Neutrosophic Sets and Systems

Volume 53 Article 19

6-15-2023

A Neutrosophic based C-Means Approach for Improving Breast Cancer Clustering Performance

Ahmed Abdelhafeez

Hoda K. Mohamed

Ali Maher

Ahmed Abdelmonem

NSS
Neutrosophic Sets and Systems, Vol. 53, 2023
University of New Mexico

A Neutrosophic based C-Means Approach for Improving Breast Cancer Clustering Performance

Ahmed Abdelhafeez[1,*], Hoda K Mohamed [2], Ali Maher [3], and Ahmed Abdelmonem[4]

[1] Faculty of Information Systems and Computer Science, October 6th University, Cairo, 12585, Egypt 1; aahafeez.scis@o6u.edu.eg

[2] Faculty of Engineering, Ain shams University, Cairo, 11566, Egypt 3; Hoda.korashy@eng.asu.edu.eg

[3] Military Technical College, Cairo, 18711, Egypt 2; ali_mtc@hotmail.com

[4] Faculty of computers and informatics Zagazig University, Zagazig, 44511, Egypt 4; ahmed.abdelmon3m15@gmail.com

*Correspondence: aahafeez.scis@o6u.edu.eg

Abstract: Breast cancer is among the most prevalent cancers, and early detection is crucial to successful treatment. One of the most crucial phases of breast cancer treatment is a correct diagnosis. Numerous studies exist about breast cancer classification in the literature. However, analyzing the cancer dataset in the context of clusterability for unsupervised modeling is rare. This work analyzes pointedly the breast cancer dataset clusterability via applying the widely used c-means clustering algorithm and its evolved versions fuzzy and neutrosophic ones. An in-depth comparative study is conducted utilizing a set of quantitative and qualitative clustering efficiency metrics. The study's outcomes divulge the presented neutrosophic c-means clustering superiority in segregating similar breast cancer instances into clusters.

Keywords: Breast cancer dataset clusterability; Fuzzy c-means clustering; Neutrosophic c-means clustering; t-SNE; Silhouette coefficient.

1. Introduction

One of the biggest problems in the healthcare system is cancer-related death. It ranks among the major causes of death among women [1]. More people have died from breast cancer than from any other disease, including tuberculosis and malaria.

Initial analysis of this condition can reduce the rate of mortality, which is on the rise [2]. Breast cancer is the sixth foremost reason of mortality globally, according to the Globocan 2020 data, and it is diagnosed in one out of every four women worldwide [3].

Making a precise diagnosis of malignancies is crucial. Most breast tumors are caused by benign (non-cancerous) alterations, however, if a benign tumor is assumed as a malignant one, it might have disastrous consequences. The most crucial actions to lowering breast cancer mortality are early detection and receiving state-of-the-art cancer therapy. Early-stage, mild breast cancer that hasn't spread can be treated successfully and quickly. Routine screening tests represent the most

dependable method for identifying breast cancer in its earliest stages [4]. In an extraordinarily rich information environment, healthcare has extraordinarily little knowledge. Healthcare systems contain a vast amount of data, and it is crucial to find and establish connections with hidden data. The International Classification of Diseases (ICD) divided the foremost origins of death into five categories, with breast cancer being part of two of them [5]. According to a McKinsey report, the amount of data is increasing at a pace of 50% annually. Data science has now formally emerged as an especially important field. According to research, the phrase "data science" describes a systematic examination of the structure, properties, and evaluation of information along with the role that data play in society [6]. Statistics knowledge can be exploited from a diversity of areas, even though machine learning procedures are the most frequently used healthcare datasets.

A data analysis method called machine learning teaches a computer what results from various methods. The most popular machine learning algorithms are decision trees, k-means clustering, and neural networks [7].

The incidence of breast cancer among women, particularly those between the ages of 35 and 55, is rising because of the inhabitants of industrialized and developing nations changing their lifestyles from traditional to modern. By identifying breast tumors in their initial stages, it is possible to keep track of the prevalence of the illness [8]. Breast cancer screening methods include self- and professional breast exams, Magnetic Resonance Imaging (MRI), ultrasound, and mammography [9]. The mammogram, which includes the backdrop, the breast region, adipose tissue, breast masses, and microcalcifications with high intensities, is the result of the mammography procedure [10]. Radiologists may make mistakes or overlook crucial signs as the need for mammography processing increases because of weariness [11].
In [12]. The DCE-MRI enables a highly accurate follow-up for breast tumors. Fuzzy spatial clustering was used by Militello et al. To segment masses on DCE-MRI breast scans, and the results were superior to those of other traditional methods.

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset, a highly well-known cancer dataset, was used as the basis for another cluster analysis work [13]. which incorporated a multidimensional data analysis. Because the multidimensionality of data has long been a barrier to data analysis this study hypothesized that a multidimensional data set must be projected into a lower dimensional space where it will inevitably lose some of its features to be displayed due to the limits of handling more than three spatial dimensions.

