

2. Neutrosophic Valued Metric Spaces

The distance is measured via some operators which are defined in some non-empty sets. In general, operators in metric spaces have zero values, depending on the set and value.

2.1. Operators

Definition 8. [20,21], *Let* A *be non-empty SVNS and* $x = \langle T_x, I_x, F_x \rangle, y = \langle T_y, I_y, F_y \rangle$ *be two SVNNs. The operations that addition, multiplication, multiplication with scalar* $\alpha \in \mathbb{R}^+$, *and exponential of SVNNs are defined as follows, respectively:*

$$x \oplus y = \left\langle T_x + T_y - T_x T_y, I_x I_y, F_x F_y \right\rangle$$

$$x \odot y = \left\langle T_x T_y, I_x + I_y - I_x I_y, F_x + F_y - F_x F_y \right\rangle$$

$$\alpha x = \left\langle 1 - \left(1 - T_x\right)^\alpha, I_x^{\;\alpha}, F_x^{\;\alpha} \right\rangle$$

$$x^\alpha = \left\langle T_x^{\;\alpha}, 1 - \left(1 - I_x\right)^\alpha, 1 - \left(1 - F_x\right)^\alpha \right\rangle$$

From this definition, we have the following theorems as a result:

Theorem 1. *Let* $x = \langle T_x, I_x, F_x \rangle$ *be an SVNN. The neutral element of the additive operator of the set* A *is* $0_A = \langle 0,1,1 \rangle$.

Proof. Let $x = \langle T_x, I_x, F_x \rangle$ and $0_A = \langle T_0, I_0, F_0 \rangle$ are two SVNN and using Definition 8 we have

$$x \oplus 0_A = \left\langle T_x + T_0 - T_x T_0, I_x I_0, F_x F_0 \right\rangle = \left\langle T_x, I_x, F_x \right\rangle$$
$$\Rightarrow \left\langle T_0, I_0, F_0 \right\rangle = \left\langle 0,1,1 \right\rangle = 0_A$$

(There is no need to show left-hand side because the operator is commutative in every component). □

To compare the neutrosophic values based on a neutral element, we shall calculate the score and accuracy functions of a neutral element $0_A = \langle 0,1,1 \rangle$, respectively:

$$s_0 = \frac{1 + T_0 - 2I_0 - F_0}{2} = -1 \quad \text{and} \quad h_0 = \frac{2 + T_0 - I_0 - F_0}{3} = 0$$

Theorem 2. *Let* $x = \langle T_x, I_x, F_x \rangle$ *be an SVNN. The neutral element of the multiplication operator of the* A *is* $1_A = \langle 1,0,0 \rangle$.

Proof. Let $x = \langle T_x, I_x, F_x \rangle$ and $1_A = \langle T_1, I_1, F_1 \rangle$ are two SVNN and using Definition 8 we have

$$x \odot 1_A = \left\langle T_x T_1, I_x + I_1 - I_x I_1, F_x + F_1 - F_x F_1 \right\rangle = \left\langle T_x, I_x, F_x \right\rangle$$
$$\Rightarrow \left\langle T_1, I_1, F_1 \right\rangle = \left\langle 1,0,0 \right\rangle = 1_A$$

In addition, score and accuracy functions of the neutral element $1_A = \langle 1,0,0 \rangle$ are

$$s_1 = \frac{1 + T_1 - 2I_1 - F_1}{2} = 1 \text{ and } h_1 = \frac{2 + T_1 - I_1 - F_1}{3} = 1, \text{ respectively. } \square$$

2.2. Neutrosophic Valued Metric Spaces

In this section, we consider the metric and generalized metric spaces in the neutrosophic meaning.

Definition 9. *Ordering in the Definition 6 gives an order relation for elements of the conglomerate SVNN. Suppose that the mapping $d : X \times X \to A$, where X and A are SVNS, satisfies:*

(I) $0_A \le d(x,y)$ and $d(x,y) = 0_A \Leftrightarrow s_x = s_y$ and $h_x = h_y$ for all $x, y \in X$.

(II) $d(x,y) = d(y,x)$ for all $x, y \in X$.

Then d is called a neutrosophic valued metric on X, and the pair (X,d) is called neutrosophic valued metric space. Here, the third condition (triangular inequality) of the metric spaces is not suitable for SVNS because the addition is not ordinary addition.

