Online ISSN 1848-3380, Print ISSN 0005-1144
ATKAFF 57(3), 822–835(2016)

Abdullah M. Iliyasu, Chastine Fatichah, Khaled A. Abuhasel

Evidence Accumulation Clustering with Possibilitic Fuzzy C-Means base clustering approach to disease diagnosis

DOI 10.7305/automatika.2016.10.1427
UDK 004.85.021-048.44:616-079

Original scientific paper

Traditionally, supervised machine learning methods are the first choice for tasks involving classification of data. This study provides a non-conventional hybrid alternative technique (pEAC) that blends the Possibilistic Fuzzy C-Means (PFCM) as base cluster generating algorithm into the 'standard' Evidence Accumulation Clustering (EAC) clustering method. The PFCM coalesces the separate properties of the Possibilistic C-Means (PCM) and Fuzzy C-Means (FCM) algorithms into a sophisticated clustering algorithm. Notwithstanding the tremendous capabilities offered by this hybrid technique, in terms of structure, it resembles the hEAC and fEAC ensemble clustering techniques that are realised by integrating the K-Means and FCM clustering algorithms into the EAC technique. To validate the new technique's effectiveness, its performance on both synthetic and real medical datasets was evaluated alongside individual runs of well-known clustering methods, other unsupervised ensemble clustering techniques and some supervised machine learning methods. Our results show that the proposed pEAC technique outperformed the individual runs of the clustering methods and other unsupervised ensemble techniques in terms accuracy for the diagnosis of hepatitis, cardiovascular, breast cancer, and diabetes ailments that were used in the experiments. Remarkably, compared alongside selected supervised machine learning classification models, our proposed pEAC ensemble technique exhibits better diagnosing accuracy for the two breast cancer datasets that were used, which suggests that even at the cost of none labelling of data, the proposed technique offers efficient medical data classification.

Key words: Evidence accumulation clustering, K-means, fuzzy C-means, possibilitic fuzzy C-means, hybrid intelligent systems, health informatics, medical data classification, disease diagnosis.

Grupiranje zasnovano na skupljanju dokaza s vjerojatnosno-neizrazitim C-means pristupom za dijagnozu bolesti. Tradicionalno, metode nadziranog strojnog učenja predstavljaju prvi izbor za zadatke koji uključuju klasifikaciju podataka. Ovo istraživanje prikazuje nekonvencionalnu hibridnu alternativnu (pEAC) tehniku koja kombinira vjerojatnosno-neizraziti C-Means (PFCM) kao osnovni algoritam grupiranja u standardno grupiranje korištenjem grupiranja zasnovanog na skupljanju dokaza (EAC). PFCM objedinjuje zasebna svojstva vjerojatnosnog C-Means (PCM) i neizrazitog C-Means (FCM) algoritama u sofisticirani algoritam grupiranja. Usprkos ogromnim mogućnostima koje nudi ova tehnika, u smislu strukture, ona je nalik cjelovitim hEAC i fEAC tehnikama grupiranja realiziranim integracijom K-Means i FCM algoritama grupiranja u EAC tehniku. Kako bi se validirala učinkovitost, njeno ponašanje je ispitano na sintetičkim i stvarnim medicinskim podacima te su provedene usporedbe s pojedinačnim široko rasprostranjenim metodama, drugim nenadziranim tehnikama grupiranja i nekim nadziranim metodama učenja. Rezultat prikazuje kako predložena pEAC tehnika nadmašuje pojedine metode grupiranja i druge tehnike nenadziranog učenja u smislu točnosti u dijagnozi hepatitisa, kadiovaskularnih bolesti, raka dojke i dijabetesa, korištenih u eksperimentu. Značajno, u usporedbi s odabranim nadziranim modelima klasifikacije, predložena pEAC tehnika pokazuje bolju točnost dijagnoze na dvama korištenim bazama podataka za rak dojke, što ukazuje na to da čak i bez označenih podataka predložena tehnika nudi efikasnu klasifikaciju medicinskih podataka.

Ključne riječi: grupiranje zasnovano na skupljanju dokaza, K-means, neizraziti C-means, vjerojatnosno-neizraziti C-means, hibridni inteligentni sustav, medicinska informatika, klasifikacija medicinskih podataka, dijagnoza bolesti.

1 INTRODUCTION

The growth of information technology (IT) has accelerated the development of research in medical data classification and disease diagnosis, thus making the process an interesting academic pursuit with numerous challenges for researchers. The disease diagnoses procedures involve the use of data from of a variety of sources within the health

care records, such as the health condition notes, results of laboratory, radiological and pathological tests and examinations, as well as results from numerous other sources [1]. These health records of patients are collated into a dataset that is later used to make initial diagnosis of new patients. This use of medical datasets for disease diagnosis based on available training/examples data from health records of patients is called a medical data classification task [1]. Medical data classification tasks are executed using different varieties of data types including text, signal, image, DNA, voice, etc. [1-10]. Some of the available literature [1-6] focus on medical data classification tasks for ailments such as diabetes, heart disease, hepatitis, Parkinson, liver, and cancer. Similarly, EEG and ECG signals are usually used in diagnosing other diseases such as epileptic seizure, schizophrenia, Alzheimer, asthma, and arrhythmia [7-11].

Traditionally, supervised learning methods are used for data classification and evidence from available literature has proven that they are efficient in most tasks, including classification of medical data and diagnosing diseases. However, in order to realise good results, these supervised methods are known to use many parameters such as learning rate, epoch, kernel and activation parameters for tuning [2]. Therefore, it takes time to obtain the best combination of parameter values needed to achieve the best accuracy in diagnosing diseases. In addition, each dataset has different optimal parameters according to the pattern of data [3]. To further determine the optimal parameters for supervised methods, some hybrid approaches that combine different optimisation algorithms with supervised classification methods have been proposed [11].

In contrast, the unsupervised methods commonly have less parameters, which could be either the number of clusters or some threshold, that are available for use in medical data classification. Motivated by this, some literature employed unsupervised clustering methods to solve classification problems. For example, [12] compares several clustering algorithms to predict final marks based on student participation in some forums. The objective of that study was to examine whether clustering methods could be used to replicate the success recorded via traditional classification algorithms. Similarly, in [13], the OWA-weight based clustering was used to solve some classification problems using dataset from the UCI Machine Learning Repository [14]. Also, the work in [15] used the discriminative subspace clustering method to classify video fragment formats, while [16] presented the use of the fuzzy C-Means (FCM) method for handwriting recognition of numbers. [17] presented a survey on the use of several clustering methods to solve classification problems. Among others, outcomes obtained from that work suggested that clustering methods could be used as good alternatives to solving various classification tasks,emphasizing the need for care in the selection of an appropriate clustering method and necessary tradeoffs, so that the best results are obtained.

Clustering of data entails discerning the pattern of available data in order to determine groups of data points and their relationship. Such relationships include whether those data points are similar (or related) to one another or different from (or unrelated to) the data points in other groups. An ensemble clustering method combines multiple partitions generated by different clustering algorithms for improved robustness, stability, and accuracy in comparison with the single clustering methods [18]. However, finding a consensus cluster from multiple partitions is still a difficult problem that has been approached from different angles including those that are graph-based, combinatorial or statistical in nature [19-21]. The Evidence Accumulation Clustering (EAC) method, proposed by Fred and Jain [22-23], employs a Hard C-Means (HCM) method (notably, the K-Means algorithm) to map the individual data partitions in a clustering ensemble into a new similarity measure between patterns [23]. A typical EAC ensemble method based on HCM base clustering (or simply hEAC ensemble technique) is executed in three stages: namely, the splitting, combination, and merging processes or steps [22-23]. However, with its tight and well-defined boundaries, the hEAC clustering technique suffers from lack of flexibility in terms of membership of clusters and the impact resulting therefrom on achieving the convergence needed to produce a stable combined clustering outcomes [24]. Consequently, the EAC ensemble method employing FCM algorithm as base cluster generator (or simply fEAC ensemble techniques) [25] was proposed to compensate for these shortcomings. Moreover, a comparison between hEAC and fEAC found that besides faster convergence than the hEAC technique, the fEAC technique offers improved accuracy for lower number of clusters than the hEAC technique [24]. While still very successful in many applications, the fEAC technique suffers from sensitivity to initialisation and its inability to accurately correlate cluster membership with degree of belonging to a data class, which arises because estimation of centroids is influenced by noise in the data [25].