In [14], a new training dataset of breast cancer is produced using the modified k-means technique, which enhances the performance of the support vector machine model. A prediction model for breast cancer was developed using k-means and support vector machine. Using the updated k-means, a training dataset of the highest caliber was produced. Then, to group the cancerous instances of unidentified photographs, classification and accuracy were improved.

In [15], The R programming language, R visual studio, and Weka machine learning software have all been tried on the breast cancer dataset. Using various clustering algorithms were employed to

examine the proper correlation in the Breast cancer dataset. In this unsupervised learning strategy, a pretrained model or label is not necessary.

The key contribution of this proposed methodology is as follows:

• Through the application of the widely known c-means clustering technique and its advanced versions fuzzy and neutrosophic ones, this work specifically investigates the clusterability of the breast cancer dataset.

• Using a collection of quantitative and qualitative clustering efficiency metrics, extensive comparative research is carried out. In terms of silhouette score, precision, and rand index, the suggested neutrosophic c-means clustering gets the best clustering performance.

Following are the last five portions of this study: Section 2 gives a review of materials and methods, Section 3 presents the metrics and results, and Section 4 presents the overall research conclusion.

2. Materials and methods

2.1 Dataset

The efficiency of the suggested model was evaluated using the WDBC datasets, which are breast cancer datasets [16]. Data from the University of Wisconsin Hospitals have previously been gathered. Each example was assigned a benign or malignant classification. The WDBC has 569 occurrences (about 62.7% benign and 37.3% malignant) and 32 significant patient features. A patient ID, 30 tumor-specific traits, and one class indicator are among these characteristics. The distinguishing features of the tumors of the patients were gathered using ten different elements, including texture, radius, area, perimeter, smoothness, concavity, compactness, concave spots, fractal dimension, and symmetry. These traits were generated from a breast lumps fine needle aspirate (FNA) picture. A set of 30 features was created by deciding the key, recognizing data for each image, such as mean, standard error, and the least or biggest standards of these features.

The dataset from Kaggle that included information about breast cancer. Thirty-two parameters make up the dataset. All the indicators can be used to categorize cancer, and if they have significantly high values, that could indicate the presence of malignant tissue. A number called ID serves as the first argument and is used for identification. The second factor is the diagnosis of membranes, which can be either malignant or benign depending on the tissue. The correct tissue diagnosis must be established for various cancer kinds if both membranes require various therapies. Following these two, a range between the center and a point on the perimeter is shown by estimated means, standard errors, and radius means. The estimated standard error is shown by radius se. The center of the projected range has the highest value of the radius worst. Knowing the distance between the center and the point is crucial since the size affects operation. With large tumors, surgery is not an option. The gray-scale values' standard deviation is represented by the texture mean. The estimated standard deviation of gray-scale values is represented by the texture se. Gray-scale values with the largest mean standard deviation are characterized as having the worst texture. Grayscale is frequently used to locate tumors, and the standard deviation is crucial to identifying data variation and explaining how to disperse the values. While the standard error of the mean indicates the core

tumor expressed as perimeter se, the perimeter mean represents the mean value for the core tumor. The perimeter worst column displays the core tumor's maximum value. Area means, area se, and area worst point are identical to the previously mentioned mean of the cancer cell areas. Regional variations in the radius range are represented by smoothness mean, local variations in radius length are represented by smoothness se, and the biggest mean value is displayed as smoothness worst. The greatest mean value of the calculation is referred to as compactness worst. Compactness mean is a mean value of estimation of the perimeter and area. Compactness se is used to calculate the standard error of the mean. The severity of the concave regions of the shape is shown by the concavity mean, and the number of concave points in the shape is indicated by the concave points mean. Concavity se denotes the standard deviation of concave areas, whereas concave points se denotes the standard deviation of the shape's concave areas. The worst concavity and worst concave points represent the highest mean value. The fractal dimension means the calculated mean value for the coastline approximation, the fractal dimension se represents the standard error of the coastline approximation, and the fractal dimension worst represents the highest mean value [16].

Figure 1. Shows the distribution of thirty-one features for all 569 lesions using the Weka tool. Through the malignant and benign lesions, each feature was visualized to show how much affect the detection of diagnosis.

