Journal of Innovative Optical Health Sciences
Vol. 7, No. 1 (2014) 1450018 (20 pages)

DOI: 10.1142/S1793545814500187

World Scientific
www.worldscientific.com

FUZZY C-MEANS IN FINDING SUBTYPES OF CANCERS IN CANCER DATABASE

S. R. KANNAN
Department of Mathematics
Pondicherry Central University, India
srkannan.mat@pondiuni.edu.in

S. RAMTHILAGAM
Department of Mathematics
Periyar Government College, Tamil Nadu

R. DEVI
Department of Mathematics
Pondicherry Central University, India

T. P. HONG
Department of Computer Science and Information Engineering
National University of Kaohsiung, Taiwan, China

Received 8 July 2013
Accepted 27 November 2013
Published 21 January 2014

Finding subtypes of cancer in breast cancer database is an extremely difficult task because of heavy noise by measurement error. Most of the recent clustering techniques for breast cancer database to achieve cancerous and noncancerous often weigh down the interpretability of the structure. Hence, this paper tries to find effective Fuzzy C-Means-based clustering techniques to identify the proper subtypes of cancer in breast cancer database. This paper obtains the objective function of effective Fuzzy C-Means clustering techniques by incorporating the kernel induced distance function, Renyi's entropy function, weighted distance measure and neighborhood terms-based spatial context. The effectiveness of the proposed methods are proved through the experimental works on Lung cancer database, IRIS dataset, Wine dataset, Checkerboard dataset, Time Series dataset and Yeast dataset. Finally, the proposed methods are implemented successfully to cluster the breast cancer database into cancerous and noncancerous. The clustering accuracy has been validated through error matrix and silhouette method.

Keywords: Fuzzy C-Means; kernel induced distance; entropy terms; cancer database.

1. Introduction

Recently clustering techniques are widely used in analyzing subtypes of cancer in cancer medical database to help physicians for treatment plan. This paper deals with the breast cancer database that contains two different types of cancers using effective fuzzy C-means techniques. Breast cancer is the second leading cause of cancer death in the world[1,2] and it is the leading cause of cancer deaths among women aged 40–59.[3] Breast cancer is the most frequently diagnosed cancer in women in the United States. About 800,000 new cases of breast cancer is diagnosed each year among Taiwanese women, and breast cancer is the second most common form in Taiwan according to cancer report.[2] In India, cancer is one of the leading causes of deaths and due to cancer disease around 3 lakh deaths occur annually in India.[4] Hence, proper[5] diagnosing method for analyzing the breast cancer database is very important to reduce the cancer death. There are number of techniques involved to identify the subtypes of breast cancer patterns.[6–8] Recently unsupervised clustering method of Fuzzy C-means is the most widely used method in clustering cancer medical databases.[9] Unsupervised clustering divides the dataset into several clusters based on the similarity between the data objects, it does not require any prior information about the data objects for clustering them into available structures.[10–12] Due to the uncertain nature of many practical real world problems, fuzzy set theory[13] based Fuzzy clustering techniques[10,14] have been proposed by researchers. The most widely used conventional fuzzy C-means is incapable in clustering nonlinear structured medical database[15,16] due to its Euclidean norm to measure the similarity between the data points. Further it fails to incorporate any information about spatial context, and neighborhood information to cluster the medical database into meaningful subtypes of cancers. Hence, researchers have invented modified fuzzy C-means algorithms[17–19] in order to deal the data objects with different noises for facing the complicated structure of databases, but the methods did not give expected accuracy in clustering the database into available subtypes.[20] To overcome the drawbacks, this paper formulates suitable novel fuzzy C-means with effective cluster center initialization in clustering more complicated structure of medical database. The algorithms of this paper are obtained by incorporating the kernel induced distance function, Renyi's entropy, weighted distance measure and neighborhood terms-based spatial context. The kernel induced distance of the proposed objective function converts the lower dimension of the objects into higher dimension to have meaningful distance between objects. The Renyi's entropy, weighted distance measure and neighborhood terms-based spatial context of objective function of this paper tries to reduce the uncertainty in the objects, and balancing the loss of information in the objects in medical database. The proposed methods work well with the dataset which is affected by the noises such as measurement error, faulty equipment and data transmission error in dataset. The methods elegantly find the difference between cluster centers and data points by considering the information from neighboring objects with higher dimensional distance using kernel, and it finds desirable membership for an object to an appropriate cluster. The neighborhood information-based clustering algorithms are effectively applied to medical and other real life fields by researchers.[21,22] Renyi's entropy with proposed objective function of fuzzy C-means tries to quantify the diversity and uncertainty of the boundary of the subtypes of cancerous portions. The entropy with proposed method helps to measure the disorder in the amount of information that may be gained by proposed systems in clustering the medical database. There is extensive literature on the applications of the Renyi's entropy in many fields from biology, medicine and genetics.[23,24] Further Renyi's entropy-based proposed algorithm performs well on datasets of nonspherical shape and capable of clustering a high-dimensional dataset. However, the random selection of initial prototypes of fuzzy C-means-based algorithms lead more number of iterations to reach the termination criterion,[25,26] therefore in order to avoid irrelevant initial random prototypes this paper introduces a prototype initialization method. This paper proved the effectiveness and strength of the proposed algorithms in clustering more complicated medical database through the experimental results on benchmarks and real database. The rest of this paper is organized as follows. Section 2 gives a short form of Kernel-based Fuzzy C-means clustering, validation method and methodology. In Sec. 3, this paper presents proposed algorithms $KEFCM_{wd}$ and $KFCM_{nt}$ online. Section 4 presents Prototypes or centers knowledge Method. Section 5 reports the

experimental results on Yeast, Lung cancer, IRIS, Wine, Checkerboard, Time Series and Breast cancer datasets. Section 6 gives the conclusion of this paper.

2. Related Works and Methodology

2.1. *Kernel-based Fuzzy C-means Clustering*

In the fuzzy C-means algorithm,[10] a cluster is viewed as a fuzzy set in the dataset, X. Thus, each data element in the dataset will have membership values with all clusters. The degree of membership, to which a data point belongs to a cluster, is computed from the distances of the data point to the cluster prototypes. The membership degrees basically reflect the relative inverse squared distance of the data point to the different cluster prototypes. In FCM, the proximity of each data, x_i, to the cluster prototypes, v_k, is defined as the membership u_{ik} of x_i to the kth cluster of X minimizing the following objective function:

$$J(U,V) = \sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik}^{m}\|x_i - v_k\|^2, \tag{1}$$

where $X = \{x_i\}_{i=1}^{n} \subset R^N$ is a given set of unlabeled data; $V = \{v_k\}_{k=1}^{c} \subset R^N$ are the centroids of the clusters, and $m = [1,\infty]$ is the weighting exponent which determines the fuzziness of the resultant clusters, at $m = 1$ Fuzzy C-Means fall down to Hard C-Means, fuzzy partition matrix $U = [u_{ik}]$ is generated that is of size $c \times n$ (c — number of clusters and n — number of data elements), satisfying the constraint $\sum_{k=1}^{c} u_{ik} = 1, i = 1, 2, 3, \ldots, n$. Minimization of the objective function is achieved by iteratively optimizing for $\mathbf{U}$ and $\mathbf{V}$. The cluster centers and the memberships are computed as follows:

$$v_k = \frac{\sum_{i=1}^{n} u_{ik}^{m} x_i}{\sum_{i=1}^{n} u_{ik}^{m}} k = 1, 2, 3, \ldots, c, \tag{2}$$

$$u_{ik} = \left(\sum_{j=1}^{c}\left(\frac{\|x_i - v_k\|}{\|x_i - v_j\|}\right)^{\frac{2}{m-1}}\right)^{-1},$$
$$k = 1, 2, 3, \ldots, c; \quad i = 1, 2, 3, \ldots, n. \tag{3}$$

The above FCM with distance measure has drawbacks in clustering complex and large amount of dataset. Therefore, kernel method is used to formulate the kernel versions of the Fuzzy C-means (KFCM) algorithm[27–30] to cluster complex data structure. The data point $x_i \in R^m, i = 1, 2, \ldots, n$, is transformed from the original space to a feature space H by a nonlinear mapping ϕ, it becomes the following form $\phi(x_1), \phi(x_2), \phi(x_3), \ldots, \phi(x_n)$. So the inner product in the original space can be expressed by the Mercer kernel[31] as $K(x_i, x_j) = (\phi(x_i) \cdot \phi(x_j))$. The Euclidean distance in the feature space can be represented as follows:

$$\begin{aligned} &d_H(x_i, x_j) \\ &\quad = \sqrt{\|\phi(x_i) - \phi(x_j)\|} \\ &\quad = \sqrt{\phi(x_i)\cdot\phi(x_i) - 2\phi(x_i)\cdot\phi(x_j) + \phi(x_j)\cdot\phi(x_j)}. \end{aligned} \tag{4}$$

The objective function of the KFCM algorithm can be formulated as follows:

$$J(U,V) = 2\sum_{i=1}^{n}\sum_{k=1}^{c}(u_{ik})^m(1 - K(x_i, v_k)). \tag{5}$$

2.2. *Clustering validation*

The silhouette width[32,33] is used to validate the results of proposed methods in clustering the complex dataset. The silhouette average value of clusters can vary between -1 and 1.[32–34] These silhouette average values measure the degree of confidence in the clustering assignment of a particular observation, with well-clustered observations having values near 1 and poorly clustered observations having values near -1. The silhouette width $s(i)$ of the object i is obtained using the equation $s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$. In the above equation $a(i)$ is the average distance between the ith data and all other data in the cluster. $b(i)$ is the smallest average distance between the i-data and all other data of other clusters. The silhouette average width of the clusters has been obtained by taking the average of the silhouette value of points belonging to the cluster.

2.3. *Methodology*

The methodology of this paper aims to find novel fuzzy clustering techniques in order to reduce the uncertainty in the objects by loss of information, measurement error and data transmitter error. This paper invents kernel fuzzy C-means-based entropy with weighted distance measure [KEFCM_{wd}] and kernel fuzzy C-means-based entropy term with neighborhood term [KFCM_{nt}]. Further this paper

invents the center or prototype initialization method to speed up the clustering algorithms. In order to evaluate the performance of the proposed methods yeast dataset, lung cancer dataset, IRIS dataset, wine dataset, time series dataset, checkerboard dataset and breast cancer dataset are involved in the experimental works. To identify the strength of the clustering methods, the clustering accuracy is obtained by silhouette method and error matrix.