The Possibilistic FCM (PFCM) clustering approach, which is a hybrid unsupervised method that combines the FCM algorithm with the Possibilistic C-Means (PCM) algorithm, is obtained by relaxing the partitioning constraint of the FCM algorithm so that a 'possibilistic' type of membership is realised [26, 27]. Furthermore, by integrating a penalty term into the PCM, the relative values of degrees of membership that are needed for enhanced parameter estimation [26] are guaranteed. The resulting PFCM clustering algorithm combines the intuitionistic degrees of membership of FCM with the possibilistic typicality of PCM

in order to avoid noise sensitivity and to overcome coincident clusters, while maintaining the usual cluster centres for each cluster.

In addition to such uses, clustering approaches have the ability to recognise the distribution of the data and its pattern in order to ameliorate the detection of noise and other outliers that can increase the misclassification of data.

Motivated by the aforementioned capabilities of the PFCM algorithm and other unsupervised clustering approaches, in this study, we explore the integration of the of the PFCM clustering method as the base cluster generating algorithm for the EAC ensemble technique. Furthermore, the resulting pEAC technique is tailored towards use in data classification. Similar to existing unsupervised clustering methods (i.e. the hEAC and fEAC ensemble techniques), the PFCM algorithm is embedded into the splitting step of the proposed pEAC ensemble clustering technique. While the focus is on profiting from the combined advantages of the PCM and FCM (which coalesce into the PFCM algorithm), unlike in [24], in this study, preference is given to using the Active Link (AL) hierarchical agglomerative linkage in the merging step. In addition to evaluation of the proposed pEAC ensemble clustering technique using synthetic data, its performance is validated using real medical datasets used in the diagnosis of hepatitis, heart, diabetes, and breast cancer ailments. Additionally, the results are compared alongside those obtained for individual runs of well-known clustering methods, as well as other unsupervised ensemble techniques and some supervised learning methods.

The remainder of the paper is outlined thus: Section 2 presents a brief overview of the EAC as an ensemble or consensus clustering method. The layout and requirements for using the proposed pEAC ensemble clustering technique in medical data classification are presented in Section 3. The experimental results validating the use of the proposed method for disease diagnosis are presented and discussed in Section 4.

2 REVIEW OF THE EVIDENCE ACCUMULATION CLUSTERING (EAC) ENSEMBLE TECHNIQUE AND UNSUPERVISED BASE CLUSTERING METHODS

The Evidence Accumulation Clustering (EAC) [21-23] is an ensemble clustering method for combining multiple clustering approaches in order to achieve better performance than is obtainable using single clustering methods. Ensemble clustering uses a consensus of several clustering solutions and merges them into a single consensus solution, so that improved robustness and stability can be achieved in comparison with the single clustering method [18, 19]. There are several approaches to accumulate evidence of clustering methods: the first involves combining the results of different clustering algorithms; the second requires resampling the data using different techniques, such as bagging and boosting, so that different results are created; and the third applies a clustering algorithm many times, each with different initialisation [21-23]. The EAC ensemble method, as proposed by Fred and Jain in [21-23], uses the third approach by applying HCM algorithms, such as K-Means algorithm, which will henceforth be referred to simply as the hEAC technique, as the underlying clustering algorithm to produce clustering ensembles, as explained in the sequel.

Table 1. Pseudo-code for the hEAC-AL ensemble technique [21,22]

Input: *n*: *d*-dimensional patterns; *k_min*: minimum initial number of clusters; *k_max*: maximum initial number of clusters; *N*: number of ensemble clusterings. *Output*: Data partitioning.
Initialisation:Set *co_assoc* to a null *n x n* matrix. **1.** Do *N* times: **1.1.** Randomly select *k* in the interval [*k_min*; *k_max*]. **1.2.** Randomly select *k* cluster centers. **1.3.** Run the K-means algorithm with the above *k* and initialisation, and produce a partition *P*. **1.4.** Update the co-association matrix: for each pattern pair, (*i, j*), in the same cluster in *P*, set $co\_assoc(i,j) = co\_assoc(i,j) + \frac{1}{N}$ **2.** Detect consistent clusters in the co-association matrix using the AL technique: compute the AL dendrogram and identify the final clusters as the ones with the highest lifetime.

2.1 Unsupervised ensemble clustering techniques

The hEAC ensemble clustering method is executed in three broad steps or processes, namely: the splitting, combination, and merging processes [21-23]. In the first step, the data is split into a large number of clusters and different partitions are obtained by random initialisations of the K-means algorithm. Following this, in the combination process, a voting mechanism is used to combine the clustering results through a co-association matrix. This matrix is built using outcomes of multiple random runs of the HCM algorithm with specified number of clusters. Next, in order to recover natural clusters, the Single Link (SL) or Average Link (AL) hierarchical agglomerative set is applied in the merging process so that the recovery of the final partition is facilitated [23]. Table 1 presents a pseudo-code for the hEAC with AL link set (i.e. hEAC-AL ensemble technique).

As explained in [21, 23] and extracted in Table 1, the input parameters for the hEAC ensemble clustering technique are the minimum initial number of clusters (*k_min*), maximum number of clusters (*k_max*), and number of ensemble or consensus clusters (*N*). The value of *k_min* is preset to start from 2 while the value of *k_max* is set using a rule of thumb, i.e. $k\_max = \sqrt{n}$, where *n* is the number of input patterns. A brief description of each part of the three steps of the hEAC ensemble technique is presented in the following subsections.

2.2 EAC with HCM as base cluster generator (hEAC ensemble clustering technique)

As presented in earlier parts of this section, by viewing each clustering result as an independent evidence of data organisation, the idea of EAC was proposed to combine the result of multiple clusters into one partition. Since the original EAC uses the HCM algorithm to generate its base clustering, this class of ensemble techniques is simply referred to as the hEAC clustering technique.

Detailed descriptions of the three steps of the hEAC clustering ensemble technique mentioned earlier are presented in the remainder of this sub-section.

2.2.1 Splitting Step: Producing Clustering Ensemble

In the splitting step of the hEAC technique, an input dataset is decomposed into *k* number of clusters using a HCM clustering method, such as the K-means algorithm, with various clustering results obtained via random initialisations of the algorithm. The K-Means algorithm is a simple clustering method that has low computational requirements (time) for clustering the data points into *k* clusters [21] where the value of *k* is determined prior to running the algorithm after which it is then used to generate the *k* initial centroid clusters. Each cluster is associated with a centroid (center point) and each data point is assigned to the cluster with the closest centroid [21]. Finally, after all data points are assigned to a centroid, which is (typically) the mean of the data points in the cluster [21], they are updated. These processes are run iteratively until the centroid of clusters are somewhat unchanged, thereby producing *k* partitions of the dataset [21].