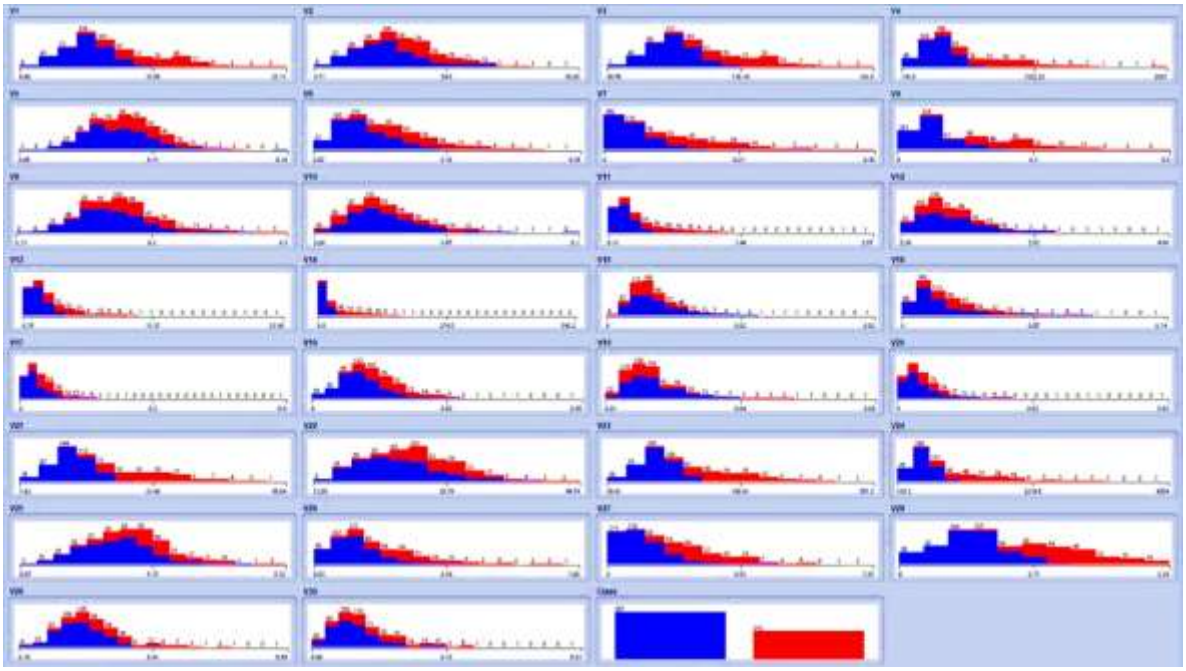


Figure 1 Distribution of dataset features.

The data is available for download in.csv format. Then, the CSV extension was updated to the Weka-compatible Attribute Relation Data Format (ARFF) extension. The data was then subjected to extensive preprocessing. There are 569 instances in the collection. The dataset is then further normalized using the min-max normalization approach in Weka software so that all feature values fall within the range [0, 1]. Being an unsupervised learning technique, clustering solely uses feature values. This indicates that the dataset's final column, the category label, is not normalized. We first

eliminate the ID number. The Hopkins Statistic Index is then used to analyze the dataset to determine whether there is a strong propensity for clustering among the data points. Then, using Python programming language and Weka software tools, we apply several clustering techniques. Hopkins Statistic Index = 0.6809 shows the dataset is heavily clustered, according to our results.

2.2 K-means clustering

K-means is a clustering method that can group enormous volumes of data with a processing time that is both quicker and more efficient. The k-means algorithm, in contrast, has a flaw that is dependent on the initial value cluster that establishes the center. K-means clustering provides superior topical remedies as trial outcomes. However, the testing procedure calls for the data to be close together. In order to get a high degree of similarity among the cluster points, this can be divided into a number of clusters. The k-means algorithm is also multisided, according to (celebi et al. 2013) K-means are too straightforward to modify at each stage of the process because they are predicated on the conditions for iteration termination. They are also easy to measure in terms of distance. The first data point collection from each cluster's midway is crucial since the k-mean cluster is a local optimization [17]. The objectives of these adjustments are the best precision and the fastest convergence. If the initial point is selected from the cluster's midway, the k-mean cluster algorithm will also be limited to the optimum site. Additionally, a starting point for the k-mean clustering method will be chosen at random from the middle, up to style k. The initial centroid cluster, which is chosen at random, will have an impact on the total number of centroid cluster iterations. Therefore, by locating the centroid cluster in the high starting data points, it can be fixed to achieve higher execution.

Two familiar features of the K-means clustering technique. The first is that as a precondition parameter for clustering, it requires the usage of a specified cluster starting value, or "k centroid." However, in most cases, without prior knowledge, we are unable to determine the optimal initial number of clustering that a given data set can produce. Connecting each point to the closest cluster is the other feature.

2.3 The Fuzzy C-Means Clustering (FCM)

In their work, Dunn and Bezdek devised the fuzzy c-means method (FCM). Finding the optimal participation and clustering center to minimize the optimal solution is the main notion.

To set up the membership vector, the method must first decide on the number of clusters to create. After then, both the Center of Clustering and the Membership vector are regularly revised. Centers of various clusters and levels of membership may be produced when the optimal solution is smaller than some threshold.

These are some of the algorithm's drawbacks: Having a high degree of, sensitivity to, and depending on, the initial grouping. It is simple for the algorithm to become wedged in a local least if the starting cluster center is distant from the global optimum clustering center.