Theorem 3. *Let (X,d) be a neutrosophic valued metric space. Then, there are relationships among truth, indeterminacy and falsity values:*

(I) $0 < T(x,y) - 2I(x,y) - F(x,y) + 3$ and if $s_o = s_d$ then $0 < T(x,y) - I(x,y) - F(x,y) + 2$.

(II)   If $d(x,y) = 0_A \Leftrightarrow T(x,y) = 0, I(x,y) = F(x,y) = 1$.

(III) $T(x,y) = T(y,x)$, $I(x,y) = I(y,x)$, $F(x,y) = F(y,x)$ so, each distance function must be symmetric.

where $T(.,.)$, $I(.,.)$ and $F(.,.)$ are distances within themselves of the truth, indeterminacy and falsity functions, respectively.

Proof.

(I)
$$0_A < d(x,y) \Leftrightarrow \langle 0,1,1 \rangle < \langle T(x,y), I(x,y), F(x,y) \rangle$$
$$\Leftrightarrow s_0 < s_d \Leftrightarrow -1 < \frac{1 + T(x,y) - 2I(x,y) - F(x,y)}{2}$$
$$\Leftrightarrow 0 < T(x,y) - 2I(x,y) - F(x,y) + 3$$

(II)
$$d(x,y) = d(y,x) \Leftrightarrow \langle T(x,y), I(x,y), F(x,y) \rangle = \langle T(y,x), I(y,x), F(y,x) \rangle$$
$$\Leftrightarrow T(x,y) = T(y,x), I(x,y) = I(y,x), F(x,y) = F(y,x)$$
$\square$

Example 3. *Let A be non-empty SVNS and $x = \langle T_x, I_x, F_x \rangle, y = \langle T_y, I_y, F_y \rangle$ be two SVNNs. If we define the metric $d : X \times X \to A$, as:*

$$d(x,y) = \langle T(x,y), I(x,y), F(x,y) \rangle = \langle |T_x - T_y|, 1 - |I_x - I_y|, 1 - |F_x - F_y| \rangle$$

then

(I)
$$0 < |T_x - T_y| - 2\left(1 - |I_x - I_y|\right) - \left(1 - |F_x - F_y|\right) + 3$$
$$\Rightarrow 0 < |T_x - T_y| + 2|I_x - I_y| + |F_x - F_y|$$
Then it satisfies the first condition.

(II)   Since the properties of the absolute value function, this condition is obvious.

So, (X, d) *is a neutrosophic-valued metric space.*

3. Neutrosophic Valued G -Metric Spaces

Definition 10. *Let X and A be a non-empty SVNS. A function* $G : X \times X \times X \to A$ *is called neutrosophic valued* G *-metric if it satisfies the following properties:*

(1) $G(x, y, z) = 0_A$ *if and only if* $x = y = z$,

(2) $G(x, x, y) \neq 0_A$ *whenever* $x \neq y$,

(3) $G(x, x, y) \leq G(x, y, z)$ *for any* $x, y, z \in X$ *, with* $z \neq y$,

(4) $G(x, y, z) = G(x, z, y) = ...$ *(symmetric for all elements).*

The pair (*X*, *G*) is called a neutrosophic valued *G*-metric space.

Theorem 4. *Let (X, G) be a neutrosophic valued G-metric space then, it satisfies followings:*

(1) $T(x, x, x) = 0, I(x, x, x) = F(x, x, x) = 1.$

(2)   *Assume* $x \neq y$ *, then* $T(x, y, z) \neq 0, I(x, y, z) \neq 1, F(x, y, z) \neq 1.$

(3) $0 < T(x, y, z) - T(x, x, y) + 2\big(I(x, x, y) - I(x, y, z)\big) + F(x, x, y) - F(x, y, z)$

(4) $T(x, y, z), I(x, y, z)$ and $F(x, y, z)$ *are symmetric for all elements.*

where $T(., ., .)$ *,* $I(., ., .)$ *and* $F(., ., .)$ *are G-distance functions of truth, indeterminacy and falsity values of the element of the set, respectively.*

Proofs are made in a similar way to neutrosophic valued metric spaces.

Example 4. *Let* X *be non-empty SVNS and the G-distance function defined by:*

$$G(x, y, z) = \frac{1}{3}\big(d(x, y) \oplus d(x, z) \oplus d(y, z)\big)$$

where $d(., .)$ *is a neutrosophic valued metric. The pair (X, G) is obviously a neutrosophic valued G-metric space because of* $d(., .)$ *. Further, it has commutative properties.*

4. Relative Weighted Neutrosophic Valued Distances and Cohesion Measures

The relative distance measure is a method used for clustering of data sets, []. We define the relative weighted distance, which is a more sensitive method for big data sets.