2.2.2 Evidence combination Step: Generating the Co-Association Matrix

In order to produce partitions with different number of clusters, a voting mechanism is used to combine the clustering results so that a new measure of similarity is created between patterns [22]. The underlying assumption is that patterns belonging to a "natural" cluster are very likely to be co-located (i.e. in proximity) within the same cluster [21]. Taking the co-occurrences of pairs of patterns in the same cluster as votes for their association, the data partitions produced by multiple runs of the K-means algorithm are mapped into an $n \times n$ co-association matrix [22] using equation (1).

$$co\_assoc\,(i,j) = \frac{n_{ij}}{N} \tag{1}$$

where *N* is the number of clusters and n_{ij} is the number of times the pattern pair (*i*, *j*) is assigned to the same cluster among the *N* clusters.

2.2.3 Merging Step: Recovering Natural Clusters

The hEAC ensemble clustering technique uses Single Link (SL) or Average Link (AL) linkage set to merge clusters based on the outcome of the co-association matrix that was presented in the preceding step. The SLand ALare examples of hierarchical clustering methods that produce a set of nested clusters organised as a hierarchical tree or a dendrogram [22, 23]. Any desired number of clusters can be obtained by 'cutting' the dendrogram at the proper level. In order to recover natural clusters, the SL or AL set cuts the dendrogram over a similarity matrix at some threshold (*t*), thus merging the clusters produced in the splitting process. There are two main types of hierarchical clustering techniques, i.e. agglomerative and divisive techniques [23]. The agglomerative approach starts with the data points as individual clusters and then the closest pair of clusters at each step are merged until only one cluster (or*k*clusters) is/are left [23]. The divisive approach starts with one cluster that is split at each step until each cluster contains a data point (or there are *k* clusters) [23].

In the merging step, the hEAC techniques uses the agglomerative approach to merge the clusters using one of several methods, i.e. including single link (SL), complete link (CL), average link (AL), etc. [23]. The SL set merges two clusters based on the two most similar (closest) data points in the different clusters, while the CL merges two clusters based on the two least similar (most distant) data points in the different clusters. Finally, the AL merges two clusters based on the average of pairwise similarity between data points in the two clusters [23].

One drawback of the hEAC technique arises during the merging process because, like ether methods that use co-association matrices, it requires a large number of base clustering in order to achieve reliable results. Additionally, the hEAC ensemble techniques have poor convergence, which ultimately affects the accuracy of the results obtained.

2.3 EAC with FCM as base cluster generator (fEAC ensembleclustering technique)

With its tight and well-defined boundaries, the hEAC ensemble technique that was presented earlier in this section lacks intuitionistic flexibility in terms of membership of its clusters [25]. By reducing the variances from clusters, FCM yields to more compact clusters, as well as providing additional information about the spatial distribution of data that is obtained via fuzzy partitioning. These are some of the qualities that motivated the integration of FCM as base cluster generator for EAC ensemble clustering technique (i.e. fEAC ensemble clustering technique), which has contributed towards improvement in the speed of convergence as a factor of the number of runs of the base clustering generator, which is crucial in producing a stable system.

2.4 Drawbacks of hEAC and fEAC ensemble clustering techniques

A few of the problems associated with the hEAC technique, such as the large size of base clustering in the merging step and poor convergence, were highlighted earlier as motivations for proposing the fEAC technique. While, the fEAC ensemble technique offers better performance than in the hEAC technique, the FCM that is at its nexus exposes the fEAC technique to increased susceptibility to noise, sensitivity to initialisation and poor handling of coincident clusters all of which impact on the performance of the fEAC technique. Additionally, in FCM, the membership of a given cluster of two points that are equidistant from the prototype of the cluster can be significantly different, whereas membership of two points in a given cluster can be equal even though the two points are arbitrarily far away [28]. This enunciates the shortcomings of the FCM (and the resulting fEAC ensemble technique) in terms of detection of noise and outlier points.

As an unsupervised approach, the fEAC ensemble technique is further affected by the inability of the FCM base cluster generator to accurately correlate cluster membership with degree of belonging to a data class [22], which arises because estimation of centroids is influenced by noise in the data.

The fragility inherent to both the hEAC and fEAC techniques as highlighted in this section provide further the motivation for the need to integrate a more robust and effective base cluster generating algorithm into the EAC ensemble technique, which, among others, is one of the main objectives of this study. Therefore, the base cluster generator envisioned in our proposed technique should be able to compensate for the identified lapses by ensuring that convergence is accelerated, so that accuracy is improved and effects of noise points are suppressed.

In the next section, we highlight a few properties of the Possibilistic Fuzzy C-Means (PFCM) [26] clustering algorithm and discuss its integration as base cluster generating algorithm of our proposed EAC ensemble technique, which, for uniformity, we shall refer to as the pEAC ensemble technique. Moreover, since it has been shown that the performance of both the hEAC and fEAC ensemble techniques varies with type of data [27] our new ensemble clustering technique should offer a wider range of utility.

Before then, we should emphasise that, structurally, our proposed pEAC ensemble clustering technique resembles the standard unsupervised ensemble techniques (i.e. the hEAC and fEAC techniques) differing mainly in terms of its replacement of the HCM and FCM base cluster generators with the more efficient PFCM clustering algorithm. In this manner, performance-wise, we expect a more robust and effective ensemble clustering, which we envisage would make it more useful for our applications in medical dataset classification.

3 METHODOLOGY FOR THE PROPOSED PFCM-BASED ENSEMBLE CLUSTERING TECHNIQUE

This section highlights our proposed blending of the Possibilitic Fuzzy C-Means (PFCM) as base cluster generator of the Evidence Accumulation Clustering (EAC) ensemble method (henceforth simply referred to as pEAC ensemble clustering technique) for the purpose of efficient use of medical datasets for diagnosis of diseases.

3.1 Possibilistic Fuzzy C-Means Clustering Method

Earlier in Section 2, we highlighted the structure of the unsupervised ensemble clustering methods (hEAC and fEAC) and disclosed a few of their shortcomings. Here, we present an introduction of the PFCM clustering algorithm and advance arguments supporting its integration into the EAC ensemble method for later use in classification of medical datasets.

The PFCM was proposed to compensate for individual deficiencies in the Fuzzy C-Means (FCM) and Possibilistic C-Means (PCM) clustering [29] algorithms, making it a good candidate to use as EAC base clustering generator, which forms the core of our proposed pEAC ensemble clustering technique.

Conceived as a hybrid clustering algorithm, the PFCM combines the intuitionistic degree of membership inherent to FCM algorithms with the possibilistic typicality found in PCM algorithms. To accomplish this, the partitioning constraint in the FCM algorithm is relaxed, yielding a 'possibilistic' type of membership, which, together with the penalty term imposed on the PCM algorithm, guarantees the realisation of relative values for degrees of membership as needed to enhance parameter estimates [26, 30]. Additionally, the PFCM clustering algorithm is useful in

overcoming sensitivity to noise and in avoiding coincident clusters, while still being able to produce the usual point prototypes or cluster centres for each cluster [26, 30].