3. Proposed KEFCM$_{\text{wd}}$ and KFCM$_{\text{nt}}$ Algorithms

3.1. *Proposed KEFCM$_{\text{wd}}$ algorithm*

The modified objective function from the standard fuzzy C-means[10] is given by

$$J(U, V) = \sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik}^{m}\|\psi(x_i) - \psi(v_k)\|^2, \quad (6)$$

where ψ stands as map $x- > \psi(x) \in F$, $x \in X$. The common ground of kernel-based FCM is to map the input data element into a feature space with higher dimension via a nonlinear transformation and then perform FCM in that feature space. And the distance function can be expressed using inner product space as $\|\psi(x_i) - \psi(v_k)\|^2 = \langle\psi(x_i), \psi(x_i)\rangle + \langle\psi(v_k), \psi(v_k)\rangle - 2\langle\psi(x_i), \psi(v_k)\rangle$, where $i = 1, 2, \ldots, n$, and $k = 1, 2, \ldots, c$. We adopt hyper tangent Function to evaluate the distance, i.e., $\psi(x_i, v_k)$ expressed as Hyper Tangent function

$$\psi(x_i, v_k) = 1 - \tanh\left(\frac{-\|x_i - v_k\|^2}{w_k}\right)$$

where w_k is the weighted mean distance in cluster k, and is given by

$$w_k = \left\{\frac{\sum_{i=1}^{n} u_{ik}\|x_i - v_k\|^2}{\sum_{i=1}^{n} u_{ik}}\right\}^{\frac{1}{2}}. \quad (7)$$

Using the expression (7) we obtained $\psi(x_i, x_i) = 1$ and $\psi(v_k, v_k) = 1$, so the distance function can be rewritten as

$$\|\psi(x_i) - \psi(v_k)\|^2 = 2(1 - \psi(x_i, v_k)). \quad (8)$$

From Eqs. (6) and (8), we have the kernelized fuzzy C-means given by

$$J(U, V) = 2\sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik}^2 \cdot (1 - \psi(x_i, v_k)). \quad (9)$$

In order to cluster effectively the more complicated dataset which have been corrupted by the noises such as measurement error, faulty equipment and data transmission error, the Renyi's entropy fuzzy C-means-based hyper tangent kernel algorithm [KEFCM$_{\text{wd}}$] is introduced as

$$J(U, V) = 2\sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik}^2 \cdot (1 - \psi(x_i, v_k)) + \frac{1}{|1 - z|}\sum_{i=1}^{n}\sum_{k=1}^{c} \ln u_{ik}^{z}. \quad (10)$$

Here z is the resolution parameter. The KEFCM$_{\text{wd}}$ objective function is optimized to obtain effective membership grades to the objects which are close to their prototypes. Using the Lagrange multiplier to the objective function of KEFCM$_{\text{wd}}$, the equation for obtaining prototypes and membership grades are calculated. In order to derive the KEFCM$_{\text{wd}}$ with respect to membership, the objective function of KEFCM$_{\text{wd}}$ has been modified as

$$J(U, V, \lambda) = 2\sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik}^2 \cdot (1 - \psi(x_i, v_k)) + \frac{1}{|1 - z|}\sum_{i=1}^{n}\sum_{k=1}^{c} \ln u_{ik}^{z} - \sum_{i=1}^{n}\lambda_i\left(\sum_{k=1}^{c} u_{ik} - 1\right), \quad (11)$$

where $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)$.

Optimizing Eq. (11) by using $\frac{\partial J(U,V,\lambda)}{\partial u_{ik}} = 0$,

$$u_{ik} = \frac{\lambda_i}{4(1 - \psi(x_i, v_k)) + \frac{z}{|1-z|}}. \quad (12)$$

Using the constraint $\sum_{k=1}^{c} u_{ik} = 1$, we get

$$\lambda_i = \frac{1}{\sum_{j=1}^{c} 4(1 - \psi(x_i, v_j)) + \frac{z}{|1-z|}}. \quad (13)$$

substitute (13) in (12), we get

$$u_{ik} = \frac{4(1 - \psi(x_i, v_k)) + \frac{z}{|1-z|}}{\sum_{j=1}^{c} 4(1 - \psi(x_i, v_j)) + \frac{z}{|1-z|}}. \quad (14)$$

The general equation is used to attain membership grades for data elements for getting meaningful groups. The accuracy of clustering results mainly depends on the cluster centers. Now optimizing

the following objective function of KEFCM_{wd}, this paper obtains the equations for updating the prototypes.

$$J(U,V,\lambda) = 2\sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik}^2 \tanh\left(\frac{-\|x_i - v_k\|^2}{w_k}\right) + \frac{1}{|1-z|}\sum_{i=1}^{n}\sum_{k=1}^{c} \ln u_{ik}^z - \sum_{i=1}^{n}\lambda_i\left(\sum_{k=1}^{c} u_{ik} - 1\right). \quad (15)$$

Since $\psi(x_i, v_k) = 1 - \tanh(\frac{-\|x_i - v_k\|^2}{w_k})$, optimizing the above objective function with respect to v_k using the necessary condition of Lagrangian method $\frac{\partial J(U,V,\lambda)}{\partial v_k} = 0$, we have

$$\frac{\partial J(U,V,\lambda)}{\partial v_k} = \sum_{i=1}^{n} u_{ik}^2\left(1 - \tanh^2\left(\frac{-\|x_i - v_k\|^2}{w_k}\right)\right) \times \left(\frac{x_i - v_k}{w_k}\right). \quad (16)$$

Simplifying this Eq. (16), we get

$$v_k = \frac{\sum_{i=1}^{n} \frac{u_{ik}^2}{w_k}\psi(x_i, v_k)\left(1 + \tanh\left(\frac{-\|x_i - v_k\|^2}{w_k}\right)\right)x_i}{\sum_{i=1}^{n} \frac{u_{ik}^2}{w_k}\psi(x_i, v_k)\left(1 + \tanh\left(\frac{-\|x_i - v_k\|^2}{w_k}\right)\right)}. \quad (17)$$

3.2. *Proposed $KFCM_{nt}$ algorithm*

In order to avoid assigning same memberships for an object which has similar intensity values for more than one cluster and to assign appropriate membership to the objects which have been corrupted by heavy noise, this paper incorporates neighborhood term with fuzzy C-means given by

$$J(U,V,\lambda) = 2\sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik}^2(1 - \psi(x_i, v_k)) + \sum_{i=1}^{n}\sum_{k=1}^{c}\frac{u_{ik}^2}{1+\varphi}(1 - \psi(\tilde{g}_i, A_k)). \quad (18)$$

Here $\tilde{g}_i$ denotes the geometric mean of the neighboring elements of x_i. The parameter φ is the neighborhood regularizer term that controls the effect of neighborhood term. A_k represents the average value of the elements in the kth cluster. In essence, the addition of the second term in (18), equivalently, aims at deriving effective finding memberships for objects. By an optimization way, the objective function can be minimized under the constraint $\sum_{k=1}^{c} u_{ik} = 1$. The proposed method can be expressed using Lagrangian multiplier as

$$J(U,V,\lambda) = 2\sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik}^2(1 - \psi(x_i, v_k)) + \sum_{i=1}^{n}\sum_{k=1}^{c}\frac{u_{ik}^2}{1+\varphi}(1 - \psi(\tilde{g}_i, A_k)) - \sum_{i=1}^{n}\lambda_i\left(\sum_{k=1}^{c} u_{ik} - 1\right). \quad (19)$$

Optimizing the Eq. (19) in terms of u_{ik} using the necessary condition of Lagrangian method $\frac{\partial J(U,V,\lambda)}{\partial u_{ik}} = 0$,

$$u_{ik} = \frac{\lambda_i}{2\left[2(1 - \psi(x_i, v_k)) + \frac{1}{1+\phi}(1 - \psi(\tilde{g}_i, A_k))\right]}. \quad (20)$$

Using the constraint $\sum_{k=1}^{c} u_{ik} = 1$ we get

$$\frac{\lambda_i}{2} = \frac{1}{\sum_{j=1}^{c}\left[2(1 - \psi(x_i, v_j)) + \frac{1}{1+\phi}(1 - \psi(\tilde{g}_i, A_j))\right]}. \quad (21)$$

Substitute (21) in (20), we have the following equation for updating memberships.

$$u_{ik} = \frac{\left(2(1 - \psi(x_i, v_k)) + \frac{1}{1+\phi}(1 - \psi(\tilde{g}_i, A_k))\right)^{-1}}{\sum_{j=1}^{c}\left(2(1 - \psi(x_i, v_j)) + \frac{1}{1+\phi}(1 - \psi(\tilde{g}_i, A_j))\right)^{-1}}. \quad (22)$$

The incorporation of neighboring information has strengthened the above membership function which gives the accurate result. To obtain the updating prototype equation, the objective function in (20) can be written as

$$J(U,V,\lambda) = 2\sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik}^2 \tanh\left(\frac{-\|x_i - v_k\|^2}{w_k}\right) + \sum_{i=1}^{n}\sum_{k=1}^{c}\frac{u_{ik}^2}{1+\varphi}\tanh\left(\frac{-\|\tilde{g}_i - A_k\|^2}{w_k}\right) - \sum_{i=1}^{n}\lambda_i\left(\sum_{k=1}^{c} u_{ik} - 1\right).$$

Optimizing the above objective function with respect to v_k using the necessary condition of Lagrangian method $\frac{\partial J(U,V,\lambda)}{\partial v_k} = 0$, we have

$$\frac{\partial J(U,V,\lambda)}{\partial v_k} = \sum_{i=1}^{n} u_{ik}^2 \left(1 - \tanh^2\left(\frac{-\|x_i - v_k\|^2}{w_k}\right)\right) \times \left(\frac{x_i - v_k}{w_k}\right). \quad (23)$$

Simplifying this Eq. (23), we obtain the cluster center updating equation as follows,

$$v_k = \frac{\sum_{i=1}^{n} \frac{u_{ik}^2}{w_k} \psi(x_i, v_k)\left(1 + \tanh\left(\frac{-\|x_i - v_k\|^2}{w_k}\right)\right) x_i}{\sum_{i=1}^{n} \frac{u_{ik}^2}{w_k} \psi(x_i, v_k)\left(1 + \tanh\left(\frac{-\|x_i - v_k\|^2}{w_k}\right)\right)}. \quad (24)$$

This effective updating center equation is used to find proper structure of the clusters from the dataset and it induces the robustness to reduce computational complexity.

The above algorithms can uniformly be summarized in the following steps.

Algorithm 1

Step 1. Fix the number of prototypes or clusters and then select initial prototypes.

Step 2. Use Eqs. (14) and (22) for obtaining membership partition matrix.

Step 3. Update the centers using Eqs. (17) and (24).
Repeat Steps 2 and 3 until the following termination criterion is satisfied:
$\|V^{(k+1)} - V^{(k)}\| < \varepsilon$, where $V^{(k+1)}$ and $V^{(k)}$ are the vector of cluster centroids at $(k+1)$th and (k)th iterations.

4. Prototypes or Centers Knowledge Method

The random selection of prototypes can lead the clustering process with more number of iterations, because the random selection is sometimes completely irrelevant and far away to the cluster. This farther representation of prototypes usually takes more iteration to complete the algorithm and many times it causes errors in clustering results. This section shows a mathematical computation to choose the prototypes from the information learned in the given data.

Let X be a finite set in the N-dimensional space $\Re^N$. $X = \{x_1, x_2, \ldots, x_n\}$, $x_k \in \Re^N$, $k = 1, 2, \ldots, n$. Consider each point in X is with dimensionality N.

Step 1. Compute the expected value μ of each point of N-dimensional in given dataset. Find the variance σ^2 using means expected value μ.

Step 2. Get $C = \frac{\sigma^2}{k}$, where k is the number of cluster and C is the number of objects in every cluster.

Step 3. Find *Median* (C).

Step 4. Assign the median of each cluster as cluster center.

Algorithm 2

Step 1. Fix the number of prototypes or centers of clusters and then select initial prototypes, using Prototypes knowledge Method.

Step 2. Use step 2 and 3 of algorithm 1.
Repeat Steps 2 and 3 until the following termination criterion is satisfied:
$\|V^{(k+1)} - V^{(k)}\| < \varepsilon$, where $V^{(k+1)}$ and $V^{(k)}$ are the vector of cluster centroids at $(k+1)$th and (k)th iterations.

5. Experimental Results

To evaluate the performance of proposed methods, this section implements the proposed methods with Yeast Dataset, Lung Cancer Database, IRIS dataset, Wine dataset, Time series dataset and Checkerboard dataset. Finally, the proposed methods have been successfully implemented with the breast cancer database for dividing it into two subtypes of cancers. The algorithms were implemented in programming language [R] on Workstation (HP Z800 INTEL Xeon HEX (6) Dual Core Processor).