Let $x_i = \left\langle T_{x_i}, F_{x_i}, I_{x_i} \right\rangle \in A$(non-empty SVNS)$, i = 0..n$ be   SVNNs.   Then   neutrosophic weighted average operator of these SVNNs is defined as:

$$M_a(A) = \sum_{i=1}^{n} \chi_i x_i = \left\langle 1 - \prod_{i=1}^{n}\left(1 - T_{x_i}\right)^{\chi_i}, \prod_{i=1}^{n}\left(I_{x_i}\right)^{\chi_i}, \prod_{i=1}^{n}\left(F_{x_i}\right)^{\chi_i} \right\rangle$$

where χ_i is weighted for the *i* th data. For a given a neutrosophic data set $W = \left\{w_1, w_2, w_3, ..., w_n\right\}$ and a neutrosophic valued metric *d*, we define a relative neutrosophic valued distance for choosing another reference neutrosophic data and compute the relative neutrosophic valued distance as the average of the difference of distances for all the neutrosophic data $w_i \in W$.

Definition 11. *The relative neutrosophic valued distance from a neutrosophic data* w_i *to another neutrosophic data* w_j *is defined as follows:*

$$RD\left(w_i \| w_j\right) = \frac{1}{n} \sum_{w_k \in W} \left(d\left(w_i, w_j\right) \ominus d\left(w_i, w_k\right)\right)$$

Here, since *T*, *I*, *F* values of SVNNs cannot be negative, we can define the expression $d\left(w_i, w_j\right) \ominus d\left(w_i, w_k\right)$ as the distance between these two neutrosophic-valued metrics. Furthermore, the distance of metrics is again neutrosophic-valued here so, a related neutrosophic-valued distance can be defined as:

$$
\begin{aligned}
d\left(w_i, w_j\right) \ominus d\left(w_i, w_k\right) &= \left\langle T(w_i, w_j), I(w_i, w_j), F(w_i, w_j)\right\rangle \ominus \left\langle T(w_i, w_k), I(w_i, w_k), F(w_i, w_k)\right\rangle \\
&= \left\langle 1 - \left|T(w_i, w_j) - \left(T(w_i, w_k) - 1\right)^2\right|, 1 - \left|I(w_i, w_j) - I(w_i, w_k)^2\right|, 1 - \left|F(w_i, w_j) - F(w_i, w_k)^2\right|\right\rangle
\end{aligned}
\tag{3}
$$

The difference operator $\ominus$ generally is not a neutrosophic-valued metric (or G-metric). We used some abbreviations for saving space.

$$
\begin{aligned}
RD\left(w_i \| w_j\right) &= \frac{1}{n} \sum_{w_k \in W} \left(d\left(w_i, w_j\right) \ominus d\left(w_i, w_k\right)\right) \\
&= d\left(w_i, w_j\right) \ominus \frac{1}{n} \sum_{w_k \in W} d\left(w_i, w_k\right) \\
&= \left\langle T(w_i, w_j), I(w_i, w_j), F(w_i, w_j)\right\rangle \ominus \frac{1}{n}\left(d\left(w_i, w_1\right) \oplus d\left(w_i, w_2\right) \oplus \ldots \oplus d\left(w_i, w_n\right)\right) \\
&= \left\langle T(w_i, w_j), I(w_i, w_j), F(w_i, w_j)\right\rangle \\
&\quad \ominus \frac{1}{n}\left[\left\langle T(w_i, w_1), I(w_i, w_1), F(w_i, w_1)\right\rangle \oplus \ldots \oplus \left\langle T(w_i, w_1), I(w_i, w_1), F(w_i, w_1)\right\rangle\right] \\
&= \left\langle T(w_i, w_j), I(w_i, w_j), F(w_i, w_j)\right\rangle \\
&\quad \ominus \frac{1}{n}\left[\left\langle \sum_{k \in W} T(w_i, w_k) - \prod_{k \in W} T(w_i, w_k), \prod_{k \in W} I(w_i, w_k), \prod_{k \in W} F(w_i, w_k)\right\rangle\right] \\
&= \left\langle T(w_i, w_j), I(w_i, w_j), F(w_i, w_j)\right\rangle \\
&\quad \ominus \left\langle 1 - \left[1 - \sum_{k \in W} T(w_i, w_k) + \prod_{k \in W} T(w_i, w_k)\right]^{1/n}, \prod_{k \in W} I(w_i, w_k)^{1/n}, \prod_{k \in W} F(w_i, w_k)^{1/n}\right\rangle \\
&= \left\langle T_1, I_1, F_1\right\rangle \ominus \left\langle T_2, I_2, F_2\right\rangle \\
&= \left\langle 1 - \left|T_1 - \left(T_2 - 1\right)^2\right|, 1 - \left|I_1 - I_2^2\right|, 1 - \left|F_1 - F_2^2\right|\right\rangle
\end{aligned}
$$