Mathematically, the objective function of the PFCM is in the form presented in equation (2).

$$J(U,T,V;X) = \sum_{k=1}^{n}\sum_{i=1}^{c}(au_{ik}^{m} + bt_{ik}^{\eta}) \times \|x_k - v_i\|_A^2 + \sum_{i=1}^{c}\gamma_i\sum_{k=1}^{n}(1-t_{ik})^{\eta} \quad (2)$$

Equation (2) is subject to the following constraints [26]:

$$\sum_{i=1}^{c} u_{ik} = 1, \quad \forall\, k \quad (3)$$

$$0 \le u_{ik}, t_{ik} \le 1 \quad (4)$$

where the variables a and b are membership and typicality values of the objective functions; the parameter γ_i is a constant value greater than zero; a and b are constants whose values must also be greater than zero. In order to assign the same weight of membership (u_{ik}) and typicality (t_{ik})), the values of a and b are preset as 1. Similarly, m and η are constants whose values must be greater than 1 and like in the FCM algorithm (for the degree of fuzziness) their values are preset to 2 [26]; the membership (u_{ik}) and typicality (t_{ik}) values have similar definitions to those in the standard FCM and PCM algorithms [25]; and x_k is the data value. Equations (5) through (7) are used to calculate the membership value (u_{ik}), the typicality value (t_{ik}), and the cluster center (v_i) [26].

$$u_{ik} = \left(\sum_{j}^{c}\left(\frac{D_{ikA}}{D_{jkA}}\right)^{\frac{2}{m-1}}\right)^{-1}; 1 \le i \le c; 1 \le k \le n \quad (5)$$

$$t_{ik} = \frac{1}{1 + (\frac{b}{\gamma_i}D_{ikA}^2)^{\frac{1}{\eta-1}}}; 1 \le i \le c; 1 \le k \le n \quad (6)$$

$$v_i = \frac{\sum_{k=1}^{n}(au_{ik}^{m} + bt_{ik}^{\eta})x_k}{\sum_{k=1}^{n}(au_{ik}^{m} + bt_{ik}^{\eta})}; 1 \le i \le c \quad (7)$$

where D_{ikA} is distance between data (x_k and the cluster center (v_i) [26].

In fuzzy clustering, a data point belongs to every cluster with some weight usually between 0 and 1, but with the additional constraint that the weights must add up to 1. Therefore, unlike in HCM clustering methods, there is no empty cluster in PFCM-based clustering because each data point should be assigned into one or more clusters. Similar to the FCM algorithm, the PFCM determines the initialisation of c centroid clusters randomly and then the new centroid clusters are updated iteratively. To determine the final cluster, each data point belongs to the cluster that has maximum level value of membership function.

3.2 EAC with PFCM as base cluster generator (pEAC ensemble clustering technique)

The system architecture depicting our proposed EAC ensemble clustering method with PFCM algorithm as base cluster generator (i.e. the pEAC ensemble clustering technique) is presented in Figure 1.

As seen from the figure, similar to the use of HCM and FCM as base cluster generators (i.e. in hEAC and fEAC ensemble clustering techniques), in the splitting step of the proposed technique, data is divided into a large (say, c) pool of smaller clusters that are mapped into $n \times n$ co-association matrices using the evidence accumulation technique in the combination step. The final clusters are obtained by applying the AL hierarchical link set, thus merging the hitherto smaller clusters obtained in the earlier splitting step.

Based on the foregoing, the final data partition is chosen as the one with highest lifetime based on which the optimal centroid clusters are determined. Depending on the dataset, the Euclidean distance measure (in equation 6) is used to suggest an inference regarding the final diagnosis of an ailment.

$$d(x,c) = \sqrt{\sum_{i=1}^{N}(x_i - c_i)^2} \quad (8)$$

where *xi* is *i*th data point and *ci* is *i*th centroid of clusters.

A well-known shortcoming of the SL linkage set is its quadratic space and time complexities,which are related to the requirements for processing an $n \times n$ co-association matrix, especially for large values of n [30]. To circumvent this kind of effect, the AL hierarchical linkage set is preferred in this study, hence its use as shown in the merging step of the proposed pEAC ensemble clustering technique (i.e. in Figure 1).

Since clustering assumes that there is no a priori knowledge of samples of known classes, then hierarchical clustering could be considered as an example of unsupervised classification where all patterns in the data are classified using the top-down or bottom-up approach [28]. Therefore, it is acceptable to conclude that unsupervised classification offers the advantages of reduced need for a priori knowledge, less human error, better recognition of classes, etc. Moreover, it has been shown that, for certain tradeoffs, the clustering approach to classification is very effective [27, 28].

Motivated by this and in order to exploit the performance improvements offered by the PFCM clustering method over HCM and FCM clustering approaches, this study utilises the pEAC ensemble technique presented in this section for medical datasets classification tasks and applications in disease diagnosis.

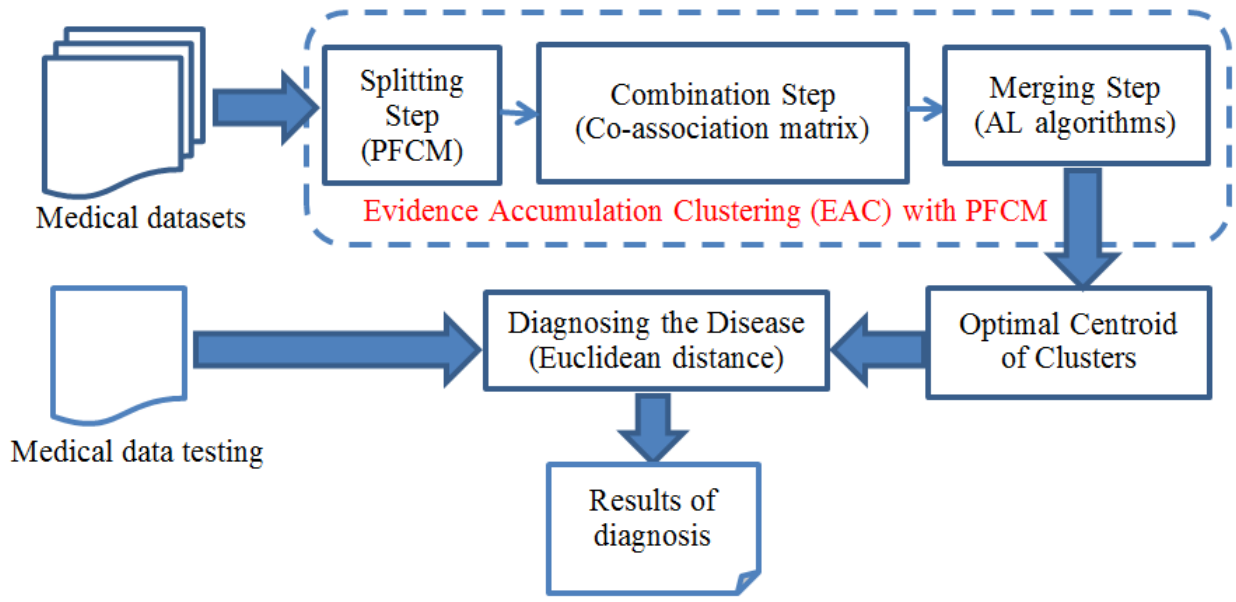

Fig. 1. System architecture for the proposed EAC the pEAC ensemble clustering technique (pEAC) for disease diagnosis.

Our proposed pEAC ensemble clustering approach offers accelerated convergence, which would lead to improved accuracy, as well as better tolerance to noise and outliers than the hEAC and fEAC unsupervised ensemble clustering techniques, which we validate via a series of experiments in the next section.

4 EXPERIMENTAL VALIDATION OF PROPOSED SYSTEM

To evaluate the performance capabilities of our proposed pEAC ensemble clustering technique, two experimental scenarios were used. In the first scenario, similar to [24], we utilise synthetic datasets consisting of the Half ring and Spiral datasets. Using real clinical datasets, our second performance evaluation scenario is further divided into two segments: the first evaluates the performance of the proposed pEAC ensemble technique alongside other unsupervised clustering methods, whereas the performance of the proposed technique is compared alongside some supervised (machine learning) models in the second segment.

Unless stated otherwise, throughout the experiments that are reported in this section, our input variables are the number of clusters (N), minimum number of clusters (k_min), and maximum number of clusters (k_max). Additionally, throughout the experiments reported in this section, we used variables: $N = 100$, $k\_min = 2$, and $k\_max = \sqrt{n}$ (where n is size of the dataset).