2.4 The Neutrosophic Sets

Smarandache introduced the neutrosophic concept, a generalization of previously expanded concepts, to overcome the shortcomings of conventional fuzzy clustering and enhance its capacity to manage and communicate unclear knowledge. When applied to fuzzy clustering, the neutrosophic theory is able not just to portray non-deterministic difficulties more accurately, but as well as provide solutions to those problems that remain open.

The central tenet of neutrosophic thought is the premise that every vantage point has some element of veracity, doubt, and fallacy. For this reason, the concepts of and were proposed as neutrosophic elements to signify the seriousness, ambiguity, and humorlessness of occurrences. True, indeterminate, and false outcomes are the names given to these agnostic components.

2.5 The Neutrosophic C-Means Clustering (NCM)

Conventional fuzzy clustering approaches in clustering algorithms can only explain the degree to which each group exists. It is challenging to distinguish which category a given sample belongs to and which divisions it joins, especially for the samples located in the border area among distinct groups. The neutrosophic c-means clustering method was introduced by Guo et al. To address these issues, which is an improvement on the FCM based on neutrosophic theory (NCM).

We propose a fresh special combination, A, which unites the determinant and indeterminate clusters. Let $A = C_j \cup B \cup R, j = 1, 2, \ldots, c, where\ C_j$ Is an indeterminate cluster, B refers to clusters near the edges, R relates to erratically sampled data, and is the union process. Clusters B and R both fall within the category of being agnostic. T indicates membership in the determinant cluster, I in the perimeter cluster, and F in the noisy set of data. With uncertainty in clustering in mind, we construct a new goal function and class membership as follows:

$$\begin{aligned} J(T,I,F,C) = \langle & \sum_{i=1}^{n}\sum_{k=1}^{c}(w_1 T_{ik})^m \parallel x_i - v_k \parallel^2 \\ & + \sum_{i=1}^{n}\sum_{k=1}^{\binom{c}{2}}(w_2 I_{2ik})^m \parallel x_i - \overline{v_{2k}} \parallel^2 \\ & + \sum_{i=1}^{n}\sum_{k=1}^{\binom{c}{3}}(w_3 I_{3ik})^m \parallel x_i - \overline{v_{3k}} \parallel^2 \\ & + \sum_{i=1}^{n}\sum_{k=1}^{\binom{c}{4}}(w_4 I_{4ik})^m \parallel x_i - \overline{v_{4k}} \parallel^2 \rangle \\ & + \sum_{i=1}^{n}\sum_{k=1}^{\binom{c}{5}}(w_5 I_{5ik})^m \parallel x_i - \overline{v_{5k}} \parallel^2 \\ & + \sum_{i=1}^{n}\sum_{k=1}^{\binom{c}{c}}(w_c I_{cik})^m \parallel x_i - \overline{v_{ck}} \parallel^2 \\ & + \sum_{i=1}^{n}(\overline{w_{c+1}} F_i)^m \end{aligned} \tag{1}$$

2.6 Hopkins statistic

The Hopkins statistic (Lawson and Jurs 1990) calculates the likelihood that a particular data set was produced by a uniform data distribution to evaluate the tendency of a data set to cluster [18]. In other words, it evaluates the data's spatial randomness.

Use the Hopkins score from clustered to estimate the likelihood of cluster formation before doing clustering. The outcome was two clusters, indicating the data is eligible for clustering. Unsupervised data has no notion of how many supposed clusters there are, therefore, assumptions range from two to six. Figure 2. Shows the Silhouette values vs. the number of clusters.

However, after clustering, the silhouette score used to measure cluster quality varied for each cluster. The formula is defined as follows:

$$\mathbf{H} = \frac{\sum_{i=1}^{n} \boldsymbol{y_i}}{\sum_{i=1}^{n} \boldsymbol{x_i} + \sum_{i=1}^{n} \boldsymbol{y_i}} \tag{2}$$

How should I interpret the Hopkins data?

In the case of a uniform distribution of D, $\sum_{i=1}^{n} \boldsymbol{y_i}$ and $\sum_{i=1}^{n} \boldsymbol{x_i}$, would be close to one another, and H would therefore be about 0.5. However, if clusters are present in D, the distances between manufactured points ($\sum_{i=1}^{n} \boldsymbol{y_i}$) would be expected to be much greater than those between genuine points ($\sum_{i=1}^{n} \boldsymbol{x_i}$), increasing the value of H.

Noting from figure 2. Through cluster numbers from two to six, we pick up the highest silhouette coefficient, which is determined by the number of two clusters, which suggests that two clusters are the optimum choice for data clustering. The average Silhouette Score plot of the number of clusters fluctuates between two and six and the highest silhouette value is 0.58, demonstrating that the breast cancer dataset is well matched to the given cluster when the cluster size is two.