where T_1, I_1, F_1 and T_2, I_2, F_2 are the first, second, and third elements of SVNN in the previous equation, respectively.

Definition 12. *The relative weighted neutrosophic valued distance from a neutrosophic data w_i to another neutrosophic data w_j is defined as follows:*

$$RD_\chi\left(w_i \| w_j\right) = \sum_{\substack{w_k \in W \\ i \neq j,\, j \neq k,\, i \neq k}} \chi_w\left(d\left(w_i, w_j\right) \ominus d\left(w_i, w_k\right)\right)$$

$$= \chi_{ij} d\left(w_i, w_j\right) \ominus \sum_{\substack{w_k \in W \\ i \neq j,\, j \neq k,\, i \neq k}} \chi_{ik} d\left(w_i, w_k\right)$$

$$= \chi_{ij} \left\langle T(w_i, w_j), I(w_i, w_j), F(w_i, w_j)\right\rangle$$

$$\ominus \left(\chi_{i1} \left\langle T(w_i, w_1), I(w_i, w_1), F(w_i, w_1)\right\rangle \oplus \ldots \oplus \chi_{in} \left\langle T(w_i, w_n), I(w_i, w_n), F(w_i, w_n)\right\rangle\right)$$

$$= \left\langle 1 - \left(1 - T(w_i, w_j)\right)^{\chi_{ij}}, I(w_i, w_j)^{\chi_{ij}}, F(w_i, w_j)^{\chi_{ij}}\right\rangle$$

$$\ominus \left(\begin{array}{l} \left\langle 1 - \left(1 - T(w_i, w_1)\right)^{\chi_{i1}}, I(w_i, w_1)^{\chi_{i1}}, F(w_i, w_1)^{\chi_{i1}}\right\rangle \oplus \ldots \\ \oplus \left\langle 1 - \left(1 - T(w_i, w_n)\right)^{\chi_{in}}, I(w_i, w_n)^{\chi_{in}}, F(w_i, w_n)^{\chi_{in}}\right\rangle \end{array}\right)$$

$$= \left\langle 1 - \left(1 - T(w_i, w_j)\right)^{\chi_{ij}}, I(w_i, w_j)^{\chi_{ij}}, F(w_i, w_j)^{\chi_{ij}}\right\rangle$$

$$\ominus \left\langle \sum_{\substack{k=1 \\ k \neq i,j}}^{n} \tilde{T}_{ik} - \prod_{\substack{k=1 \\ k \neq i,j}}^{n} \tilde{T}_{ik}, \prod_{\substack{k=1 \\ k \neq i,j}}^{n} \tilde{I}_{ik}, \prod_{\substack{k=1 \\ k \neq i,j}}^{n} \tilde{F}_{ik}\right\rangle$$

$$= \left\langle T_1, I_1, F_1\right\rangle \ominus \left\langle T_2, I_2, F_2\right\rangle$$

$$= \left\langle 1 - \left|T_1 - (T_2 - 1)^2\right|, 1 - \left|I_1 - I_2^2\right|, 1 - \left|F_1 - F_2^2\right|\right\rangle$$

where $\tilde{T}_{ik} = 1 - \left(1 - T(w_i, w_k)\right)^{\chi_{ik}}$, $\tilde{I}_{ik} = I(w_i, w_k)^{\chi_{ik}}$, $\tilde{F}_{ik} = F(w_i, w_k)^{\chi_{ik}}$.