5.1. *Experimental results on yeast dataset*

Yeast Dataset[35] is composed of 1484 data, each data has 9 attributes (8 predictive, 1 name). The eight predictive are: (1) McGeoch's method for signal sequence recognition; (2) von Heijne's method for signal sequence recognition; (3) score of the ALOM membrane spanning region prediction program; (4) Score of discriminant analysis of the amino acid content of the N-terminal region of mitochondrial and nonmitochondrial proteins; (5) Presence of

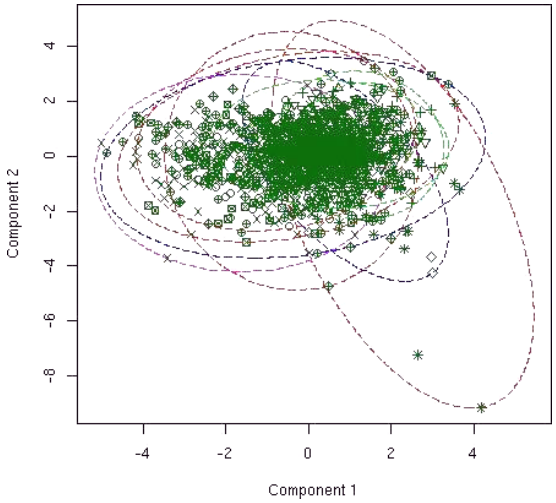

Fig. 1. Size of Clusters by KFCM.

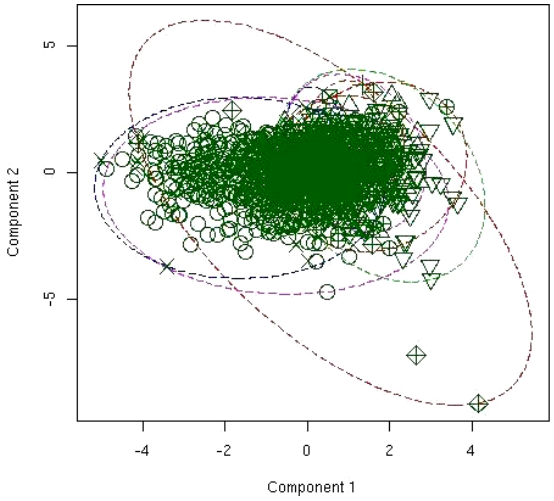

Fig. 3. Size of Clusters by $KEFCM_{wd}$.

"HDEL" substring; (6) Peroxisomal targeting signal in the C-terminus; (7) Score of discriminate analysis of the amino acid content of vacuolar and extracellular proteins; and (8) Score of discriminant analysis of nuclear localization signals of nuclear and non-nuclear proteins. The data has 10 classes: cytosolic or cytoskeletal, nuclear, mitochondrial, membrane protein, no N-terminal signal, membrane protein-uncleaved, membrane protein-cleaved signal, extracellular, vacuolar, peroxisomal, endoplasmic reticulum lumen. The research in clustering Yeast Dataset is extensively active in recent years[35–37] to improve the clustering accuracy. KFCM[29] clustering results based on 10 classes in yeast data are plotted in Fig. 1. The divided 10 classes by KFCM are visualized in Fig. 2. The results of proposed $KEFCM_{wd}$, and Proposed $KFCM_{nt}$

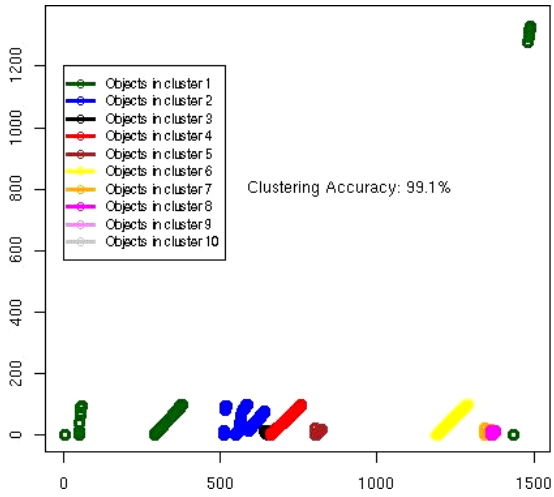

Fig. 4. Reallocated 1484 Data by $KEFCM_{wd}$.

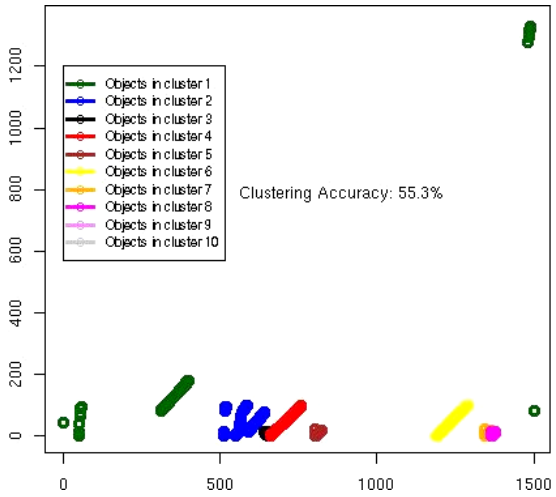

Fig. 2. Reallocated 1484 Data by KFCM.

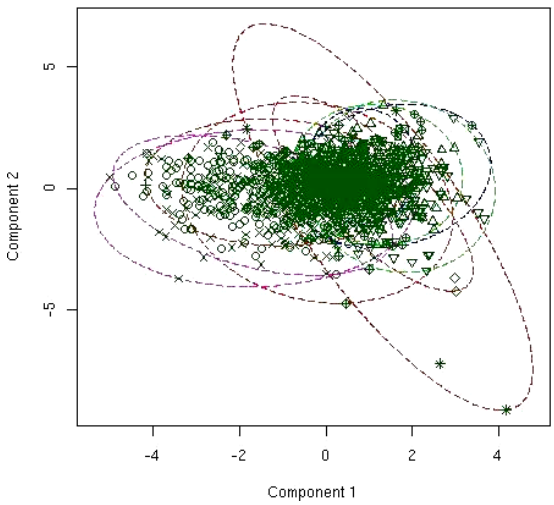

Fig. 5. Size of Clusters by $KFCM_{nt}$.

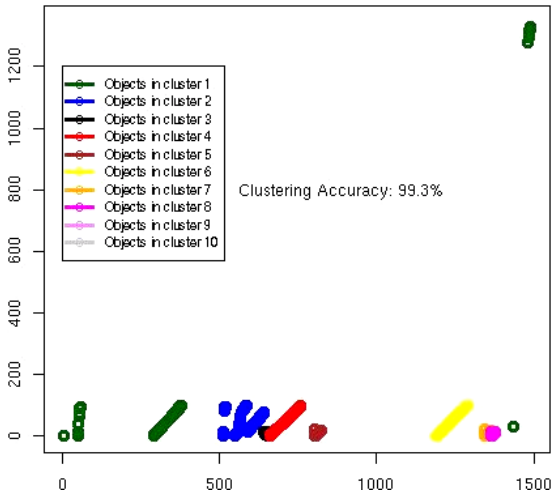

Fig. 6. Reallocated 1484 Data by $KFCM_{nt}$.

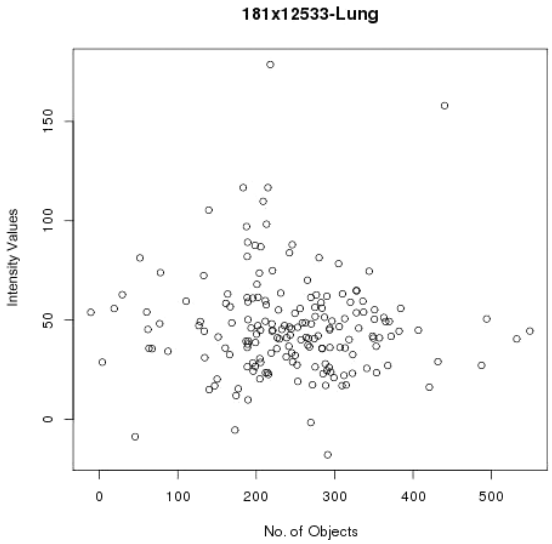

Fig. 7. Lung database.

methods based on 10 classes in yeast dataset are shown in Figs. 3–6. The objects in 10 classes of yeast dataset by Proposed $KEFCM_{wd}$ and $KFCM_{nt}$ algorithms are given in Figs. 3 and 5, respectively. The captured size of 10 classes of yeast dataset by proposed $KEFCM_{wd}$ and proposed $KFCM_{nt}$ are shown in Figs. 4 and 6, respectively. The clustering accuracy of KFCM, proposed $KEFCM_{wd}$ and $KFCM_{nt}$ algorithms on clustering 10 classes in yeast database are listed in Table 1. This paper shows from Table 1, that the proposed methods improve the clustering accuracy more than the KFCM, because of the neighborhood terms, and weighted distance with Renyi's entropy.

The Error Matrix Table 2 gives the accuracy between reference classes and the obtained classes in yeast dataset by the methods involved in this experiment study. From Table 2, the best clustering accuracy was obtained for proposed methods during the experiment on yeast dataset with 10 clusters.

Table 1. Silhouette average values in clustering Yeast dataset.

	KFCM	$KEFCM_{wd}$	$KFCM_{nt}$
Accuracy	55.3%	99.1%	99.3%

Table 2. Error matrix on Yeast dataset.

	KFCM	$KEFCM_{wd}$	$KFCM_{nt}$
Accuracy	51%	98.7%	99.1%

5.2. *Experimental results on lung cancer*

This subsection has employed a version of the dataset in Ref. 38 which has samples from two cancer types [malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA)]. The Lung Dataset which is given in Fig. 7 consists 181 of human tissue samples and each sample are described by 12,533 genes.

To show the effectiveness of the proposed methods, $KEFCM_{wd}$ and $KFCM_{nt}$ in clustering Lung cancer database, this subsection compares the results of proposed methods with the results obtained by GKFCM[28] and KFCM on same dataset. The partitions with three clusters for the two types of cancers for the algorithms of GKFCM, KFCM, $KEFCM_{wd}$ and $KFCM_{nt}$ are illustrated in Figs. 8–11, respectively. The GKFCM and KFCM take more iteration to complete the process of algorithm in clustering two subtypes of cancers. Further the existing methods provide poor accuracy in clustering the Lung cancer Database. On the other hand, the proposed algorithms predict the two subtypes of cancers correctly due to its robust objective functions.