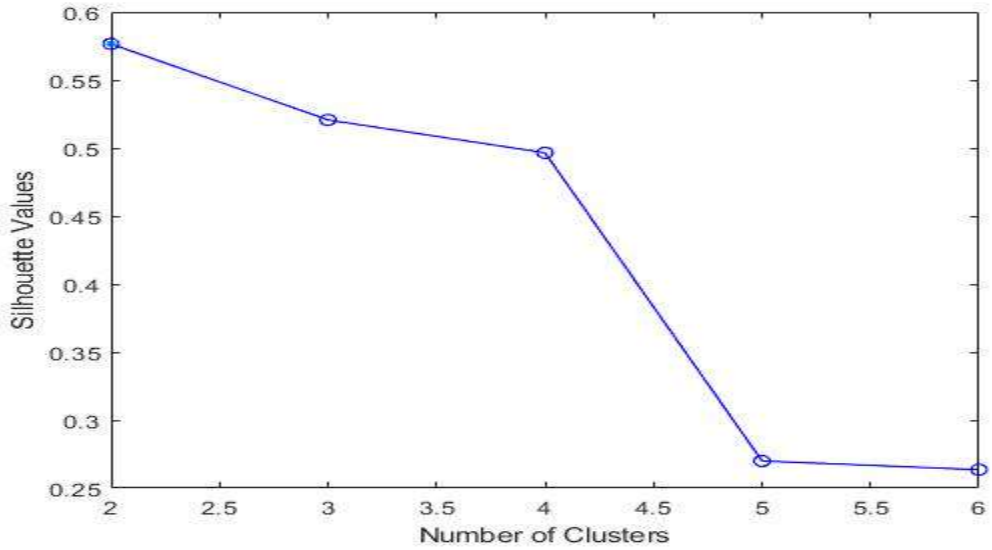
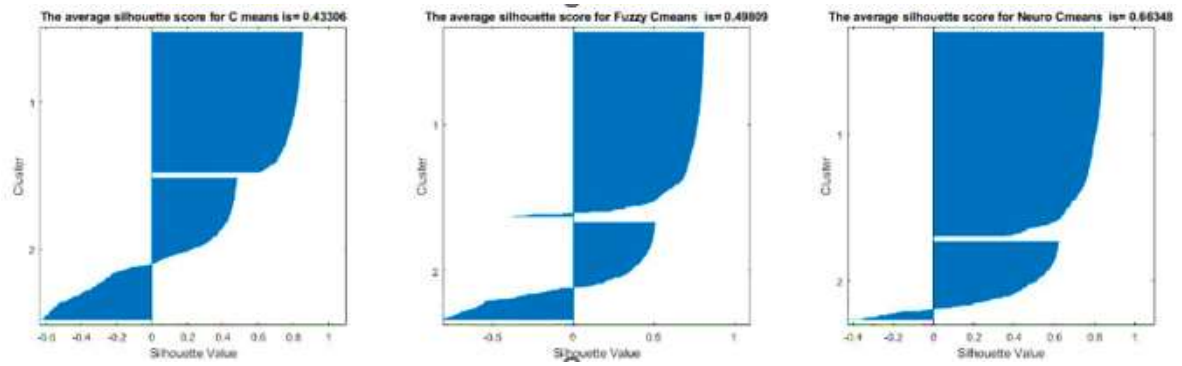


Figure 2. Silhouette values vs. the number of clusters.

2.7 Silhouette Score Analysis

Researchers may determine how closely related each observation is to the cluster to which it has been assigned about other clusters using silhouette analysis. For each observation in the data, this metric (silhouette width) runs from -1 to 1, and it can be interpreted as follows [19]:
I) Values that are near 1 indicate that the allocated cluster is a good fit for the observation.
II) Values near 0 point to a possible borderline match between two groups of the observation.
III) Values near -1 point to the possibility that the observations were placed in the incorrect cluster.
In the study, we will use three well known clustering methods, investigating which one will be superior in detecting cancer cases for the aforementioned dataset. Applying k-means, fuzzy c-means, and neutrosophic c-means clustering methods.

In figure 3, We investigated two clusters of the provided data: a benign cluster and a malignant cluster. Clusters C1 and C2 are home to all 569 instances. The two clusters' average Silhouette values are 0.43 for the c-means cluster on the left, 0.5 for the fuzzy cluster in the middle, and 0.66 for the neutrosophic cluster on the right. When the Silhouette width has the highest value, which is the neutrosophic c-means in the outcomes from the three approaches, we can obtain the best clustering result. The silhouette score is shown in Table 1.



(a) (b) (c)

Figure 3. Silhouette Score for (a) k-means (left) , (b) fuzzy c-means (middle) , and (c) neutrosophic c-means clustering methods (right).

Table 1. The silhouette score of the three models.