Brief descriptions of the synthetic and real clinical datasets for the two experimental scenarios and discussions on the results obtained are presented in the remainder of this section.

4.1 Evaluation of proposed technique using synthetic datasets

Our first experimental validation is based on two types of synthetic datasets, i.e. the half rings and spiral datasets, which are generated and utilised to assess the performance of the proposed pEAC clustering approach.

The Half rings dataset, shown in Figure 2, consists of two clusters with uneven sparseness that is made up of 100 data points in the upper cluster and 300 data points in the lower cluster.

Similarly, as shown in Figure 3, the Spiral dataset consists of two spiral-shaped clusters comprising of 200 data points.

4.1.1 Experimental results based on Synthetic datasets

The experimental results for the Half ring dataset with variations in parameter pair values, i.e. the threshold (t) and number of cluster (k) as presented in Table 3. As seen therein, these results indicate that the best accuracy in final cluster results for this dataset is obtained with parameter value pairings of $t = 0.2$ and $k = 10$; $t = 0.1$–0.3 and $k = 15$; $t = 0.1 - 0.2$ and $k = 20$ with the optimal values shown highlighted in the table.

The results suggest that the satisfactory solution would be the identification of two clusters that produce the best

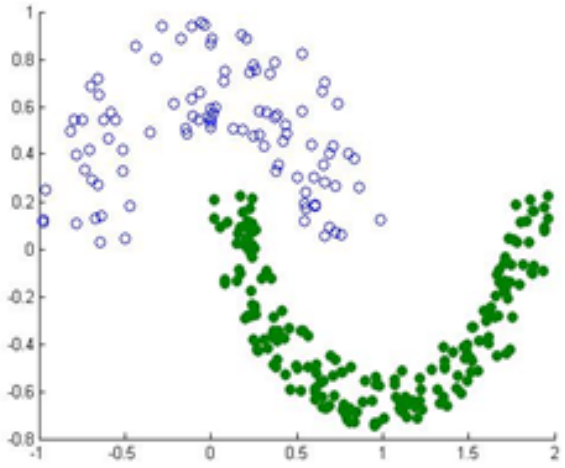

Fig. 2. Half ring dataset.

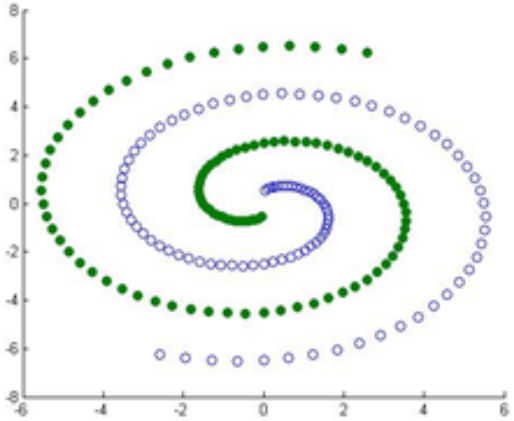

Fig. 3. Spiral dataset.

accuracy in the clustering results. Figure 4 presents a visualisation of the clustering outcome for optimal parameter pair values $k = 15$ and $t = 0.2$.

Similar to the case of the Half ring dataset, the experimental result for the Spiral dataset with variation in parameter pair values for threshold (t) and number of cluster (k) is presented in Table 4. These results indicate that the best accuracy of final cluster results for this dataset are obtained for pairing of parameter values $t = 0.6$ and $k = 20$; $t = 0.7$ and $k = 15$ and $k = 20$.

These optimal parameter values produce the best accuracy in clustering results (shown highlighted in the table), whereas Figure 5 presents the clustering outcome for optimal parameter pair values $k = 15$ and $t = 0.7$. Meanwhile, the visualisation of the clustering outcome for a non-optimal clustering results for parameter pair values $k = 3$ and $t = 0.9$ is presented in Figure 6.

Table 2. Results for Half ring synthetic dataset for variations in parameter values t and k

Threshold	The number of clusters (k) as input in PFCM algorithm **3** | **5** | **10** | **15** | **20** The final result of number of clusters in AL link set				
0.1	1	1	1	**2**	**2**
0.2	1	1	**2**	**2**	**2**
0.3	1	1	3	**2**	4
0.4	1	1	5	4	4
0.5	1	2	7	4	5
0.6	5	5	7	6	8
0.7	5	6	7	7	12
0.8	5	7	10	25	36
0.9	5	11	25	48	92

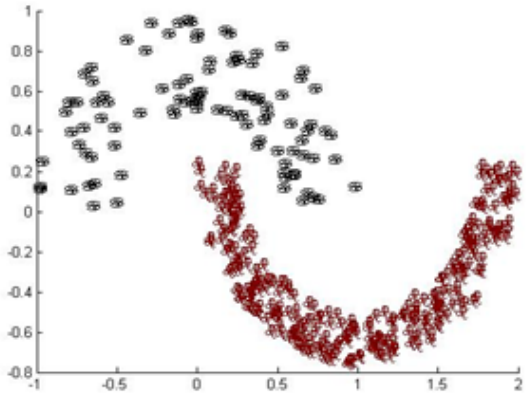

Fig. 4. Clustering outcome for Half ring synthetic dataset for optimal parameter pair values k = 15 and t = 0.2.

Motivated by the visual accuracy of the clustering outcomes (Figures 4, 5 and 6) and the impact of choices in parameter pair values (t and k) on the clustering outcome, we present, in the next subsection, results of our second experimental scenario wherein similar evaluations of the result of applying the proposed pEAC technique on real clinical datasets will be discussed.

4.2 Evaluation of the proposed technique using real clinical datasets

To further establish the efficacy of our proposed ensemble clustering technique, we employed clinical datasets that are commonly used in diagnosing a range of liver, cardiovascular, cancer and metabolic ailments. With this dataset,

Table 3. Results for Spiral synthetic dataset for variations in parameter values t and k

Threshold	The number of clusters (k) as input in PFCM algorithm **3** | **5** | **10** | **15** | **20** The final result of number of clusters in AL link set				
0.1	1	1	1	1	1
0.2	1	1	1	1	1
0.3	1	1	1	1	1
0.4	1	1	1	1	1
0.5	1	1	1	1	1
0.6	1	1	1	1	**2**
0.7	1	1	1	**2**	**2**
0.8	1	1	4	17	28
0.9	2	32	77	124	150

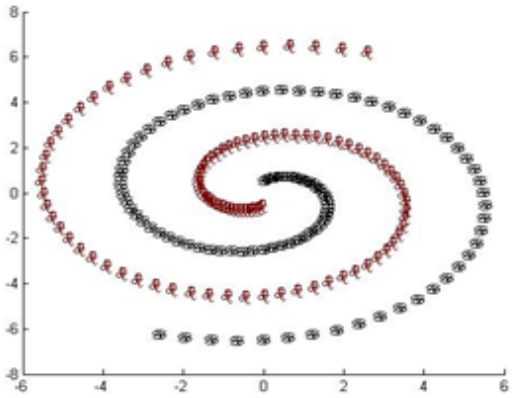

Fig. 5. Clustering outcome for Spiral synthetic dataset for optimal parameter pair values k = 15 and t = 0.7.

we further assessed the effectiveness of our proposed technique along two scenarios – i.e. performance evaluation alongside comparisons with unsupervised (standard) clustering methods and also with supervised (machine learning) models. The outcomes are presented in this and the next subsections.