The results on Table 3 show that the classes of GKFCM and KFCM algorithms exhibited poor clustering performance than that of the other classes obtained by proposed methods $KEFCM_{wd}$ and $KFCM_{nt}$. The accuracy test (average accuracy

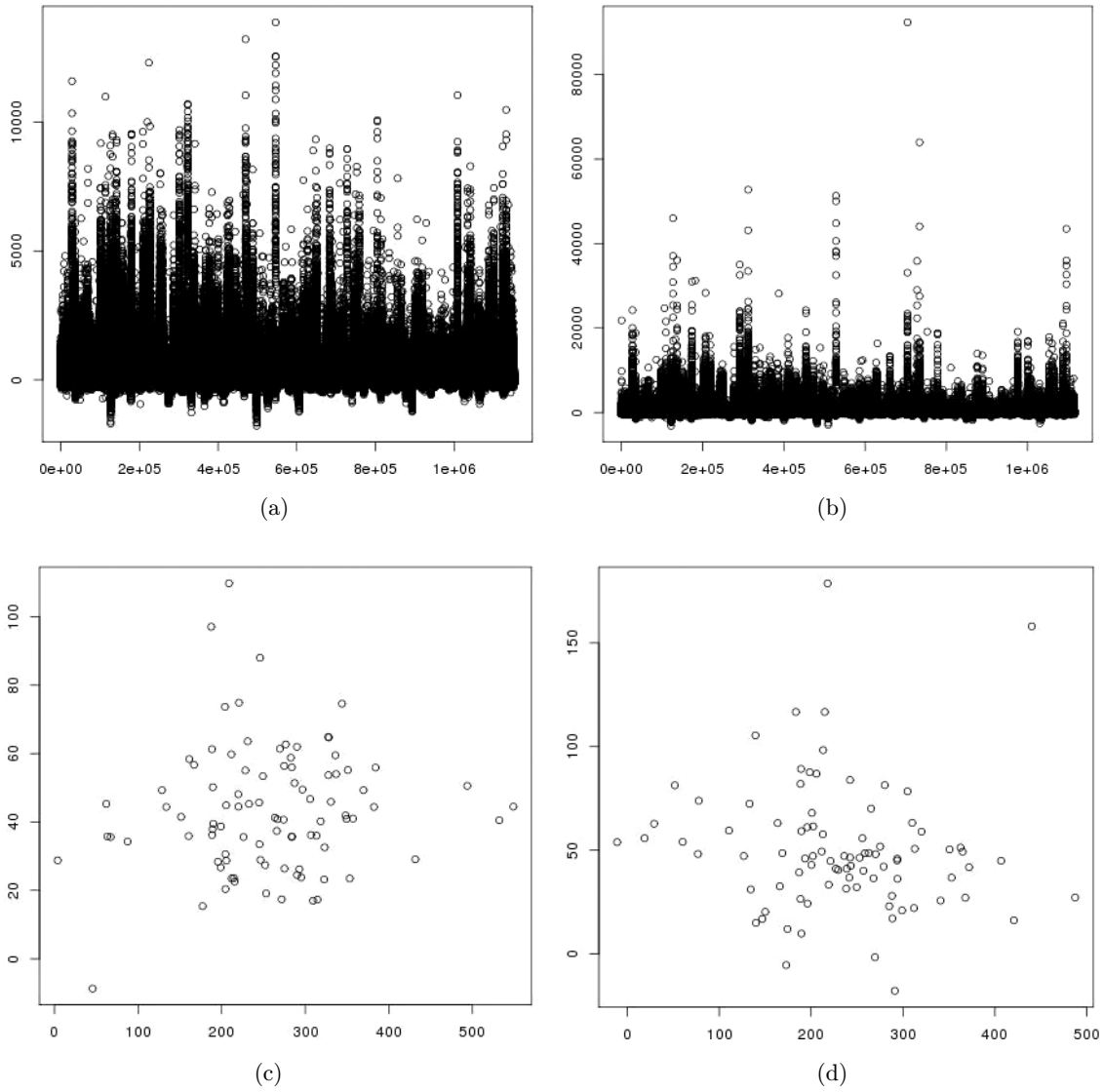

Fig. 8. Subtype of Cancers by GKFCM: (a) Gene expression in Cluster 1 (b) Gene expression in Cluster 2 (c) Cluster 1 in 181 Lung Cancer Dataset (d) Cluster 2 in 181 Lung Cancer Dataset.

value) indicated that GKFCM and KFCM receive low accuracy values when compared to those of the other methods involved in this experiment. The proposed methods obtained good accuracy, less running time and less number of iterations for clustering the Lung cancer dataset into three clusters.

From the results on Lung cancer dataset, this paper proves the impact of proposed methods via number of iterations, accuracy of clustering results and visual inspection of separation of clusters, that the proposed methods can have more capability to cluster the similar expression of genes in Lung cancer database.

Membership Comparison Test

The resulted membership of objects in each cluster on clustering Lung cancer database into two subtypes of cancers have been plotted in Fig. 12 to find the effect of membership equations of proposed methods in obtaining strong memberships to objects. It is observed from Figs. 12(a) and 12(b) that the GKFCM and KFCM provide weak memberships to the objects in Cluster 1 and Cluster 2 and the methods have less difference between the memberships of the objects between the first and second clusters. From Figs. 12(c) and 12(d) we can find that the proposed methods have provided

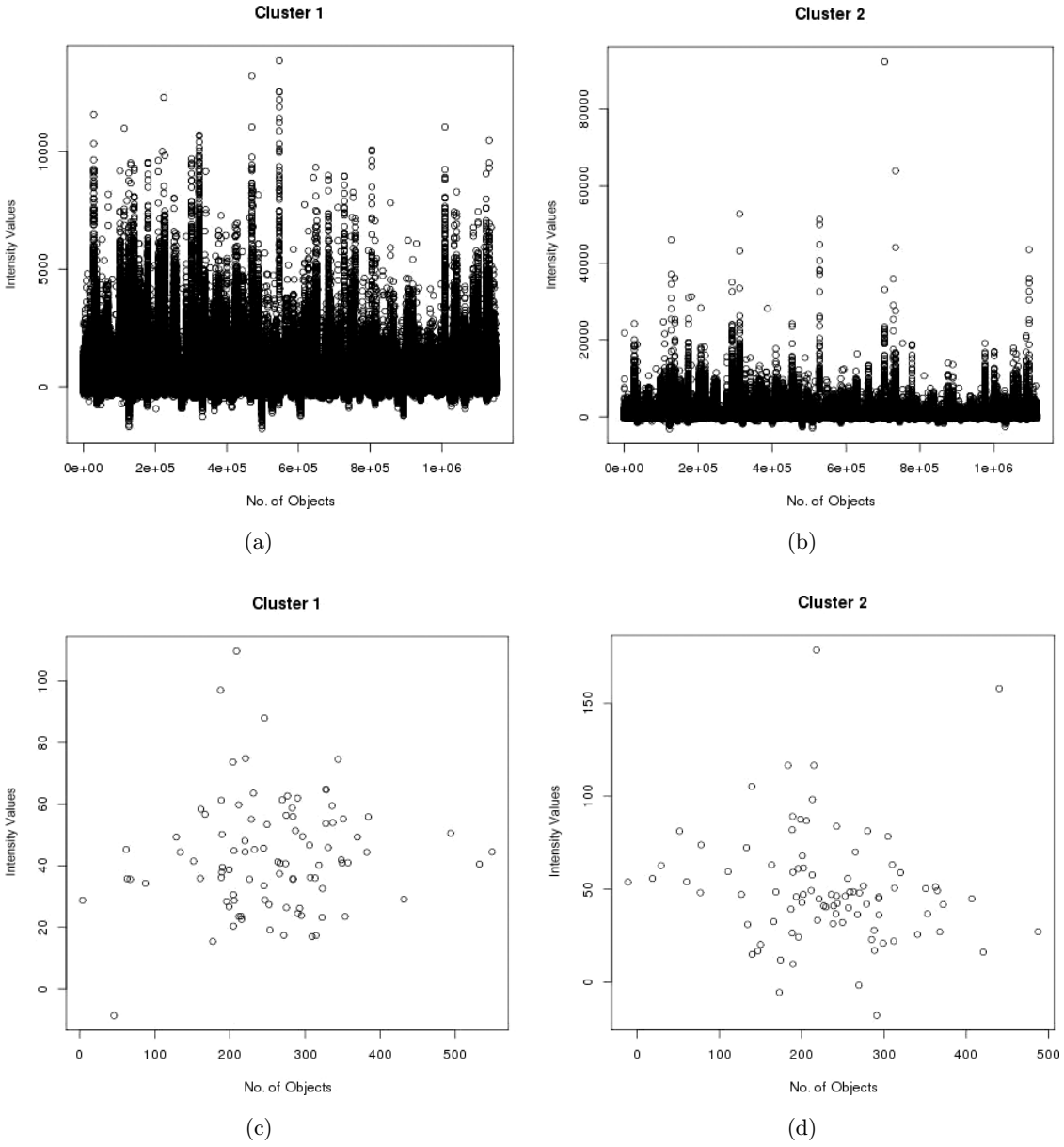

Fig. 9. Subtype of Cancers by KFCM: (a) Gene expression in Cluster 1 (b) Gene expression in Cluster 2 (c) Cluster 1 in 181 Lung Cancer Dataset (d) Cluster 2 in 181 Lung Cancer Dataset.

strong membership than GKFCM and KFCM for placing the objects in Cluster 1 and Cluster 2.

5.3. *Experimental results on benchmark datasets*

This subsection implements the proposed methods with Wine dataset, IRIS dataset, Checkerboard Dataset and Synthetic time series dataset in order to evaluate the performance of the proposed methods. The 178 instances with 13 constituents of Wine dataset[39] have been used by many researchers for comparing various clustering techniques.[40–42] The wine data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivators. Table 4 is listed the analysis determined the quantities of 13 constituents. The three classes of 150 instances of IRIS often used in the field of cluster analysis and data mining.[40,43,44] The 486 black with three attributes of checkerboard dataset[39] is widely used in performing the clustering methods.[45] Time series data analysis is most widely used at present in many areas and with special purposes, it is mainly used in

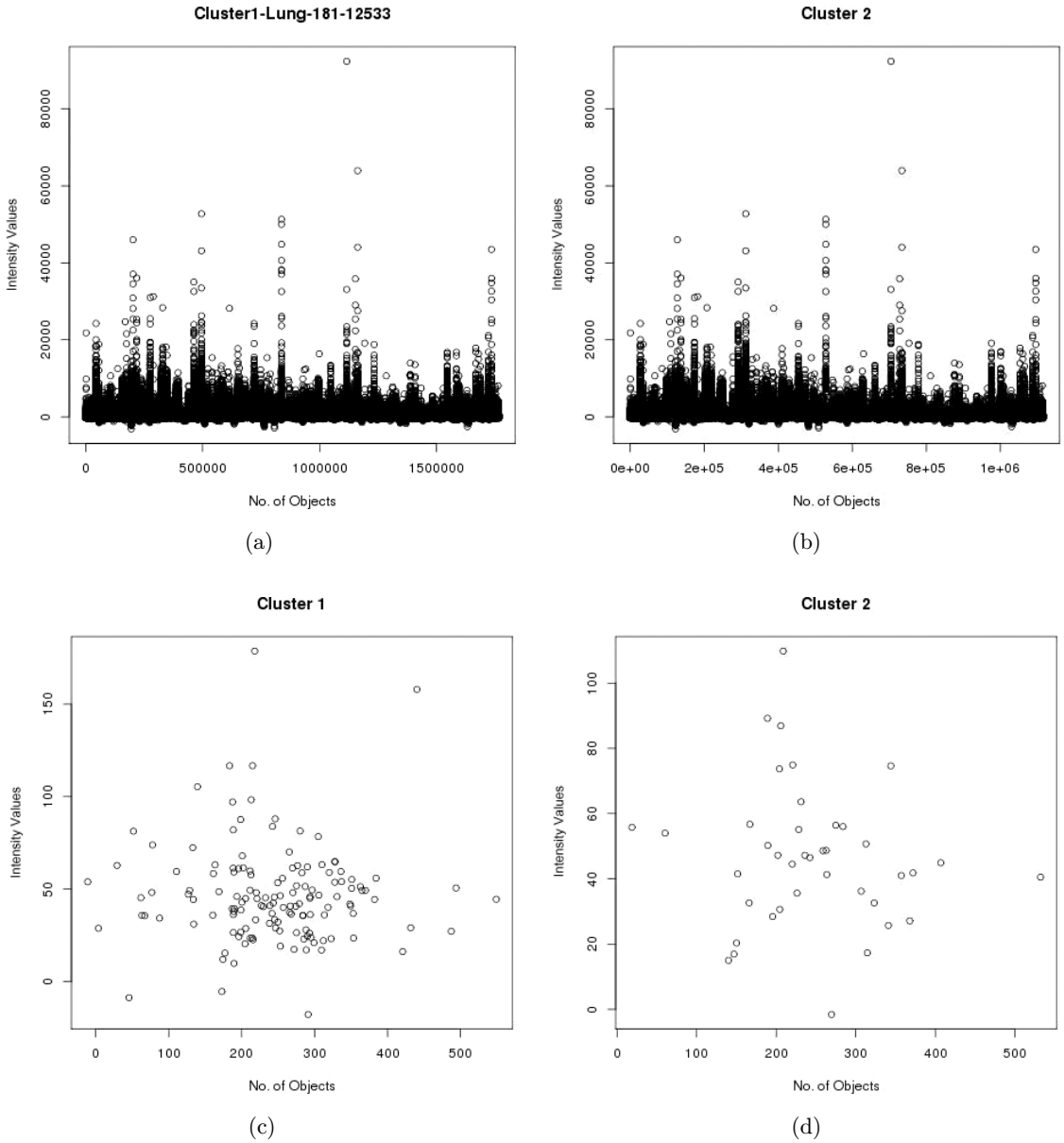

Fig. 10. Subtype of Cancers by KEFCM$_{wd}$: (a) Gene expression in Cluster 1 (b) Gene expression in Cluster 2 (c) Cluster 1 in 181 Lung Cancer dataset (d) Cluster 2 in 181 Lung Cancer dataset.

