Appl Intell (2011) 34: 258–278
DOI 10.1007/s10489-009-0194-7

Hybrid ensemble approach for classification

Brijesh Verma · Syed Zahid Hassan

Published online: 19 September 2009

Abstract This paper presents a novel hybrid ensemble approach for classification in medical databases. The proposed approach is formulated to cluster extracted features from medical databases into soft clusters using unsupervised learning strategies and fuse the decisions using parallel data fusion techniques. The idea is to observe associations in the features and fuse the decisions made by learning algorithms to find the strong clusters which can make impact on overall classification accuracy. The novel techniques such as parallel neural-based strong clusters fusion and parallel neural network based data fusion are proposed that allow integration of various clustering algorithms for hybrid ensemble approach. The proposed approach has been implemented and evaluated on the benchmark databases such as Digital Database for Screening Mammograms, Wisconsin Breast Cancer, and Pima Indian Diabetics. A comparative performance analysis of the proposed approach with other existing approaches for knowledge extraction and classification is presented. The experimental results demonstrate the effectiveness of the proposed approach in terms of improved classification accuracy on benchmark medical databases.

Keywords Classifiers · Ensembles · Hybrid systems · Neural networks · Medical data classification

1 Introduction

In the last few decades, medical disciplines have become increasingly data-intensive. The advances in digital technology have led to an unprecedented growth in the size, complexity, and quantity of collected data—medical reports and associated images. According to Damien McAullay [1], "there are 5.7 million hospitals admissions, 210 million doctor's visits, and a similar number of prescribed medicines dispensed in Australia annually". All records are captured electronically. There are billions of medical records transaction occurrences world wide every year.

On the other hand, patient-centred medical applications (electronic patient records, personal health record, electronic medical records, etc.) are also on the verge of becoming practical, further increasing data growth and leading to a data-rich but information-poor healthcare system. Thus, it has become crucial for researchers to investigate and propose a novel approach that can appropriately utilize such valuable data to provide useful evidence as a basis for future medical practice. The paramount important factor is to utilize the collected data that suit specific and useful purposes which leads to enable the discovery of new 'knowledge' that provides insights to assists healthcare analyst and policy makers to make strategic decisions and predict future consequences by taking into account the actual outcomes of current operative values. In addition, the World Health Organization [2] identifies some possible needs for the discovery of knowledge from medical data repositories; this includes, but is not limited to, medical diagnosis and prognosis, patient health planning and development, healthcare system monitoring and evaluation, health planning and resource allocation, hospital and health services management, epidemiological and clinical research, and disease prevention.

Lately, this abundance of healthcare data has resulted in a large number of concerted efforts to inductively discover 'useful' knowledge from the collected data, and indeed interesting results have been reported by many researchers. However, despite the noted efficacy of knowledge discovery

B. Verma (✉) · S.Z. Hassan
School of Computing Sciences, CQUniversity, Rockhampton, QLD 4702, Australia
e-mail: b.verma@cqu.edu.au

methods, the challenge facing healthcare practitioners today is about data usability and impact—i.e. the use of 'appropriate' clustering algorithms with the right data to discover value-added 'action-oriented' knowledge in terms of data-mediated decision-support services.

Notably, recent advances in areas such as neural networks, evolutionary algorithms, statistical modelling and visualization tools have made it possible to transform any kind of raw data into high level knowledge. Neural networks are tools which can learn from unknown complex data and predict or classify new data which they have never seen before (e.g. data taken for cancer diagnosis). Neural networks have many characteristics which can significantly improve data classification algorithms. Neural networks can be defined as algorithms that (i) extracts rules from raw data, (ii) creates knowledge base by learning and adapting from raw data, and (iii) fuses/combines data/decisions from different sources. Neural networks are capable of learning from raw data using supervised and unsupervised learning. Once neural networks are trained, they are able to generalise new data and help in decision making such as prediction and classification. Evolutionary algorithms are used as data optimisation, selection of most or least significant part of data, and extraction of significant features from raw data.