The dataset used in our evaluation were sourced via the UCI Machine Learning Repository [14] and it consists of the Hepatitis dataset, which contains 19 attributes from 155 individuals out of which 32 indicate 'dead' and 123 indicate 'alive' classes; the heart-stat log dataset is used for the cardiovascular ailments and it consists of 13 attributes from 270 individuals out of which 120 indicate 'presence' of a heart disease, while 150 indicate 'absence' of heart disease, i.e. healthy subjects; two sets of data are used for breast cancer. The first, the Breast cancer dataset is made up of 9 attributes from 286 individuals out of which

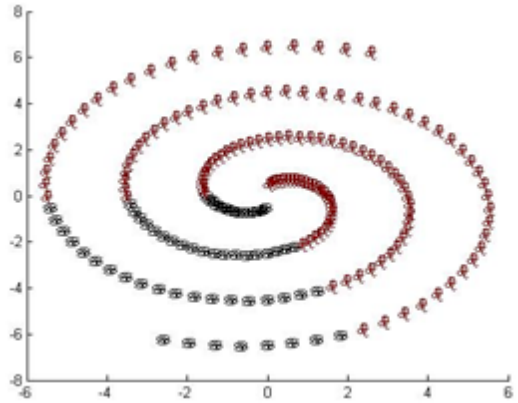

Fig. 6. Clustering outcome for Spiral synthetic dataset for non-optimal parameter pair values k = 3 and t = 0.9.

Table 4. Description of medical datasets

Dataset Name	**Data size**	**Number of Attributes**	**Class labels**
Hepatitis	155	19	Dead Alive
Heart-statlog	270	13	Absent Present
Breast cancer	286	9	No occurrence Occurrence
Wisconsin Breast cancer	699	9	Benign Malignant
Diabetes	768	8	Negative Positive

201 indicate 'no-recurrence' and 85 indicate 'recurrence' of cancer, while the second breast cancer dataset, the Wisconsin Breast-cancer dataset, comprises of 699 data points represented by 9 attributes, with two class labels indicating the tissue type, i.e. 'benign' (458) and 'malignant' (241); and finally, the Diabetes dataset, which has 768 data points represented by 8 attributes out of which 500 indicate 'negative' presence (i.e. absence) of diabetes and 268 are labeled as 'positive' to indicate the presence of diabetes. This information about the composition of the clinical datasets is further summarised in Table 4.

As mentioned earlier, the second experimental scenario for benchmarking the performance of our proposed pEAC ensemble clustering technique consists of two segments that are undertaken by comparing results from our proposed technique with those from other unsupervised learning techniques (first segment) and those from supervised machine learning models (second segment).

In the latter case, we used the "arff" data format of the Weka application

(http://repository.seasr.org/Datasets/UCI/arff/) for the Naive Bayes (NB), J48 (i.e. Decision trees with C4.5 algorithm in the Weka Application) and the Support Vector Machine with Sequential Minimal Optimisation (SMO) machine learning models. The Hepatitis, Breast cancer, and Wisconsin Breast cancer dataset had a few missing values, which were compensated for using the in-built mean or modus operation from the imputation process of the Weka application employed in our experiments. We used this feature to obtain the missing values prior to using of the proposed pEAC ensemble clustering technique.

Results of evaluations from the two experimental scenarios are presented and discussed in the remainder of this section.

4.3 Experimental results assessing the proposed pEAC Ensemble Clustering Technique versus Unsupervised Learning Methods

We used the 10-folds cross-validations to ascertain the performance of our proposed system alongside other unsupervised clustering methods in terms of (1) the number of correct diagnosis relative to existing labels in the datasets, for which we obtain the centroid of clusters, and (2) in terms of predictions of the unknown labels.

Table 5 presents the results obtained when our proposed ensemble clustering technique is compared alongside single (individual) runs of three unsupervised clustering algorithms (i.e. KM, FCM and PFCM) and the two standard unsupervised ensemble methods (i.e. hEAC and fEAC techniques)) for the five types of medical datasets. As seen from the table, our pEAC ensemble technique is more robust (better results) than individual runs of the clustering methods and the other unsupervised ensemble clustering algorithms for all the medical datasets that were used.

Additionally, as presented in Table 6, the speed up offered by our proposed pEAC ensemble clustering technique is seen to match that offered by the other unsupervised ensemble clustering methods that were employed in the experiment.

4.4 Experimental results assessing the proposed pEAC Ensemble Clustering Technique versus Supervised Machine Learning Models

Tables 7, 8, and 9 summarise the results of the performance of our proposed pEAC ensemble technique in comparison with the supervised learning methods in terms of accuracy of disease diagnosis and requirements in terms computational resources (speed and memory). In this experiment, for the J48 and SMO we used the default parameter settings in the Weka application. The parameter setting used for J48 (Decision Tree C4.5) method specifies a confidence

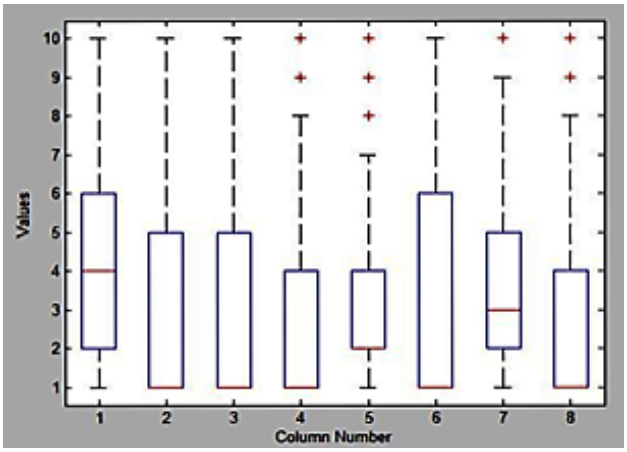

Fig. 7. Boxplot showing outlier detection in Wisconsin breast cancer dataset.

threshold for pruning set at the default value of 0.25, minimum number of instances per leaf set at default 2, using of binary splits only. The parameter settings used for SOM (SVM) method specifies *C* at default 1, the tolerance at default value of 1.0e-3, epsilon at default value of 1.0e-12, and setting the kernel is set as polynomial.

The results reported in Table 7 show that the supervised machine learning models (NB, J48, SMO) outperformed our proposed pEAC ensemble clustering approach in terms of diagnosing accuracy for the Hepatitis, Heart-statlog, and Diabetes clinical datasets. Remarkably, however, in the case of the two breast cancer datasets (i.e. the Breast cancer and Wisconsin Breast cancer) our proposed approach offers better disease diagnosing accuracy than the all the supervised learning techniques (results shown in bold) that were employed in the experiment.

A Box plot, which is a simple method used to graphically illustrate grouping in numerical data through their quartiles (by plotting outliers as individual points) is used to detect outliers in the Wisconsin breast cancer dataset. As seen in Figure 7, the Wisconsin breast cancer dataset has two outliers in the fourth column, three outliers in the fifth column, one outlier in the sixth column, and two outliers in the last column. The column (horizontal axis) represents the number of attributes (or features), while the vertical axis plots values of the attributes. However, as presented in Tables 8 and 9, the gains recorded in diagnosing accuracy by using the proposed pEAC ensemble clustering technique (for the two breast cancer datasets in Table 7) were realised at the expense of other computational resources that were used, particularly, speed (Table 8) and memory (Table 9).