Definition 13. *The relative weighted neutrosophic valued distance (from a random neutrosophic data* w_i *) to a neutrosophic data* w_j *is defined as follows:*

$$RD_\chi\left(w_j\right) = \sum_{w_i \in W} \chi_i RD_\chi\left(w_i \| w_j\right)$$

$$= \sum_{w_i \in W} \chi_i \left[\sum_{w_k \in W} \chi_w\left(d\left(w_i, w_j\right) \ominus d\left(w_i, w_k\right)\right)\right]$$

$$= \sum_{w_i \in W} \chi_i \left[\sum_{w_k \in W} \chi_w\left(\delta\left(d_{ij}, d_{ik}\right)\right)\right]$$

Definition 14. *The relative weighted neutrosophic valued distance from a neutrosophic data set* W_1 *to another neutrosophic data set* W_2 *is defined as follows:*

$$RD_\chi\left(W_1 \| W_2\right) = \sum_{x \in W_1} \chi_x \sum_{y \in W_2} \chi_y RD_\chi\left(x \| y\right)$$

Definition 15. *(Weighted cohesion measure between two neutrosophic data) The difference of the relative weighted neutrosophic-valued distance to* w_j *and the relative weighted neutrosophic-valued distance from* w_i *to* w_j *, i.e.,*

$$\rho_\chi(w_i, w_j) = RD_\chi(w_j) \ominus RD_\chi\left(w_i \| w_j\right) \tag{4}$$

is called the weighted neutrosophic-valued cohesion measure between two neutrosophic data w_i *and* w_j *. If* $\rho_\chi(w_i, w_j) \geq 0_W \left(\text{resp. } \rho_\chi(w_i, w_j) \leq 0_W\right)$ *then* w_i *and* w_j *are said to be cohesive (resp. incohesive).*

So, the relative weighted neutrosophic distance from w_i *and* w_j *is not larger than the relative weighted neutrosophic distance (from a random neutrosophic data) to* w_j.

Definition 16. *(Weighted cohesion measure between two neutrosophic data sets) Let* w_i *and* w_j *are elements of the neutrosophic data sets U and V, respectively. Then the measure*

$$\rho_\chi(U,V) = \sum_{w_i \in U} \chi_u \sum_{w_j \in V} \chi_v \rho_\chi(w_i, w_j) \tag{5}$$

is called weighted cohesion neutrosophic-valued measure of the neutrosophic data sets U and V.

Definition 17. *(Cluster) The non-empty neutrosophic data set W is called a cluster if it is cohesive, i.e.,* $\rho(W,W) \geq 0_W$.

5. Clustering via Neutrosophic Valued *G*-metric Spaces

In this section, we can cluster neutrosophic big data thank to defined weighted distance definitions in Section 4 and *G*-metric definition.

Definition 18. *The neutrosophic valued weighted G-distance from a neutrosophic data* W *to a neutrosophic big data set* U *is defined as follows:*

$$G(w, y, z) = \sum_{y \in U} \chi_u \sum_{z \in U} \chi_u \left(d(w,y) \oplus d(w,z) \ominus d(y,z) \right) \tag{6}$$

Algorithm (K-sets algorithm)

Input: A neutrosophic big data set $W = \{w_1, w_2, ..., w_n\}$, a neutrosophic distance measure $d(.,.)$, and the number of sets *K*.

Output: A partition of neutrosophic sets $\{U_1, U_2, ..., U_K\}$.

1. Initially, choose arbitrarily *K* disjoint nonempty sets $U_1, U_2, ..., U_K$ as a partition of W.
2. for *i* from 1 to n do

   begin

   Compute $G(x_i, y_k, z_k)$ for each set U_k.

   Find the set to which the point x_i is closest in terms of *G*-distance.

   Assign point x_i to that set.

   end
3. Repeat from 2 until there is no further change.

---

6. Application and Example

We will give an example of the definition of the data that could have this kind of data and fall into the frame to fit this definition. We can call a data set a big data set if it is difficult and/or voluminous to define, analyze and visualize a data set. We give a big neutrosophic data example in accordance with this definition and possible use of *G*-metric, but it is fictional since there is no real neutrosophic big data example yet. It is a candidate for a good example that one of the current topics, image processing for big data analysis. Imagine a camera on a circuit board that is able to distinguish colors, cluster all the tools it can capture in the image and record that data. The camera that can be used for any color (for example white color vehicle) assigns the following degrees:

(I) The vehicle is at a certain distance at which the color can be detected, and the truth value of the portion of the vehicle is determined.