The main problem with data classification methods is that each method has its own approach to deal with data structure, shape, and validity. This limitation affects the performance of classification systems. To overcome the limitations of traditional data classification algorithms, the need of a combination of diverse algorithms has widely been recognized [3, 4]. The numbers of hybrid clustering endeavours have been initiated all over the globe. The notion of clustering ensemble has extensively reported in the literature [5, 6]. The clustering ensemble incorporates a set of algorithms, whereby the algorithms decisions are typically combined by weighted/unweighted voting to discover new clusters [7]. The limitations associated with many existing hybrid approaches are: (i) the fusion of various classifiers is done based on simple majority of decisions instead of fusion of confidences from each method/classifier, (ii) the clustering is done as two cluster problem such as disease or no disease instead of sub-clusters within disease and no disease, (iii) the classification is done as two class learning problem instead of multi-class, and (vi) the performance is not consistent and sometime difficult to explain in medical domain so the decision is questioned by medical practitioners.

The aim of this research is to present and investigate a novel hybrid ensemble classification approach which is an effective combination of various clustering methods, in order to utilize the strengths of each individual technique and compensate for each other's weaknesses. More specifically, the proposed clustering strategy is formulated to cluster extracted features into 'soft' clusters using unsupervised learning strategies and fuse the cluster decisions using parallel fusion in conjunction with a neural classifier. In summary, this research poses the three basic, yet challenging, questions: why use hybrid and not standard individual data clustering/classification approaches? What are the main characteristics that a hybrid intelligent system should have in order to become the method of choice for a given application? What is the most efficient clustering strategy to discover strong clusters from various groups of clusters partitions?

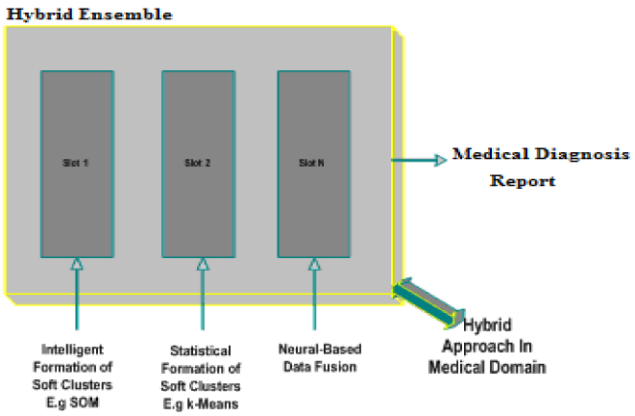

Fig. 1 Shows the context of hybrid ensemble in medical domain

To understand the context in which our approach is presented, the framework is viewed as supporting a large initiative to deliver strategic hybrid data classification/prediction services or ensemble (see Fig. 1). The ensembler contains various clustering tasks and methods to provide specialized data driven services. Proposed clustering strategy helps to populate the ensembler with the relevant clustering methods and specific data, coupled with the ability to produce soft clusters autonomously for neural-based data fusion. For evaluation purposes, this research focuses on medical domain.

2 Related work

The statistical, intelligent and hybrid algorithms have been used for data mining/classification/prediction in medical domain. The unsupervised learning algorithms have drawn prominent attention in medical data classification due to the nature of its problem domain, where the databases consist of complex, large and unlabelled data samples. Application of clustering ensemble techniques have started to emerge in several application domains, such as medical diagnostics [8], image classifications [9], document clustering [10], etc. Notably, the structure of medical data repositories, which consist of complex, large and unlabelled data samples, seems to be a good candidate for unsupervised learning algorithms. The unsupervised learning algorithms such as self-organizing map (SOM), k-Means, K-NN, have been

reported in various literatures ranges from feature selection, extraction, classification to data visualization. For example, self-organizing map (SOM) is used to identify the clusters in breast cancer diagnosis [11], to predict biopsy outcomes [12] and to model selection of mammography features [13]. It is demonstrated in [13] that how clustering algorithms can make impact on gene functions discovering and find the sample tissues to identify the causes of cancer.

Approaches to combine clustering algorithms differ in two main esteems, the way in which clustering algorithms are combined and the way contributing component clusters are obtained. Clustering ensemble approaches such as bagging and boosting have been proved as effective methods to improve the learning accuracy of the hybrid data mining systems [14]. A combination of data partitions obtained from multiple bootstrap algorithms are presented in [15, 16]. Both works used Hungarian algorithm to label each bootstrap sample by means of a single reference partition. The reference partition is set by a clustering of the entire data set [17]. The voting techniques is applied to determine strong clusters—the patterns which are closely related to each other. In [18], the final clusters, decision of multiple k-Means are defined by using bagging approach rather than by explicit labelling. In this approach, the bagging is achieved by grouping k-Means centers and assigning the clusters to the closest group. The k-Means centers are "bagged" and clustered by a hierarchical procedure. The partitioned clusters do not keep information about the individual cluster labels but only information about cluster sample. In [51], a clustering algorithm is proposed which combined the advantages of fuzzy sets and rough sets.