Model 1	Silhouette score
K-Means	0.43306
Fuzzy c-means	0.49809
Neutrosophic c-means	0.66348

2.8 T-Distributed Stochastic Neighbor Embedding

T-Distributed Stochastic Neighbor Embedding (t-SNE) has emerged as a powerful standard for visualizing high-dimensional datasets in a variety of biological data sets, especially for large datasets. Using this method will help each class have a clearer image. T-SNE encompasses a variety of fields, including Bioinformatics, music analysis, computer security, and cancer biology. Similar to SNE, t-SNE chooses two distinct similarity measures for the two-dimensional embedding and the high-dimensional information. The objective of this stage is to produce a 2-dimensional embedding with a KL divergence between the vector of similarities between points in pairs over the entire dataset and the similarities between points in the encoding that is as little as possible. T-SNE is used to solve the nonconvex optimization problem utilizing gradient descent and random initialization.

Figure 4. Shows the three dimensions of T-SNE visualization (best viewed in color) for the four clustering methods actual clusters (right-bottom), c-means (left-upper), fuzzy c-means (right-upper), and neutrosophic c-means (left-bottom), respectively. By visualizing, it becomes evident that neutrosophic c- means is the best option because it is close to the actual clustering. C-means, on the other hand, is the clustering approach that is farthest from the actual means; as a result, fuzzy c-means is the second-closest method.

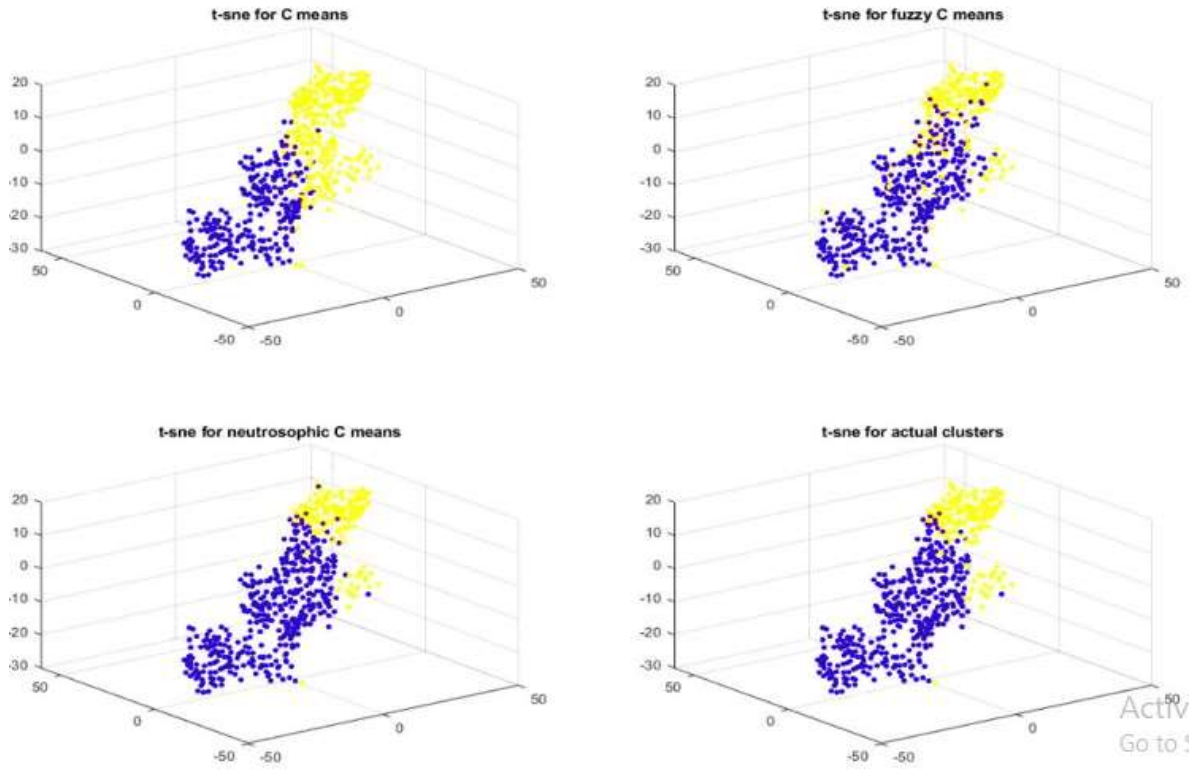


Figure 4. T-SNE graphs for c-means (left-upper), fuzzy c-means (right-upper), neutrosophic c-means (left-bottom), and actual clusters (right-bottom), (best viewed in color).

3. Results

3.1. Performance metrics*:*

Achieving high intra-cluster identity and low inter-cluster commonality is the primary focus of clustering methods (objects in the same cluster are more similar than the objects in different clusters).