the area of data mining.[46–50] The time series dataset used in this subsection contains a total 300 control chart time series generated by the process in Alcock and Manolopoulos,[51] with three classes as follows: (Class one) 1–100 Normal, (Class two) 101–200 Cyclic, (Class three) 201–300 increasing trend. For visualization, the wine, IRIS, checkerboard data and time series dataset are given in Figs. 13–16. Since the benchmark datasets are having known number of clusters, this subsection corrupts the intensities of objects of datasets in order to run the proposed algorithm to cluster them into appropriate clusters.

The obtained size of clusters and accuracies on Wine dataset and IRIS dataset are shown in Figs. 17 and 18 and the allocated objects are given in Figs. 19 and 20.

The clustering accuracies using Silhouette width of KFCM, Proposed KEFCM$_{wd}$ and Proposed KFCM$_{nt}$ algorithms on Wine, IRIS, Checkerboard

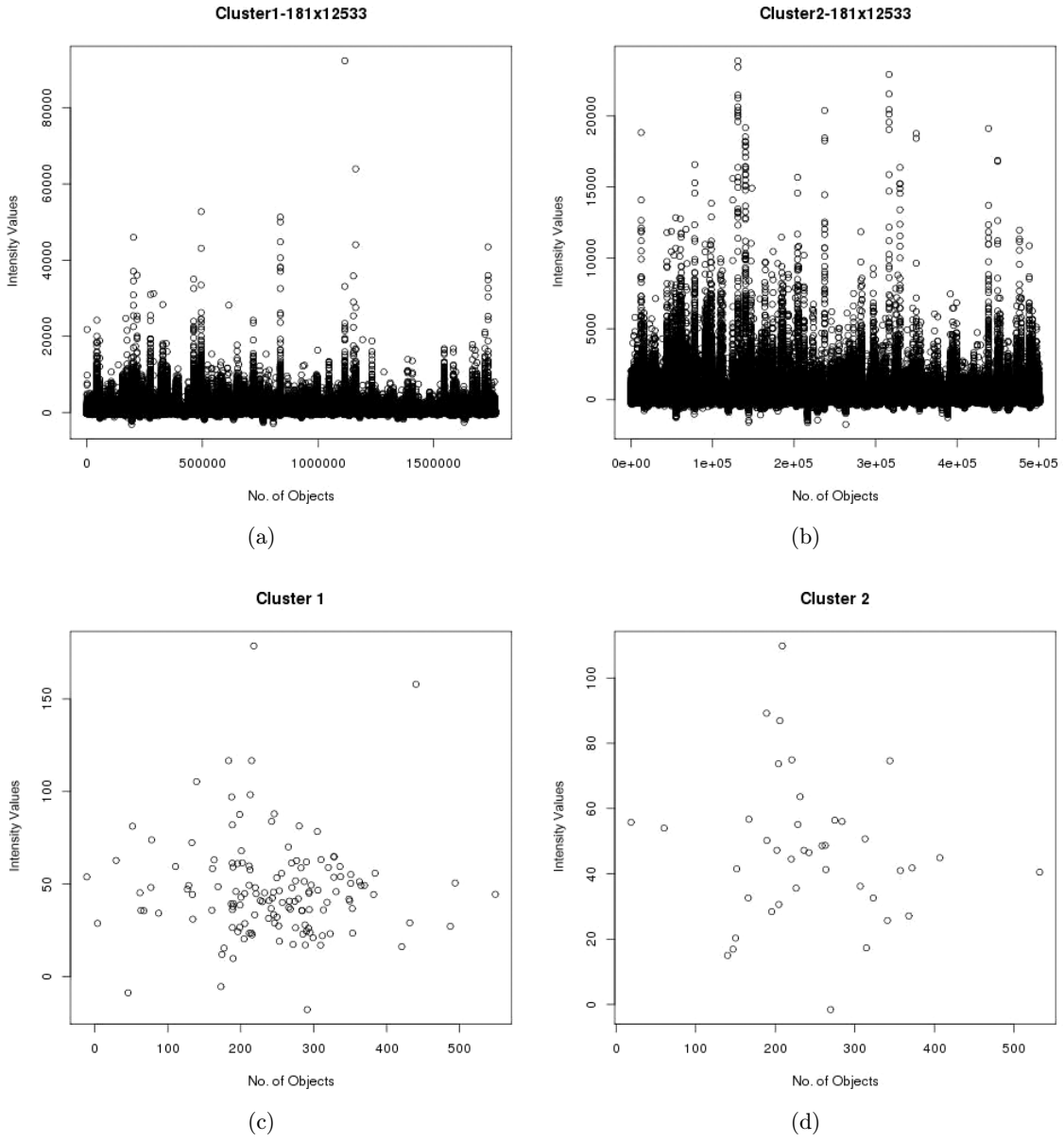

Fig. 11. Subtype of Cancers by $KFCM_{nt}$: (a) Gene expression in Cluster 1 (b) Gene expression in Cluster 2 (c) Cluster 1 in 181 Lung Cancer dataset (d) Cluster 2 in 181 Lung Cancer dataset.

Table 3. Comparison of Iteration Count (Its), Running Time (RT) and clustering accuracy (SW).

	Lung		
	SW	RT	Its
GKFCM	64%	56 s	27
KFCM	61%	1 min	30
$KEFCM_{wd}$	93%	5 s	11
$KFCM_{nt}$	93%	5 s	12

and Synthetic Control Time Series dataset are listed in Table 5. The Proposed $KEFCM_{wd}$ and $KFCM_{nt}$ algorithms have obtained good clustering results due to the objective function with kernel entropy and neighborhood term.

5.4. *Experimental results with breast cancer database*

This subsection uses 699 breast cancer datasets[39] given in Fig. 21 for the purpose of experimental works using the proposed clustering methods. The

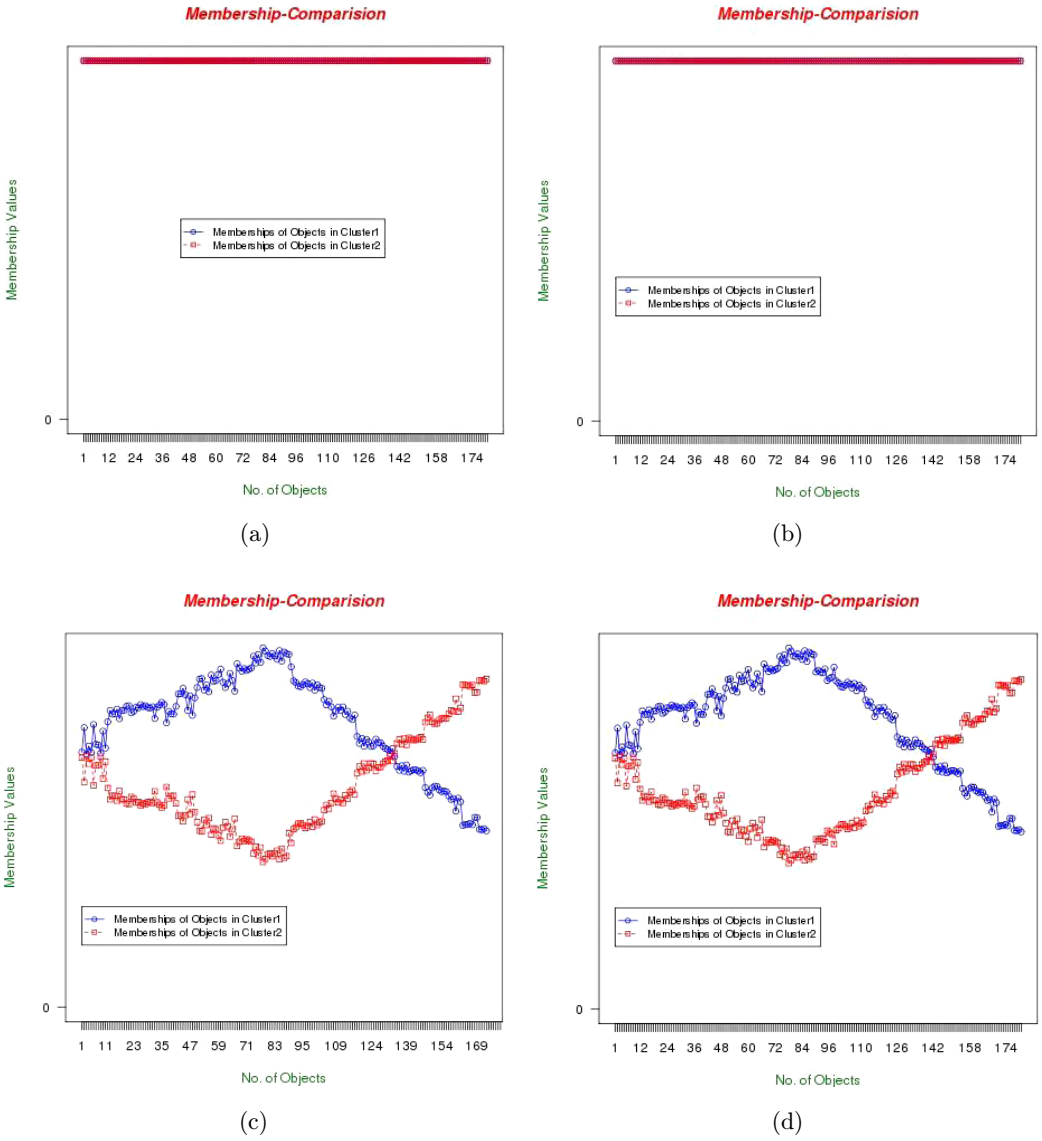

Fig. 12. Comparison of membership (a) by GKFCM (b) by KFCM (c) by $KEFCM_{wd}$ and (d) by $KFCM_{nt}$.

Table 4. The constituents of Wine dataset.