The idea of presenting the results of various clustering algorithms in the form of decision or co-association matrix was introduced by Fred et al. [19]. The values in the co-association matrix represent the strength of association between attributes by analyzing how often each pattern of objects appears in the same cluster. To determine the final clusters (strong clusters), a majority-voting algorithm is applied to the co-association matrix. The majority-voting algorithm was implemented based on the occurrence of each pattern in a similarity matrix for the data items to find clusters by linking the objects whose co-association value exceeds a certain threshold. Further work by Fred et al. [20] followed the same methodology, but despite using a fixed threshold, they applied a hierarchical single-link clustering algorithm to the co-association matrix. The k-Means algorithm with various values of k and random initializations was finally used for generating the clusters partition. Kellam et al. [21] also combined clusters decisions through a type of co-association matrix. Nevertheless, this matrix is used only to find the clusters with the highest value of occurrence in co-association matrix. As a result, only a set of strong clusters is produced which may not inherit all the properties of initial objects.

Following the same principals, the cluster ensemble approach was initiated to combine the clustering results of multiple clustering algorithms to obtain better quality and robust clustering results [22]. Martin et al. [23] emphasised the need of hybrid data mining by reporting some interesting results on individual clustering algorithms. They monitored the inability of individual clustering algorithm while dealing with data sets which were diverse in nature. Two clustering algorithms: k-Means and single-link algorithms were considered, to find two Spirals and two Globular clusters. It was observed that none of the clustering algorithm was able to discover given three clusters.

Lately, the need of a combination of diverse clustering algorithms has widely been recognized. The numbers of hybrid clustering endeavours have been initiated all over the globe. The notion of clustering ensemble is extensively reported in the hybrid data mining literature [24, 25]. In clustering ensemble, a set of classifiers are incorporated by the ensembler, whereby individual classifier's decisions are typically combined by weighted/unweighted voting to discover new clusters.

Many clustering ensemble approaches are proposed and investigated on benchmark databases, it is hard to say which cluster ensemble is the best. Different clustering algorithms present different types of knowledge concerning the clustering criterion; most clustering criteria in various algorithms are compensative rather than competitive in data analysis. Researchers believe that an effective combination of several clustering algorithms is an important step to improve the clustering quality.

Xiahua and Illhoi [26] discussed some of the major issues of clustering ensembles designing—how to combine different clustering algorithms and unbiased their consensus results with regard to all the basics partitions. They further highlighted the difficulties as "the quality of a clustering combination algorithm cannot be evaluated as precisely as a combining classifier, and various clustering algorithms always produce results with large differences due to different clustering criteria", and emphasize on the new mechanism to combine the diverse clustering algorithms to obtain better clustering results.

Many ensemble approaches based on combining various learning algorithms including reinforcement learning, incremental learning, etc. are also proposed and investigated. Wiering and van Hasselt [52] investigated several ensemble approaches that combine multiple reinforcement learning algorithms in a single agent. They wanted to enhance learning speed and final performance by combining the chosen actions or action probabilities of different reinforcement learning algorithms.

Liu and Yao [53] presented a new cooperative ensemble learning system for designing neural network ensembles. The idea was to encourage different individual networks in

an ensemble to learn different parts or aspects of a training data so that the ensemble can learn the whole training data better. The individual networks are trained simultaneously rather than independently or sequentially. It can create negatively correlated neural networks using a correlation penalty term in the error function. Islam et al. [54] proposed two cooperative ensemble learning algorithms. The proposed algorithms use the negative correlation learning algorithm and train different neural networks in an ensemble. Bagging and boosting algorithms are used in NegBagg and NegBoost, respectively, to create different training sets for different neural networks in the ensemble.