In several of my investigations, the clustering methods failed to identify the optimal number of clusters. It has been shown that certain methods overestimate the size of clusters while others underestimate it. When the final class number matches the number of categories in the gold standard, we may use the typical criteria for analyzing recognition accuracy.

Equation. (3) depicts the clustering technique as a K x S matrix, where K is the expected number of clusters of the clustering method and S is the number of classes in the reference set.

Here, the element a_{ks} represents the entire number of objects that have been clustered into the k^{th} cluster and are of the sth class in the ideal distribution.

When K = S, the clustering method's estimated number of clusters exactly corresponds to the number of classes found in the reference data.

$$matrix\ k * s = \begin{matrix} k_1 \\ \dots \\ k_k \end{matrix} \begin{bmatrix} a_{11} & \cdots & a_{1s} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{ks} \end{bmatrix} \quad (3)$$

Precision

We find the benchmark class to which the most items have been allocated for each cluster. Following this, we take the sum of the largest number of items in every group and divide it by the whole number of grouped objects. precision is determined by calculating the resultant value by using K x S matrix, as seen in equation. (4).

$$precision = \frac{\sum_s \max_k a_{ks}}{\sum_k \sum_s \max_k a_{ks}} \quad (4)$$

Recall

We find the class where most items belong based on the gold standard. The complete list of grouped and unclustered items is then divided by the sum of the maximum number of objects in each gold standard class. Equation. (5) demonstrates the K x S matrix's role in deriving the recall (also called sensitivity). The number of items that are not in a cluster is denoted by U.

$$Recall = \frac{\sum_s \max_k a_{ks}}{\sum_k \sum_s \max_k a_{ks} + U} \quad (5)$$

F1-Score

According to equation. (6), the F1-score is determined by taking the mean of the accuracy and recall scores.

$$F1 = 2 * \frac{precision * recall}{precision + recall} \quad (6)$$

Rand Index

Two clustering strategies may be compared with one another using the Rand index.

The Rand Index, sometimes abbreviated as R, is determined using the following formula:

$$R\ =\ (a + b)\ /\ (nC2) \quad (7)$$

Where:

a: The frequency with which a given pair of items is assigned to the same cluster by two different techniques of clustering.

b: The frequency with which a given pair of items is found in different clusters when using two different clustering techniques.

nC2 is the count of all the non-matched pairings in a collection of n items.

3.2 results analysis

Applying the equations. (4,5,6) to compute the precision, recall, and f1 score. In the precision, the total clustered data is 569 and there is no unclustered data. In the c-means the maximum clustered data is 453, so the precision is computed by dividing the maximum clustered data by the total clustered data, the outcome is 0.796. Due to no unclustered data, the precision is equal to recall and f1 score. Table 2 shows all analyses of the precision and Rand Index.

Table 2. The overall performance analysis of the proposed model.

Model	Precision	Rand Index
C-Means	0.796	0.6748
FCM	0.8872	0.7919
NCM	0.9789	0.9586

Table 3. There are four predicted class data by the neutrosophic and fuzzy c-means clustering. In data 1, the fuzzy predicted class 0, neutrosophic predicted class 0, and the actual label is class 0, so the fuzzy and neutrosophic predicted this data truly. In data 2 the fuzzy predicted class 1, and the neutrosophic predicted class 1, also the actual data is class 1, so the fuzzy and neutrosophic predicted true labels. In data 3 the fuzzy predicted class 0, but the neutrosophic predicted class 1 and the actual labels are class 1, so the neutrosophic predicted true but the fuzzy predicted false. In data 4, the fuzzy predicted class 1, the neutrosophic predicted class 0, and the actual class is 0, so the neutrosophic predicted true and the fuzzy predicted false. Table 3. The neutrosophic predicted four true classes and the fuzzy predicted the two true classes and one false class.

Table 3. The predicted labels for Fuzzy and Neutrosophic vs. Actual label.

Data	Fuzzy Predicated Label	Neutrosophic Predicated Label	Actual Label
Data 1	0	0	0
Data 2	1	1	1
Data 3	0	1	1
Data 4	1	0	0

4. Conclusions

This paper analyzes the breast cancer dataset cluster ability via applying the widely used c-means clustering algorithm and its evolved versions fuzzy and neutrosophic ones. The conducted comparative study utilizes various metrics to fairly judge the breast cancer dataset clustering efficiency. The suggested neutrosophic c-means clustering achieves the highest clustering performance in terms of silhouette score, precision, and Rand index.