No.	Chemical name of constituent
1	Alcohol
2	Malic acid
3	Ash
4	Alcalinity of ash
5	Magnesium
6	Total phenols
7	Flavanoids
8	Nonflavanoid phenols
9	Proanthocyanins
10	Color intensity
11	Hue
12	OD280/OD315 of diluted wines
13	Proline

data consists of visually assessed nuclear features of fine needle aspirates (FNAs) taken from patients' breasts. Each data have been assigned nine-dimensional vectors by Dr. Wolberg. Each component is in the interval 1–10, with a value 1 corresponding to a normal state and 10 to a most abnormal state. The nine-dimensional vectors are: thickness, cell size, cell shape, marginal, adhesion, epithelial cell size, nuclei, chromatin, normal nucleoli and mitoses. The breast cancer data are used to make a decision on the medical condition that the cancer is malignant or benign.

GKFCM and KFCM results based on malignant and benign in 699 breast data are plotted in Figs. 22 and 24, respectively. The separated two classes for

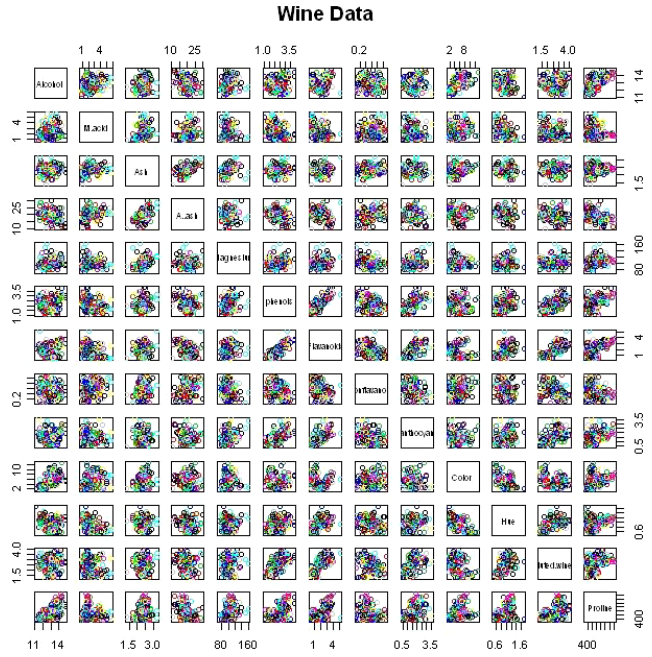

Fig. 13. 178 Wine Data.

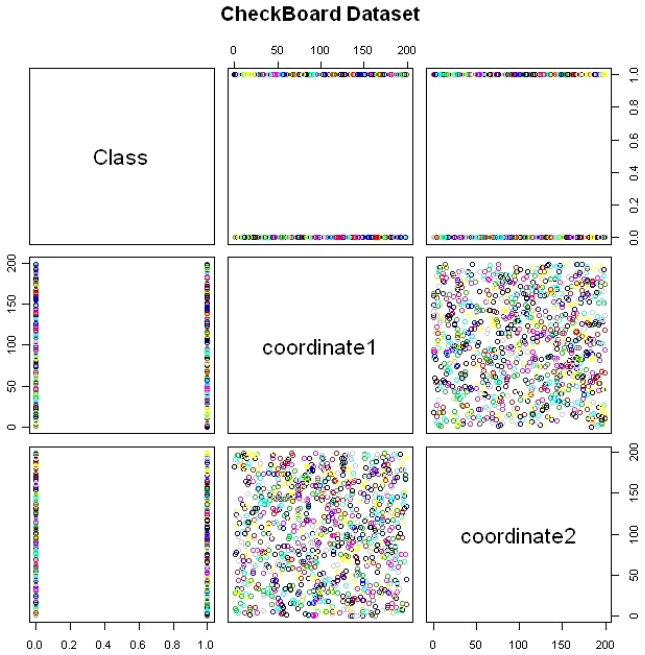

Fig. 15. 1000 Check board Dataset.

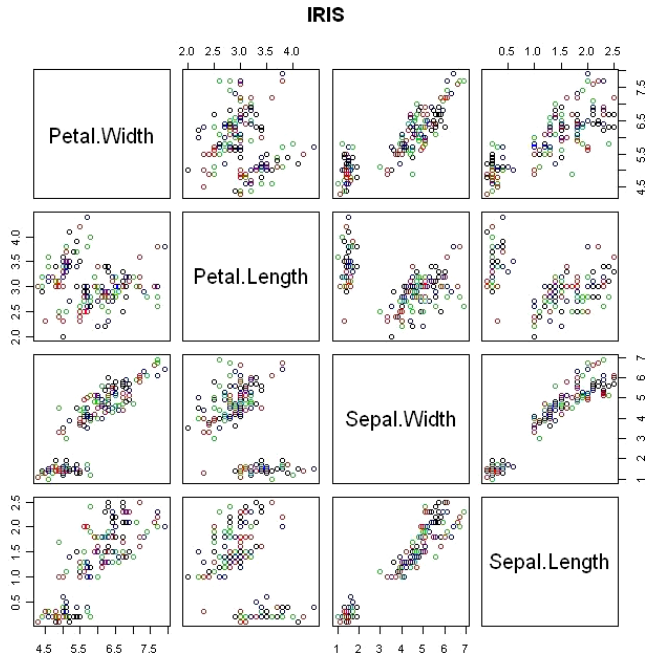

Fig. 14. 150 IRIS Dataset.

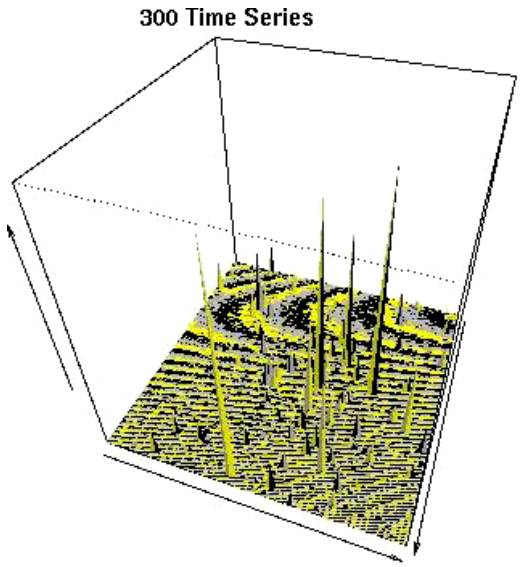

Fig. 16. 300 Time Series Dataset.

results of malignant and benign by GKFCM and KFCM are visualized in Figs. 23 and 25, respectively. The results of proposed KEFCM$_{wd}$ and KFCM$_{nt}$ methods based on malignant and benign on breast cancer dataset are shown in Figs. 26 and 28, respectively. The allocated two classes for malignant and benign of 699 breast cancer dataset by Proposed KEFCM$_{wd}$ and KFCM$_{nt}$ are given in Figs. 27 and 29, respectively. The captured size of two classes of malignant and benign by proposed KEFCM$_{wd}$ and KFCM$_{nt}$ algorithms are shown in Figs. 27 and 29, respectively. The clustering accuracy of GKFCM, KFCM, KEFCM$_{wd}$ and KFCM$_{nt}$

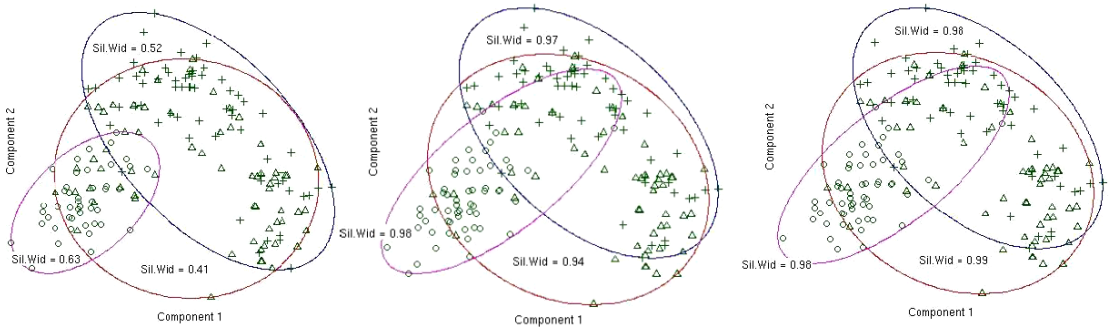

Fig. 17. Wine Dataset: (i) Result by KFCM (ii) Result by KEFCM$_{wd}$ (iii) Result by KFCM$_{nt}$.

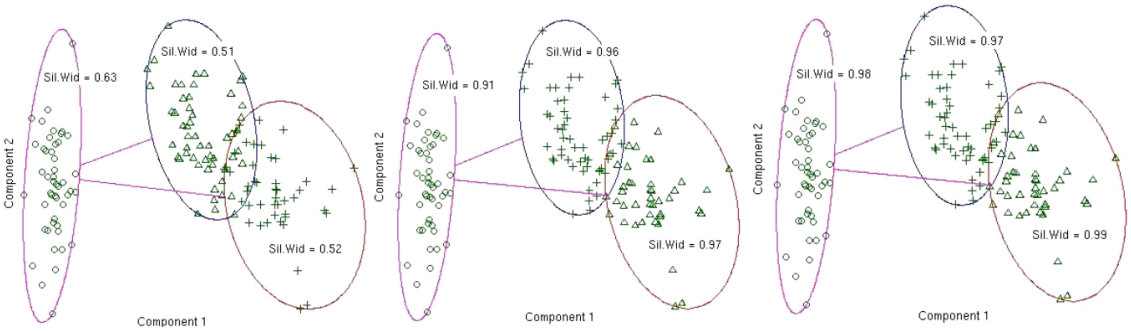

Fig. 18. IRIS Dataset: (i) Result by KFCM (ii) Result by KEFCM$_{wd}$ (iii) Result by KFCM$_{nt}$.

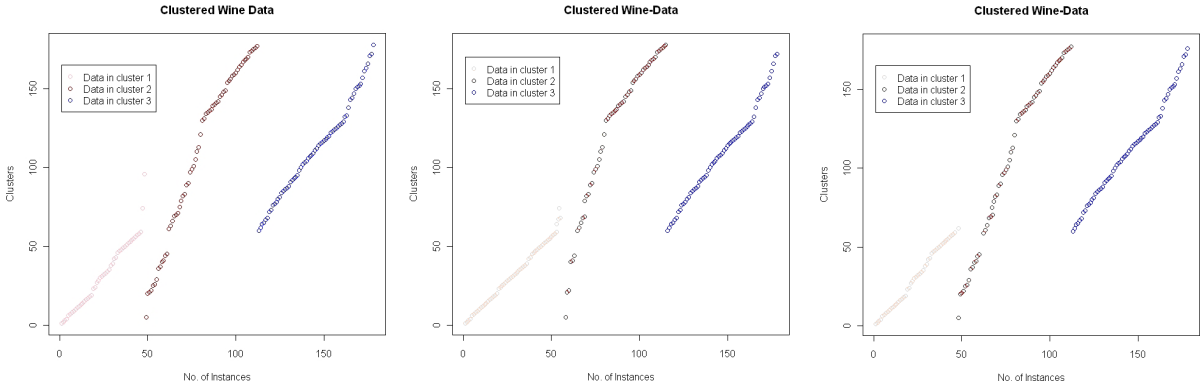

Fig. 19. Wine Dataset: (i) Result by KFCM (ii) Result by KEFCM$_{wd}$ (iii) Result by KFCM$_{nt}$.