Parikh and Polikar [55] introduced an ensemble of classifiers based on incremental learning for data fusion. Their approach sequentially generates an ensemble of classifiers that specifically seek the most discriminating information from each data set. They observed and documented that their approach for data fusion consistently outperforms a similarly configured ensemble classifier trained on any of the individual data sources across several applications. Hassan and Verma [56] introduced combination of clustering techniques using serial fusion and produced a good accuracy. Carpenter et al. [57–60] proposed ARTMAP systems which create input clusters through unsupervised learning and link them to the output patterns through an inter-ART map field using supervised learning.

3 Proposed hybrid ensemble classifier

This research presents a novel hybrid ensemble approach for classification. Hybrid ensemble approach can be defined as a process of combining various algorithms and techniques in such a way that it can utilize the strengths of each individual technique and compensate for each other's weaknesses. It is a multilayered process. The multilayered process in proposed approach consists of three separate techniques such as Self Organising Map (SOM), k-means and Multi-Layer Perceptron (MLP). First layer is to create soft clusters using intelligent clustering technique called SOM, second layer is to create soft clusters using statistical clustering technique called k-means and third layer is fusion of strong clusters using MLP technique. The proposed approach incorporates some novel ideas such as soft clusters and parallel neural fusion and these ideas are described below.

Soft clusters are defined as clusters within a class (e.g. disease) with different confidence. The creation of soft clusters is based on an idea that in a classification problem, each class can have more than one cluster called soft clusters. The incorporation of soft clusters' output values into the learning of neural network weights, might improve the learning process and the overall classification accuracy. Soft clustering idea is exact opposite to hard clustering which means one cluster per class. For example, all data can be clustered into two hard clusters such as disease and no disease. However, for soft clusters, all data will be clustered in many clusters within disease and within no disease classes.

Parallel neural fusion means that the output values of clusters from different clustering algorithms are fed simultaneously to MLP. It is done by fusing/combining all outputs together into one single vector and then feed to MLP as an input. Let $\{c_{11}, c_{12}, c_{13}\}$ be the output values from first clustering algorithm and $\{c_{21}, c_{22}, c_{23}\}$ be the output values from 2nd clustering algorithm. The parallel combination for creating an input to MLP is as follows.

$$\{c_{11}, c_{12}, c_{13}, c_{21}, c_{22}, c_{23}\}$$

The proposed approach, which is the amalgamation of self-organizing map, k-Means, and multilayer perceptron (MLP), clusters extracted features from medical repository into soft clusters using unsupervised learning strategies and combines the decisions in conjunction with a neural classifier. The idea is to observe associations in the features and fuse the decisions (made by learning algorithms) to find the strong clusters which can make impact on overall system performance. More specifically, the proposed technique incorporates a number of clustering algorithms (both intelligent and statistical), which varies in their methods of search and representation to ensure diversity in the errors of the learned models. The proposed clustering approach is formed on the basis of the following clustering hypothesis: data (features) that are relevant to same concept can be clustered together, since they tend to be more similar to each other than to non relevant data. More specifically, this hypothesis suggests that separation of relevant data from non relevant data with proper clustering algorithms. The notion of neural-based clusters fusion can be understood by its fusion hypothesis, which assumes that more similar data a cluster contains, the more reliable the cluster is for decision-making.

Particularly, the entire amount of data is introduced to each clustering algorithm presented in the ensemble. The soft clusters produced by each algorithm are recorded, combined and fused into MLP in parallel fashion. Specifically, the two types of hybrid combination are investigated in this research: Parallel Neural-based Clusters Fusion (PNCF) and Parallel Neural-based Strong Clusters Fusion (PNSCF), and as depicted in Figs. 2(a) and 2(b).

3.1 Parallel Neural-based Clusters Fusion (PNCF)

In PNCF approach, multilayer perceptron (MLP) is incorporated as a classifier with the unsupervised clustering algorithms. The MLP classifier learns with the soft clusters, generated by the clustering algorithms, and classifies them into appropriate classes, which can later be explored for further