References

1. Kocarnik, J.M.; Compton, K.; Dean, F.E.; Fu, W.; Gaw, B.L.; Harvey, J.D.; Henrikson, H.J.; Lu, D.; Pennini, A.; Xu, R. Cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life years for 29 cancer groups from 2010 to 2019: A systematic analysis for the Global Burden of Disease Study 2019. *JAMA Oncol.* **2022**, *8*, 420–444.
2. Smolarz, B.; Nowak, A.Z.; Romanowicz, H. Breast Cancer-Epidemiology, Classification, Pathogenesis, and Treatment (Review of Literature). *Cancers* **2022**, *14*, 2569.
3. HOXHA, Ilir; ISLAMI, Dafina Ademi; UWIZEYE, Glorieuse; FORBES, Victoria; CHAMBERLIN, Mary D. Forty-five Years of Research and Progress in Breast Cancer: Progress for Some, Disparities for Most. JCO Global Oncology, v. 8, **2022**. Disponível em: < https://ascopubs.org/doi/full/10.1200/GO.21.00424 >. DOI: https://doi.org/10.1200/GO.21.00424.
4. PDQ Cancer Genetics Editorial Board. *Genetics of Breast and Gynecologic Cancers (PDQ®): Health Professional Version*; National Cancer Institute (US): Bethesda, MD, USA, 2020.
5. Nikdouz, A.; Namarvari, N.; Shayan, R.G.; Hosseini, A. Comprehensive Comparison of Theragnostic Nanoparticles in Breast Cancer. Am. J. Clin. Exp. Immunol. **2022**, 11, 1–27.
6. Cleveland Clinic. Available online: https://my.clevelandclinic.org/health/diseases/3986-breast-cancer (accessed on 19 12 **2022**).
7. Rahman, M.F.; Wen, Y.; Xu, H.; Tseng, T.-L.; Akundi, S. Data mining in telemedicine. In *Advances in Telemedicine for Health Monitoring: Technologies, Design, and Applications*; IET Digital Library: London, UK, **2020**; pp. 103–131.
8. Tiggaa, N.P.; Garg, S. Prediction of Type 2 Diabetes using Machine Learning Classification Methods. In Proceedings of the International Conference on Computational Intelligence and Data Science (ICCIDS 2019), Gurgaon, India, 6–7 September **2019**.
9. American Cancer Society. American Cancer Society Recommendations for the Early Detection of Breast Cancer. **2022**. Available online: https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/american-cancer-society-recommendations-for-the-early-detection-of-breast-cancer.html (accessed on 19 December 2022).
10. Roee Zamir, Shai Bagon, David Samocha, Yael Yagil, Ronen Basri, Miri Sklair-Levy, and Meirav Galun "Segmenting microcalcifications in mammograms and its applications", Proc. SPIE 11596, Medical Imaging **2021**: Image Processing, 115962W (15 February 2021); **https://doi.org/10.1117/12.2580398**
11. Degnan, A.J.; Ghobadi, E.H.; Hardy, P.; Krupinski, E.; Scali, E.P.; Stratchko, L.; Ulano, A.; Walker, E.; Wasnik, A.P.; Auffermann, W.F. Perceptual and interpretive error in diagnostic radiology—Causes and potential solutions. *Acad. Radiol.* **2019**, *26*, 833–845.
12. C. Militello, L. Rundo, M. Dimarco, A. Orlando, V. Conti, R. Woitek, I. D'Angelo, T.V. Bartolotta, G. R usso" Semi-automated and interactive segmentation of contrast-enhancing masses on breast DCE-MRI using spatial fuzzy clustering" *Biomedical Signal Processing and Control*, *71*,**2022**, Article 103113.
13. Pantazi, S., Kagolovsky, Y., Moehr, J.R.: Cluster analysis of Wisconsin breast cancer dataset using self-organizing maps. In: Surjàn, G., Engelbrecht, R., Mcnair, P. (eds.) Health Data in the Information

Society, no. 90 in Technology and Informatics. International Congress on Medical Informatics, pp. 431–436. IOS Press, Amsterdam **2002.**

14. W. L. Al-Yaseen, A. Jehad, Q. A. Abed, and A. K. Idrees, "The Use of Modified K-Means Algorithm to Enhance the Performance of Support Vector Machine in Classifying Breast Cancer," Int. J. Intell. Eng. Syst., vol. 14, no. 2, p. 190, **2021**, doi: 10.22266/ijies2021.0430.17.
15. Chakraborty, S., Murali, B.: Investigate the correlation of breast cancer dataset using different clustering techniques. ArXiv abs/2109.01538,**2021**.
16. Mangasarian OL, Wolberg WH. Cancer diagnosis via linear programming. SIAM News 1990;23(5): 1-18. Available: http://www.cs.wisc.edu/~olvi/uwmp/cancer.html ,**2022**, Dec 15, 2022].
17. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A k-means clustering algorithm. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **1979**, *28*, 100–108.
18. Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J.: Understanding of internal clustering validation measures. In: **2010** IEEE international conference on data mining. pp. 911–916. IEEE.
19. Hodes L. **1992**. Limits of classification. 2. Comment on Lawson and Jurs. J. *Chem. Inf. Model.* ***32***(2): 157–166.

Received: Sep 1, 2022. Accepted: Dec 15, 2022