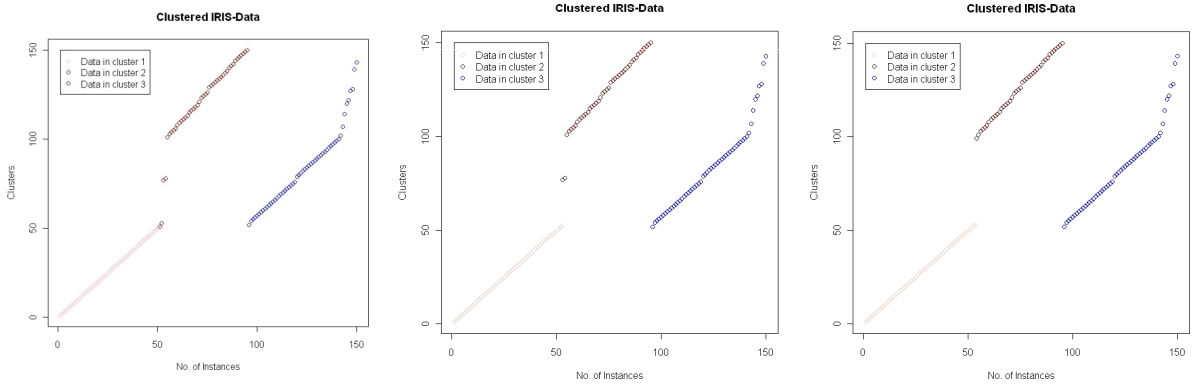

Fig. 20. IRIS Dataset: (i) Result by KFCM (ii) Result by KEFCM$_{wd}$ (iii) Result by KFCM$_{nt}$.

Table 5. Cluster results. Silhouette width [SW], Clustering Accuracy [CA], Minutes [M], Seconds [S], Running Time [RT] and Iterations [Its].

	No. of objects in cluster 1	SW	No. of objects in cluster 2	SW	No. of objects in cluster 3	SW	CA	RT	Its
				Wine data					
KFCM	56	0.52	59	0.63	63	0.41	52%	1.5 min	49
$KEFCM_{wd}$	48	0.97	65	0.98	65	0.94	96.3%	6 s	7
$KFCM_{nt}$	48	0.98	65	0.98	65	0.99	98.3%	5 s	5
				IRIS data					
KFCM	48	0.51	50	0.63	52	0.52	55.33%	1.6 min	40
$KEFCM_{wd}$	52	0.96	44	0.91	54	0.97	94.6%	9 s	7
$KFCM_{nt}$	51	0.97	44	0.98	55	0.99	98%	7 s	6

	Checkerboard			Time series		
	SW	RT	Its	SW	RT	Its
KFCM	61%	1.4 min	47	60%	1.3 min	50
$KEFCM_{wd}$	97%	8 s	7	95%	7 s	9
$KFCM_{nt}$	98.5%	7 s	7	98%	7 s	8

on clustering malignant and benign in breast cancer database is listed in Table 6. We can find from Table 6 that the proposed methods have better clustering accuracy during the experiment on breast

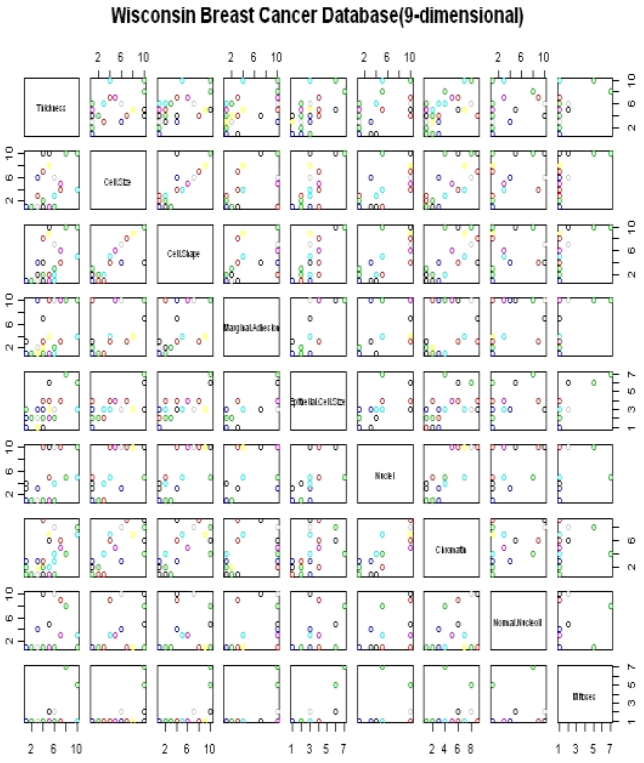

Fig. 21. 699 Breast Cancer Dataset.

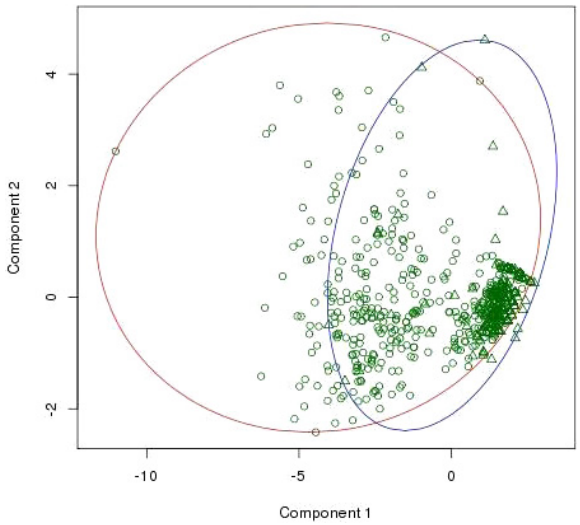

Fig. 22. Size of clusters by GKFCM.

699 dataset with two classes of malignant and benign.

The Error Matrix Table 7 gives the accuracy between the reference classes and the obtained classes by the methods involved in this experimental study with Breast Cancer database. From Tables 6 and 7, the best clustering accuracy was obtained for the proposed methods during the experiment on Breast Cancer data with two clusters.

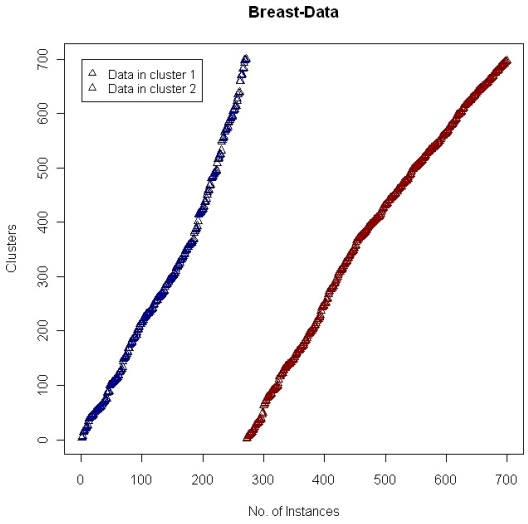

Fig. 23. Reallocated 699 Data by GKFCM.

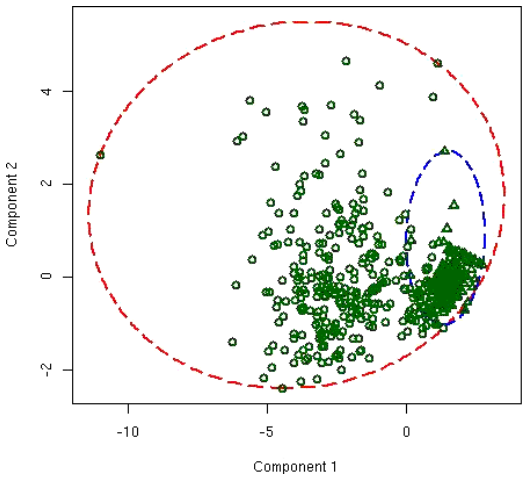

Fig. 24. Size of clusters by KFCM.

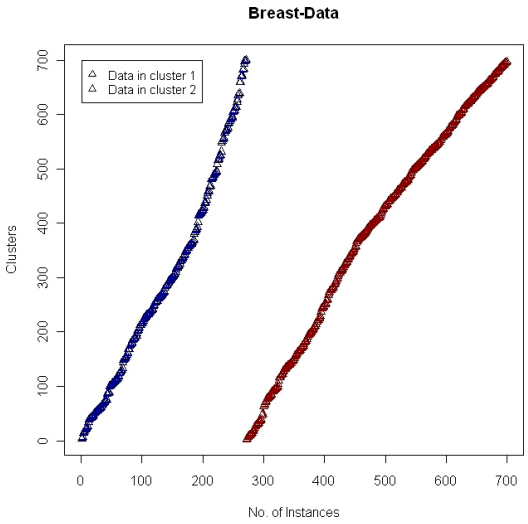

Fig. 25. Reallocated 699 data by KFCM.

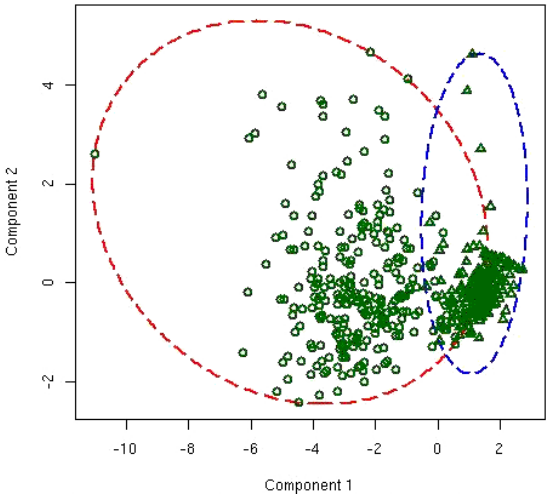

Fig. 26. Size of clusters by KEFCM_{wd}.

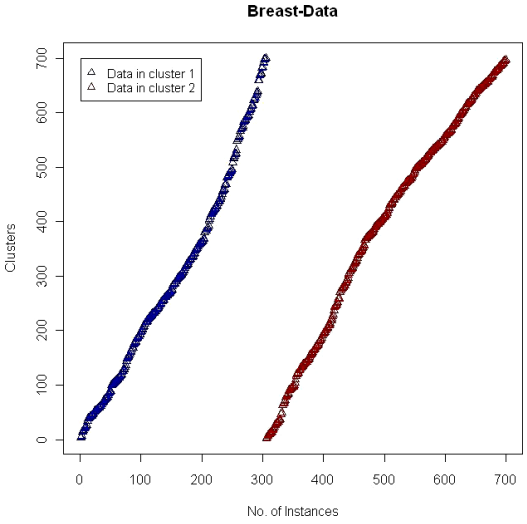

Fig. 27. Reallocated 699 data by KEFCM_{wd}.

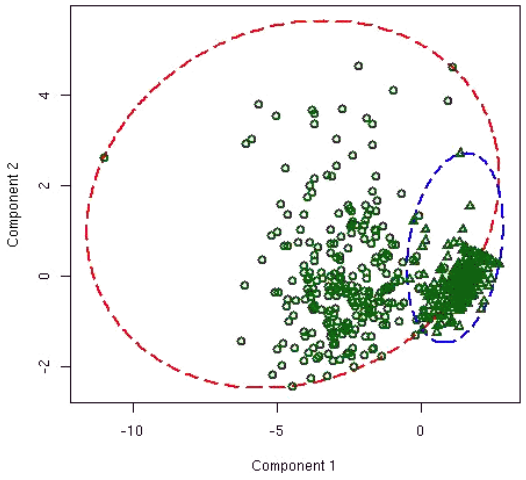

Fig. 28. Size of clusters by KFCM_{nt}.