It was shown in [31] and [32] that the respective memory requirements for the Naive Bayes and Decision Tree

Table 5. Comparison of accuracy rate in disease diagnosis (for training set) between the proposed pEAC ensemble clustering technique, individual clustering methods (KM, FCM, PFCM) and the standard unsupervised ensemble techniques (hEAC and fEAC) with 10-folds cross-validations

Dataset	**Diagnosing Accuracy (%)**					
Name	**KM**	**FCM**	**PFCM**	**hEAC**	**fEAC**	**pEAC**
Hepatitis	73.55	76.78	79.36	76.78	83.87	**84.51**
Heart-statlog	59.26	64.75	64.75	72.66	**76.66**	**76.66**
Breast cancer	74.48	74.48	75.52	75.52	77.79	**78.68**
Wisconsin Breast cancer	95.71	96.14	96.99	95.90	96.99	**98.39**
Diabetes	66.80	65.24	66.93	76.3	73.82	**77.34**

Table 6. Comparison of running time between the proposed pEAC ensemble clustering technique, individual clustering methods (KM, FCM, PFCM) and the standard unsupervised ensemble techniques (hEAC and fEAC)

Dataset	**Running time (s)**					
Name	**KM**	**FCM**	**PFCM**	**hEAC**	**fEAC**	**pEAC**
Hepatitis	0.03	0.03	0.03	0.87	0.87	0.87
Heart-statlog	0.03	0.03	0.03	0.90	0.90	0.90
Breast cancer	0.03	0.03	0.05	0.91	0.91	0.92
Wisconsin Breast cancer	0.04	0.04	0.07	1.50	1.50	1.51
Diabetes	0.08	0.08	0.15	1.90	1.90	1.95

Table 7. Comparison of accuracy rate in disease diagnosis between the proposed pEAC ensemble clustering technique and supervised machine learning models with 10-folds cross-validations (best results highlighted in bold)

Dataset	**Diagnosing Accuracy (%)**			
Name	**pEAC**	**Naive Bayes**	**J48**	**SMO**
Hepatitis	84.51	84.51	83.87	**85.16**
Heart-statlog	76.66	83.70	76.66	**84.07**
Breast cancer	**78.68**	71.67	75.52	69.58
Wisconsin Breast cancer	**98.39**	95.99	94.56	96.99
Diabetes	**77.34**	76.30	73.82	**77.34**

Table 8. Comparison of running time between the proposed pEAC and supervised machine learning models

Dataset	**Running Time (s)**				
Name	**Data size**	**pEAC**	**Naive Bayes**	**J48**	**SMO**
Hepatitis	155	0.87	0.01	0.02	0.03
Heart-statlog	270	0.90	0.01	0.02	0.03
Breast cancer	286	0.92	0.01	0.03	0.03
Wisconsin Breast cancer	699	1.51	0.01	0.04	0.04
Diabetes	768	1.95	0.01	0.07	0.08

Table 9. Comparison of space requirements between the proposed pEAC and supervised machine learning models

Medical Dataset Name	**Data size**	**No. of Attributes**	**Space complexity (bytes)**			
			hEAC	**Naive Bayes**	**J48**	**SMO**
Hepatitis	155	19	19,460	38	38	19,460
Heart-statlog	270	13	59,049	26	26	59,049
Breast cancer	286	9	66,254	18	18	66,254
Wisconsin Breast cancer	699	9	395,766	18	18	395,766
Diabetes	768	8	477,757	16	16	477,757

(C.45) methods is $O(fv)$ for systems containing f attributes (features) and v feature values. Similarly, the space (memory) complexity for the Support Vector Machine (SVM) was shown to be is $O(n^2)$ in [33], where n is number of training set size. Our results presented in Table 9 suggest that a similar quadratic space (of $O(n^2)$) is required to execute our proposed pEAC ensemble clustering technique for a training set of size n.

5 CONCLUDING REMARKS

Most of the available literature utilise supervised methods to diagnose diseases from a wide range of medical datasets. While these supervised methods are acceptable and in many instances effective, they have been associated with the high demands in terms of parameters such as learning rate, epoch, kernel and activation parameters for tuning in order to achieve best results [25]. These claims suggest that more resources are needed to obtain the optimal outcomes from different combinations of parameter values. Furthermore, supervised machine learning methods suffer from primacy associated with having one or two parameters for tuning. To overcome these shortcomings, we proposed a system that utilises an unsupervised learning method to make inferences regarding different ailments.

Our study proposed an Evidence Accumulation Clustering (EAC) technique that integrates the Possibilistic Fuzzy C-Means (PFCM) algorithm as its base cluster generator (pEAC) for the purpose of combining clustering partitions as needed to classify clinical data for disease diagnosis. Executed in three stages, in the first step of pEAC ensemble technique, the PFCM algorithm is used to split the dataset into smaller clusters. In the second step, relevant information in these clusters are co-associated and accumulated, so that the smaller clusters are combined. In the third step, the Active Link hierarchical agglomerative set is used to merge the smaller clusters that were obtained in earlier steps.

Two experimental scenarios were presented to validate the proposed ensemble clustering technique based on the use of the synthetic and real clinical datasets. The performance of the proposed technique was evaluated based on comparisons alongside two sets of unsupervised learning approaches: individual runs of well-known clustering algorithms (K-Means, FCM and PFCM) and other standard unsupervised ensemble clustering methods (hEAC and fEAC) as well as one set of selected supervised (machine learning) models (consisting up of the Naive Bayes, Decision Tree (J48), and Support Vector Machine (SVM) with Sequential Minimal Optimisation (SMO) methods).

Our results show that the proposed pEAC ensemble clustering technique outperforms the individual runs of the KM, FCM and PFCM algorithms as well as the standard (hEAC and fEAC) unsupervised ensemble techniques in terms of diagnosing accuracy and computational time (speed) for both the synthetic and real medical datasets that were used.

Remarkably, the comparison between our proposed pEAC ensemble technique and the selected supervised learning models revealed that, for tradeoffs in computational resources, the proposed technique offered gains in terms of better diagnosing accuracy for the two breast cancer datasets that were used in the experiments, thereby suggesting that the proposed ensemble clustering technique could offer efficient disease diagnosis for certain data types. Similarly, the computational requirements (speed and memory) to execute the proposed ensemble technique were able to match those offered by some of the supervised learning models that were used in the experiments.

The outcomes suggest that even at the cost of none labelling of data, for certain tradeoffs, our proposed unconventional technique could offer efficient medical data classification, thus indicating its potential applications in more advanced disease diagnosis.

ACKNOWLEDGEMENTS

This study is funded, in full, by the Prince Sattam Bin Abdulaziz University in Kingdom of Saudi Arabia via the

Deanship for Scientific Research Grant to The Computational Intelligence and Intelligent Systems (CIIS) Research Group Project Number 2016/01/6441.

[1] K. A. Abuhasel, A. M. Iliyasu, C. Fatichah, "A combined AdaBoost and NEWFM technique for medical data classification," Information Science and Applications, Lecture Notes in Electrical Engineering, Springer, Vol. 339, 2015, pp 801-809.

[2] S. Al-Muhaideb, et al., "Hybrid Metaheuristics for Medical Data Classification," Hybrid Metaheuristics, Studies in Computational Intelligence, Vol. 434, 2013, pp 187-217.

[3] M. C. Tu, et al., "Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms," Proc. of IEEE Int. Conf. on Dependable, Autonomic and Secure Computing, 12-14 Dec. 2009, Chengdu, China, pp. 183 - 187

[4] Oh S, et al., "Ensemble learning with active example selection for imbalanced biomedical data classification," Trans. Comput. Biol. Bioinform., Vol. 8, No. 2, 2011, pp. 316-25

[5] M. Seera, C. P. Lim, "A hybrid intelligent system for medical data classification," Expert Systems with Applications, Vol. 41, No. 5, 2014, pp. 2239–2249

[6] C. Fatichah, M. L. Tangel, et al., "Fuzzy feature representation for white blood cell differential counting in acute leukemia diagnosis", International Journal of Control, Automation and Systems,Vol. 13, No. 3, 2015, pp. 742-752.

[7] L. Guo, D. Rivero, J. Dorado, J. R. Rabunal, and A. Pazos," Automatic epileptic seizure detection in EEGs based on line length feature and artificial neural networks," Journal of Neuroscience Methods, Vol. 191, 2010, pp. 101–109.