(II) The rate at which the vehicle can be detected by the camera is assigned as the uncertainty value (the mixed color is the external factors such as the effect of daylight and the color is determined on a different scale).

(III) The rate of not seeing a large part of the vehicle or the rate of out of range of the color is assigned as the value of falsity.

Thus, data of the camera is clustering via *G*-metric. This result gives that the numbers according to the daily quantities and colors of vehicles passing by are determined. The data will change continuously as long as the road is open, and the camera records the data. There will be a neutrosophic data for each vehicle. So, a Big Neutrosophic Data Clustering will occur.

Here, the weight functions we have defined for the metric can be given 1 value for the main colors (red-yellow-blue). For other secondary or mixed colors, the color may be given a proportional value depending on which color is closer.

A Numerical Toy Example

Take 5 neutrosophic data with their weights are equal to 1 to make a numerical example:

$$W = \left\{ w_1 \langle 0.6, 0.6, 0.6 \rangle, w_2 \langle 0.8, 0.4, 0.5 \rangle, w_3 \langle 0.5, 0.8, 0.7 \rangle, w_4 \langle 0.9, 0.5, 0.6 \rangle, w_5 \langle 0.1, 0.2, 0.7 \rangle \right\}$$

$K = 3$ disjoint sets can be chosen $U_1 = \{ w_1, w_4, w_5 \}, U_2 = \{ w_2, w_3 \}$.

Then

$$d(w_i, w_j) = \begin{bmatrix} \langle 0,1,1 \rangle & \langle 0.2, 0.8, 0.9 \rangle & \langle 0.1, 0.8, 0.9 \rangle & \langle 0.3, 0.9, 1.0 \rangle & \langle 0.5, 0.6, 0.9 \rangle \\ \langle 0.2, 0.8, 0.9 \rangle & \langle 0,1,1 \rangle & \langle 0.3, 0.6, 0.8 \rangle & \langle 0.1, 0.9, 0.9 \rangle & \langle 0.7, 0.8, 0.8 \rangle \\ \langle 0.1, 0.8, 0.9 \rangle & \langle 0.3, 0.6, 0.8 \rangle & \langle 0,1,1 \rangle & \langle 0.4, 0.7, 0.9 \rangle & \langle 0.4, 0.4, 1.0 \rangle \\ \langle 0.3, 0.9, 1.0 \rangle & \langle 0.1, 0.9, 0.9 \rangle & \langle 0.4, 0.7, 0.9 \rangle & \langle 0,1,1 \rangle & \langle 0.2, 0.8, 0.9 \rangle \\ \langle 0.5, 0.6, 0.9 \rangle & \langle 0.7, 0.8, 0.8 \rangle & \langle 0.4, 0.4, 1.0 \rangle & \langle 0.2, 0.8, 0.9 \rangle & \langle 0,1,1 \rangle \end{bmatrix}$$

where we assume the $d(w_i, w_j)$ as in Example 3. So, we can compute the *G*-metrics of the data as in Equation (3):

$$G(w_1, U_1) = G(w_1, w_4, w_5) = \langle 0.99, 0.90, 0.91 \rangle$$

$$G(w_1, U_2) = G(w_1, w_2, w_3) = \langle 0.79, 0.72, 0.83 \rangle$$

$$G(w_2, U_1) = G(w_2, w_1, w_4) \oplus G(w_2, w_1, w_5) \oplus G(w_2, w_4, w_5) = \langle 0.9874, 0.6027, 0.6707 \rangle$$

$$G(w_2, U_2) = G(w_2, w_2, w_3) = \langle 0,1,1 \rangle$$

$$G(w_3, U_1) = G(w_3, w_1, w_4) \oplus G(w_3, w_1, w_5) \oplus G(w_3, w_4, w_5) = \langle 1, 0.4608, 0.6707 \rangle$$

$$G(w_3, U_2) = G(w_3, w_2, w_3) = \langle 0,1,1 \rangle$$

$$G(w_4, U_1) = G(w_4, w_1, w_5) = \langle 0.81, 0.64, 0.91 \rangle$$

$$G(w_4, U_2) = G(w_4, w_2, w_3) = \langle 0.97, 0.73, 0.83 \rangle$$

So, according to the calculations above, w_4 belongs to set U_1 and the other data belong to U_2. Here, we have made the data belonging to the clusters according to the fact that the truth values of the *G*-metrics are mainly low. If the truth value of *G*-distance is low, then the data is closer to the set.