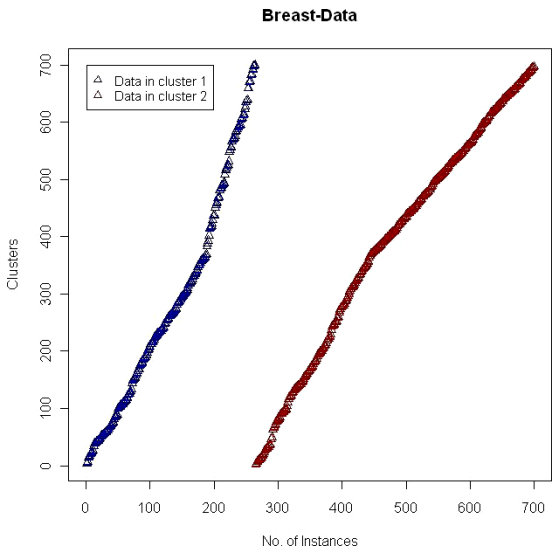

Fig. 29. Reallocated 699 data by $KFCM_{nt}$.

Table 6. Silhouette average values in clustering Breast cancer Data.

	Cluster1		Cluster 2		
	No. of items	SW	No. of items	SW	ASW
GKFCM	303	0.72	396	0.77	74.5%
KFCM	305	0.34	394	0.32	33%
$KEFCM_{wd}$	271	0.85	428	0.86	85.5%
$KFCM_{nt}$	266	0.84	433	0.87	85.5%

Table 7. Error Matrix on Breast Cancer Dataset.

	GKFCM	KFCM	$KEFCM_{wd}$	$KFCM_{nt}$
Accuracy	79%	55%	92.7%	93%

6. Conclusions

The novel fuzzy clustering algorithms $KEFCM_{wd}$ and $KFCM_{nt}$ have been developed for finding subtypes of cancers in the Breast cancer database. A center or prototypes knowledge method is introduced to speed up the convergence of the algorithms. This paper evaluated the performance of the proposed methods through the experimental works on Yeast, Lung Cancer, IRIS, Wine, Checkerboard and Time Series dataset. This paper has reported the superiority the proposed methods have shown using silhouette width, error matrix, running time, number of iterations and well-separated clusters. Finally, this paper has proved that the proposed methods are effective in clustering the breast cancer database into cancerous and noncancerous portions.

Acknowledgments

This work was supported by Indo Taiwan Joint Research Project, DST India & NSC Taiwan.

References

1. Hawes *et al.*, "DNA hypermethylation of tumors from non-small cell lung cancer (NSCLC) patients is associated with gender and histologic type," *Lung Cancer* **69**, 172–179 (2010).
2. D. M. Parkin, F. Bray, J. Ferlay, P. Pisani, "Global cancer statistics," *CA Cancer J. Clin.* **55**(2), 74–108 (2002).
3. J. Calle, "Breast cancer facts and figures 2003–2004," *American Cancer Society*, 1–27 (2004).
4. Y. N. Rao, S. Gupta, S. P. Agarwal, "National Cancer Control Programme: Current status and strategies, 50 years of cancer control in India," NCD Section, Director General of Health (2003).
5. H.-L. Chen, B. Yang, J. Liu, D.-Y. Liu, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis," *Expert Syst. Appl.* **38**, 9014–9022 (2011).
6. C. Liedtke *et al.*, "Systematic analysis of *in vitro* chemosensitivity and mib-1 expression in molecular breast cancer subtypes," *Eur. J. Cancer* **48**(13), 2066–2074 (2012).
7. F. P. Turkoz *et al.*, "Association between common risk factors and molecular subtypes in breast cancer patients," *Breast J.* **22**(3), 344–350 (2013).
8. Ramathilagam *et al.*, "Extended Gaussian kernel version of fuzzy c-means in the problem of data analyzing," *Expert Syst. Appl.* **38**(4), 3793–3805 (2011).
9. M. Kowal *et al.*, "Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images," *Comput. Biol. Med.* **43**(10), 1563–1572 (2013).
10. J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms* (Plenum Press, New York, 1981).
11. R. C. Dubes, A. K. Jain, Clustering methodology in Exploratory Data Analysis, *Advances in Computers*, M. C. Yovits, Ed., pp. 113–225 (Academic Press, New York, 1980).
12. Y. Yong, Z. Chongxun, L. Pan, "A novel Fuzzy C-means clustering algorithm for image thresholding," *Meas. Sci. Rev.* **4**, 11–19 (2004).
13. L. A. Zadeh, "Fuzzy sets," *Inf. Control* **8**, 338–353 (1965).

14. J. Dunn, "A fuzzy relative of the Isodata process and its use in detecting compact, well-separated clusters," *J. Cybernetics* **3**(3), 32–57 (1973).
15. J.-S. Lin, "Clustering problem using Fuzzy C-means algorithms and unsupervised neural networks," *Neuro-Fuzzy Pattern Recogn.* **41**, 75–99 (2000).
16. Ravi *et al.*, "Threshold accepting-based Fuzzy clustering algorithms," *Int. J. Unc. Fuzz. Knowl. Based Syst. World Scientific J.* **14**, 617 (2006).
17. Chaira *et al.*, "Intuitionistic Fuzzy C means clustering in medical image segmentation," *Adv. Pattern Recogn.* **1**, 226–230 (2007).
18. H. P. Ng *et al.*, "Fuzzy C-means algorithm with local thresholding for gray-scale images," *Int. J. Artif. Intell. Tools* **17**(04), 765–775 (2008).
19. K. Yuan *et al.*, "A novel Fuzzy C-means algorithm and its application," *Int. J. Pattern Recogn. Artif. Intell.* **19**(08), 1059–1066 (2005).
20. S. D. Hu, K. Tak, U. "A Novel video steganography based on non-uniform rectangular partition," *IEEE Int. Conf. Computational Science and Engineering CSE/ I-SPAN/IUCC 2011*, doi: 10.1109/CSE.2011.24.
21. Y. Wen, "Brain tissue classification based on DTI using an improved Fuzzy C-means algorithm with spatial constraints," *Magn. Reson. Imaging* **31**(9), 1623–1630 (2013).
22. Y. Li *et al.*, "Fast Fuzzy c-means clustering algorithm with spatial constraints for image segmentation," Advances in Neural Network Research and Applications, Springer (2010).
23. E. Çomak, "A biomedical decision support system using LS-SVM classifier with an efficient and new parameter regularization procedure for diagnosis of heart valve diseases," *J. Med. Syst.* **36**, 549–556 (2012), doi: 10.1007/s10916-010-9500-5.
24. L. Zhang *et al.*, "A novel ant-based clustering algorithm using Renyi entropy," *Appl. Soft Comput.* **13**, 2643–2657 (2013).
25. L. Bai *et al.*, "An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data," *Knowledge-Based Syst.* **24**, 785–795 (2011).
26. S. Shehroz *et al.*, "Cluster center initialization algorithm for K-means clustering," *Pattern Recognit. Lett.* **25**, 1293–1302 (2004).
27. H.-S. Tsai *et al.*, A Kernel-based Fuzzy C-means algorithm with partition index maximization, *Proc. 2010 Seventh Int. Conf. Fuzzy Systems and Knowledge Discovery, FSKD*, pp. 391–394 (2010).
28. M.-S. Yang *et al.*, "A Gaussian kernel-based fuzzy c-means algorithm with a spatial bias correction," *Pattern Recogn. Lett.* **29**, 1713–1725 (2008).
29. Saikumar *et al.*, "Robust adaptive threshold algorithm based on Kernel Fuzzy clustering on image segmentation," *The First Int. Conf. Information Technology Convergence and Services, ITCS 2012*, N. Meghanathan *et al.* Ed. SIP, JSE-2012, CS & IT 04, pp. 99–103 (2012), doi: 10.5121/csit.2012.2109.
30. S. Chen, D. Zhang, "Robust image segmentation using FCM with spatial constraints based on new Kernel-induced distance measure," *IEEE Trans. Syst. Man Cybern. B Cybern.* **34**(4), 1907–1916 (2004).
31. M. Girolami, "Mercer-based clustering in feature space," *IEEE Trans. Neural Netw.* **13**(3), 780–784 (2002).
32. G. N. Brock, V. Pihur, S. Datta, S. Datta, "clValid, an R package for cluster validation," *J. Stat. Software* **25**(4), 1–22 (2008).
33. P. J. Rousseeuw, "Silhouettes: A Graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.* **20**, 53–65 (1987).
34. A. Struyf, M. Hubert, P. Rousseeuw, "Clustering in an object-oriented environment," *J. Statistical Software* **1**(4), 1–30 (1997).
35. K. Mouhoubi *et al.*, A knowledge-driven biclustering method for mining noisy datasets, *Neural Information Processing*, Lecture Notes in Computer Science, Vol. 7665, pp. 585–593, Springer (2010).
36. D. Hou *et al.*, "An Efficient successive iteration partial cluster algorithm for large datasets, Fuzzy information and engineering 2010," *Adv. Intell. Soft Comput.* **78**, 557–562 (2010).
37. K. Das *et. al.*, "Empirical comparison of sampling strategies for classification," *Procedia Eng.* **38**, 1072–1076 (2012).
38. G. J. Gordon *et al.*, "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and Mesothelioma," *Cancer Res.* **62**, 4963–4967 (2002).
39. UCI Benchmark repository: A huge collection of artificial and real world data sets, University of California Irvine, Available at http://www.ics.uci.edu/∼mlearn.
40. C. Borgelt, R. Kruse, Speeding up Fuzzy clustering with neural network techniques, *Proc. 12th IEEE Int. Conf. Fuzzy Systems, FUZZ-IEEE'03*, St. Louis, MO, USA, IEEE Press, Piscataway, NJ, USA (2003).
41. E. J. Bredenstenier, K. P. Bennett, "Multicategory classification by support vector machines," *Comput. Optim. Appl.* **12**, 53–79 (1999).
42. J. G. Dy, C. E. Brodley, "Feature selection for unsupervised learning," *J. Mach. Learn. Res.* **5**, 845–889 (2004).
43. D.-Q. Zhang, S.-C. Chen, "Clustering incomplete data using Kernel-based Fuzzy C-means algorithm," *Neural Process. Lett.* **18**, 155–162 (2003).
44. X.-L. Yang, Q. Song, Y.-L. Wu, "A robust deterministic annealing algorithm for data clustering," *Data Knowl. Eng.* **62**, 84–100 (2007).
45. J. Gonzalez-Castillo *et al.*, "Whole-brain, time-locked activation with simple tasks revealed using

massive averaging and model-free analysis," *Proc. Natl. Acad. Sci. USA* **109**(14), 5487–5492 (2012).
46. D. Martínez-Rego, O. Fontenla-Romero, A. Alonso-Betanzos, "Efficiency of local models ensembles for time series prediction," *Expert Syst. Appli.* **38**, 6884–6894 (2011).
47. M. L. Hetland, "A survey of recent methods for efficient retrieval of similar time sequences," *Data Mining in Time Series Databases*, World Scientific (2002).
48. L. Singh, M. Sayal, "Privacy preserving burst detection of distributed time series data using linear transforms," *Proc. IEEE Symp. Computational Intelligence and Data Mining, CIDM 2007*, pp. 646–653 (2007).
49. Q. Wang, V. Megalooikonomou, C. Faloutsos, "Time series analysis with multiple resolutions," *Inf. Syst.* doi: 10.1016/j.is.2009.03.006.
50. Q. Wang, V. Megalooikonomou, "A dimensionality reduction technique for efficient time series similarity analysis," *Inf. Syst.* **33**, 115–132 (2008).
51. R. J. Alcock, Y. Manolopoulos, "Time-series similarity queries employing a feature-based approach," *7th Hellenic Conf. Informatics*, 27–29 August, 1999, Ioannina, Greece, 1999.