[8] M. Sabetia, S.D. Katebi, R. Boostani, G.W. Price, "A new approach for EEG signal classification of schizophrenic and control participants," Expert Systems with Applications, Elsevier, Vol. 38, Issue 3, 2011, pp. 2063–2071.

[9] C. Fatichah, A. M. Iliyasu, K. A. Abuhasel, et al., "Principal Component Analysis-based Neural Network with Fuzzy Membership Function For Epileptic Seizure Detection," Proc. of 10th Int. Conf. on Natural Comput., Xiamen, China, 19-21 August 2014, pp. 186-191.

[10] J. P. Betancourt, et al., "Similarity-based fuzzy classification of ECG and capnogram signals", Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol. 17, No. 2, 2013, pp. 302-310.

[11] K. A. Abuhasel, A. M. Iliyasu, C. Fatichah, "A Hybrid PSO and Neural Network with Fuzzy Membership Function Technique for Epileptic Seizure Classification", Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol. 19, No. 3, 2015, pp. 447-455.

[12] M. I. Lopez, J. M. Luna, C. Romero, and S. Ventura, "Classification via Clustering for Predicting Final Marks Based on Student Participation in Forums," Proceeding of The 5thInternational Educational Data Mining Society, 2012, pp. 148-151.

[13] C. H. Cheng, J. W. Wang, and M. C. Wu, "OWA-weighted based clustering method for classification problem," Expert Systems with Applications, Vol. 36, No. 3, 2009, pp. 4988-4995.

[14] UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets.html

[15] X., Jin, & J. Kim, "Video fragment format classification using optimized discriminative subspace clustering," Signal Processing: Image Communication, 40, 2016, pp. 26-35.

[16] X. Tong, S. Zeng, N. Shang, L. Zeng, "Hand-Written numeral recognition based on fuzzy C-Means algorithm," Proceeding of The 10th International Symposium on Distributed Computing and Applications to Business, Engineering, Science, 2010, pp. 528-532.

[17] R. Evans, B. Pfahringer, and G. Holmes, "Clustering for classification," Proceeding of The 7th International Conference on Information Technology in Asia (CITA 11), 12-13 July 2011, pp. 1-8.

[18] R. Ghaemi, Md. N. Sulaiman, H. Ibrahim, N. Mustapha, "A Survey: Clustering Ensembles Techniques," Proceedings of World Academy of Science, Engineering, and Technology, Vol. 38, February 2009, pp. 644-653.

[19] A. Topchy, A. K. Jain, and W. Punch, "Clustering Ensembles: Models of Consensus and Weak Partitions,"IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 27(12), 2005, pp. 1866-1881.

[20] A. Strehl and J. Ghosh, "Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions," J. Machine Learning Research, Vol. 3, 2002, pp. 583-617.

[21] A.L.N. Fred, A. K. Jain, "Data Clustering Using Evidence Accumulation," Proc. 16th International Conference Pattern Recognition, 2002, pp. 276- 280.

[22] A.L.N. Fred, A. K. Jain, "Combining Multiple Clusterings Using Evidence Accumulation," IEEE Transactions on Pattern Analysis And Machine Intelligence, Vol. 27, No. 6, 2005, pp. 835-849.

[23] A. L. N. Fred, A. K. Jain, "Evidence Accumulation Clustering based on the K-Means Algorithm," Proc. Structural, Syntactic, and Statistical Pattern Recognition, Joint IAPR Int'l Workshops SSPR 2002 and SPR 2002, T. Caelli, et al., eds., 2002, pp. 442-451.

[24] T. Wang, "Comparing hard and fuzzy c-means for evidence-accumulation clustering," IEEE Proc. of International Conference Fuzzy Systems, Korea, August 2009, pp. 468-473.

[25] R. Krishnapuram, J.M. Keller, "The possibilistic C-means algorithm: insights and recommendations", IEEE Trans on Fuzzy Systems, Vol. 4(3), 1996, pp. 385-393.

[26] N.R. Pal, K. Pal, J.M. Keller, and J.C. Bezdek, "A possibilistic fuzzy c-*mean*s *cluster*ing algorithm", *IEEE Transactions on Fuzzy Systems*, Vol. 13, No. 4, pp. 517-530, 2005.

[27] R. Nock and F. Nielsen, "On Weighting Clustering," IEEE Trans. on Pattern Analysis and Machine Intelligence, 28 (8), 2006, pp. 1–13.

[28] R. Campello, E. R. Hruschka, V. S. Alves, "On the efficiency of evolutionary fuzzy clustering", J Heuristics, Vol. 15, pp. 43–75, 2009

[29] R. Krishnapuram and J. Keller, "A possibilistic approach to clustering," IEEE Trans. Fuzzy Syst., vol. 1, no. 2, pp. 98–110, Apr. 1993.

[30] K. A. Abuhasel, C. Fatichah, A. M. Iliyasu, "A Bi-Stage Technique for segmenting Cervical Smear Images Using Possibilistic Fuzzy C-Means and Mathematical Morphology", J. of Medical Imaging and Health Info., 6(8), 2016, pp. 1-7.

[31] C. Elkan, "Boosting and Naive Bayesian Learning," Technical Report No. CS97-557, September 1997.

[32] R. Kohavi, M. Sahami, "Error-Based and Entropy-Based Discretization of Continuous Features," Proceedings of The Second International Conference on Knowledge Discovery and Data Mining (KDD), 1996.

[33] I. W. Tsang, J. T. Kwok, P. M. Cheung, "Core Vector Machines: Fast SVM Training on Very Large Data Sets," Journal of Machine Learning Research 6, pp. 363–392, 2005.

Abdullah M. Iliyasu (aka Abdul M. Elias) obtained his ME, PhD and Dr Eng degrees in Computational Intelligence and Systems Science from the Tokyo Institute of Technology in Japan. Presently, he is the Principal Investigator and Team Leader of the Computational Intelligence and Intelligent Systems (CIIS) Research Group at the College of Engineering, Prince Sattam Bin Abdulaziz University in the Kingdom of Saudi Arabia. He has to his credit more than 60 publications traversing the areas of Computational Intelligence, Quantum Image Processing, Cyber and Information Security, Hybrid Intelligent Systems, Health Informatics and Electronics Systems Reliability.

Chastine Fatichah received her Ph.D from Tokyo Institute of Technology, Japan in 2012. She is currently a faculty member with the Institut Teknologi Sepuluh Nopember in Surabaya, Indonesia. Her research interests include Medical Image Processing, Computational Intelligence, and Data Mining.

Khaled A. Abuhasel obtained his BSEET and MSIE degrees from the University of Central Florida (Orlando, USA) and his PhD degree from the New Mexico State University (Las cruces, USA) specialising in Industrial Engineering. His research interests include Optimisation, Systems Engineering, Healthcare Systems, and Statistical Analysis. He is currently an Assistant Professor with the Department of Mechanical Engineering, Bisha University in the Kingdom Saudi Arabia.

AUTHORS' ADDRESSES

Abdullah M. Iliyasu,
Computational Intelligence & Intelligent Systems (CIIS) Research Group, College of Engineering,
Prince Sattam Bin Abdulaziz University,
Al-Kharj 11942,
Kingdom of Saudi Arabia.
Email: a.iliyasu@psau.edu.sa
Telephone: +966-115-888-8259

Chastine Fatichah
Department of Informatics,
Institut Teknologi Sepuluh Nopember,
Kampus ITS Sukolilo, Surabaya 60111, Indonesia.
Email: chastine@cs.its.ac.id

Khaled A. Abuhasel,
Department of Mechanical Engineering,
Bisha University,
Bisha 61361,
Kingdom of Saudi Arabia.

Received: 2015-08-01
Accepted: 2016-10-18