7. Conclusions

This paper has introduced many new notions and definitions for clustering neutrosophic big data and geometric similarity metric of the data. Neutrosophic data sets have density. For example, sets having indeterminacy density or neutrosophic density and these are adding the more data and complexity. So, neutrosophic data sets are complex big data sets. Separation and clustering of these sets are evaluated according to weighted distances. Neutrosophic data sets in the last part of the paper, K-sets algorithm has been given for neutrosophic big data sets. We hope that the results in this paper can be applied to other data types like interval neutrosophic big data sets and can be analyzed in other metric spaces such as neutrosophic complex valued *G*-metric spaces etc. and can help to solve problems in other study areas.

References

1. Zadeh, L.A. Fuzzy sets. *Inf. Control* **1965**, *8*, 338–353.
2. Atanassov, K.T. Intuitionistic fuzzy sets. *Fuzzy Sets Syst.* **1986**, *20*, 87–96.
3. Smarandache, F. *Neutrosophy: Neutrosophic Probability, Set and Logic*; American Research Press: Rehoboth, NM, USA, 1998.
4. Deli, I.; Subas, Y. Single valued neutrosophic numbers and their applications to multicriteria decision making problem. *Neutrosophic Sets Syst.* **2014**, *2*, 1–13.
5. Fan, C.; Fan, E.; Ye, J. The Cosine Measure of Single-Valued Neutrosophic Multisets for Multiple Attribute Decision-Making. *Symmetry* **2018**, *10*, 154.
6. Li, Y.; Liu, P.; Chen, Y. Some single valued neutrosophic number heronian mean operators and their application in multiple attribute group decision making. *Informatica* **2016**, *27*, 85–110.
7. Lu, Z.; Ye, J. Single-valued neutrosophic hybrid arithmetic and geometric aggregation operators and their decision-making method. *Information* **2017**, *8*, 84.
8. Ye, J.; Smarandache, F. Similarity measure of refined single-valued neutrosophic sets and its multicriteria decision making method. *Neutrosophic Sets Syst.* **2016**, *12*, 41–44.
9. Guo, Y.; Şengür, A.; Ye, J. A novel image thresholding algorithm based on neutrosophic similarity score. *Measurement* **2014**, *58*, 175–186.
10. Xia, P.; Zhang, L.; Li, F. Learning similarity with cosine similarity ensemble. *Inf. Sci.* **2015**, *307*, 39–52.
11. Laney, D. 3D data management: Controlling data volume, velocity and variety. Meta group: Stamford, CT, USA, 2001.
12. Chang, C.; Liao, W.; Chen, Y.; Liou, L. A Mathematical Theory for Clustering in Metric Spaces. *IEEE Trans. Netw. Sci. Eng.* **2016**, *3*, 2–16.
13. Azam, A.; Fisher, B.; Khan, M. Common fixed point theorems in complex valued metric spaces. *Numer. Funct. Anal. Optim.* **2011**, *32*, 243–253.
14. Ahmed El-Sayed Ahmed; Saleh Omran; Abdalla J. Asad, Fixed Point Theorems in Quaternion-Valued Metric Spaces. *Abstr. Appl. Anal.* **2014**, *2014*, 9.
15. Mustafa, Z.; Sims, B. A new approach to generalized metric spaces. *J. Nonlinear Convex Anal.* **2006**, *7*, 289–297.
16. Abbas, M.; Nazir, T.; Radenović, S. Some periodic point results in generalized metric spaces. *Appl. Math. Comput.* **2010**, *217*, 4094–4099.
17. Khamsi, M.A. Generalized metric spaces: A survey. *J. Fixed Point Theory Appl.* **2015**, *17*, 455–475.
18. Smarandache, F. Neutrosophic set—A generalization of the intuitionistic fuzzy sets. *Int. J. Pure Appl. Math.* **2005**, *24*, 287–297.
19. Vasantha, W.B.; Smarandache, F. Some Neutrosophic Algebraic Structures and Neutrosophic N-Algebraic Structures; Hexis: Phoenix, AZ, USA, 2006.
20. Şahin, M.; Kargın, A. Neutrosophic triplet normed space. *Open Phys.* **2017**, *15*, 697–704.
21. Stanujkic, D.; Smarandache, F.; Zavadskas, E.K.; Karabasevic, D., *An Approach to Measuring the Website Quality Based on Neutrosophic Sets*. Infinite Study: Brussel, Belgium, 2018.