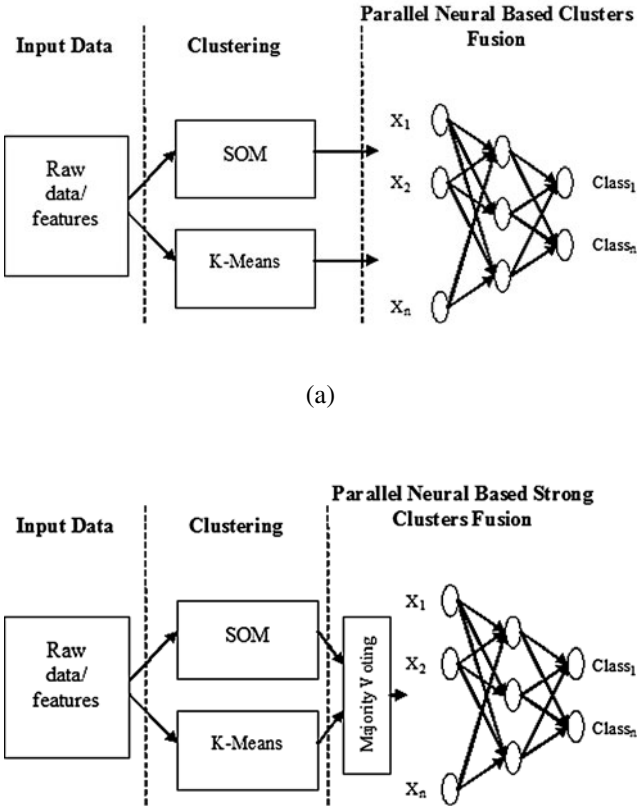

Fig. 2 (**a**) Parallel Neural-based Clusters Fusion (PNCF). (**b**) Parallel Neural-based Strong Clusters Fusion (PNSCF)

investigation and decision-making. The PNCF approach can be described as follow.

Step 1. Divide the data into training and test partitions D and D' respectively
Step 2. Initialize clustering algorithms A_i $(i = 1, \ldots, n)$ with training data D
Step 3. Cluster the partitions D into k-clusters (soft clusters) using A_i $(i = 1, \ldots, n)$
Step 4. Calculate the performance of A_i $(i = 1, \ldots, n)$ with both training and test partitions D, D' (*note that test partition is used only for evaluation purposes*)
Step 5. Combine the output of A_i $(i = 1, \ldots, n)$ in a parallel way (see 3rd paragraph in Sect. 3) and create new input vectors for MLP
Step 6. Train the MLP with input data from Step 5 and target outputs
Step 7. Calculate the performance of the PNCF using test partition.

3.2 Parallel Neural-based Strong Clusters Fusion (PNSCF)

In PNSCF approach, the ensemble of clustering algorithms is generated, whereby the output decisions of individual clustering algorithm are combined by a simple majority-voting scheme. Notably, the decisions are combined on both training and test data samples. In this majority-voting scheme, each algorithm assigns the confidence level to its generated output cluster based on the maximum cases that one cluster contains. This confidence can be considered as the 'weight' to a particular cluster. The cluster with higher confidence value considered as a strong cluster. The performance and decision of each individual clustering algorithm in the ensemble contributes in confidence consensus. Particularly, the PNSCF approach consists of following steps.

Step 1. Divide the data into training and test partitions D and D' respectively
Step 2. Initialize clustering algorithms A_i $(i = 1, \ldots, n)$ with training data D
Step 3. Cluster the partitions T into k-clusters (soft clusters) using A_i $(i = 1, \ldots, n)$
Step 4. Calculate the performance of A_i $(i = 1, \ldots, n)$ with both training and test partitions D, D', (*note that test partition is used only for evaluation purposes*)
Step 5. Find strong clusters by assigning confidence to each cluster generated by A_i $(i = 1, \ldots, n)$ based on majority-voting
Step 5. Combine the output of A_i $(i = 1, \ldots, n)$ in a parallel way (see 3rd paragraph in Sect. 3) and create new input vectors for MLP
Step 6. Train the MLP with input data from Step 5 and target outputs
Step 7. Calculate the performance of the PNSCF using test partition.

3.3 Theoretical underpinning of proposed ensemble approach

In this section, the proposed PNSCF and PNCF approaches are descried by taking mathematical underpinning into account to provide evidence of practice. Figure 3, illustrates that how input patterns are converted into high-dimensional groups of clusters (soft clusters) by using various unsupervised clustering criterions. To fuse all clusters and strong clusters in the generated soft clusters, two strategies are considered, PNSCF and PNCF, as described above.

To explicate the proposed methodology further, an unsupervised clustering-ensembler is formed by using self-organising map and k-Means clustering criterions. The

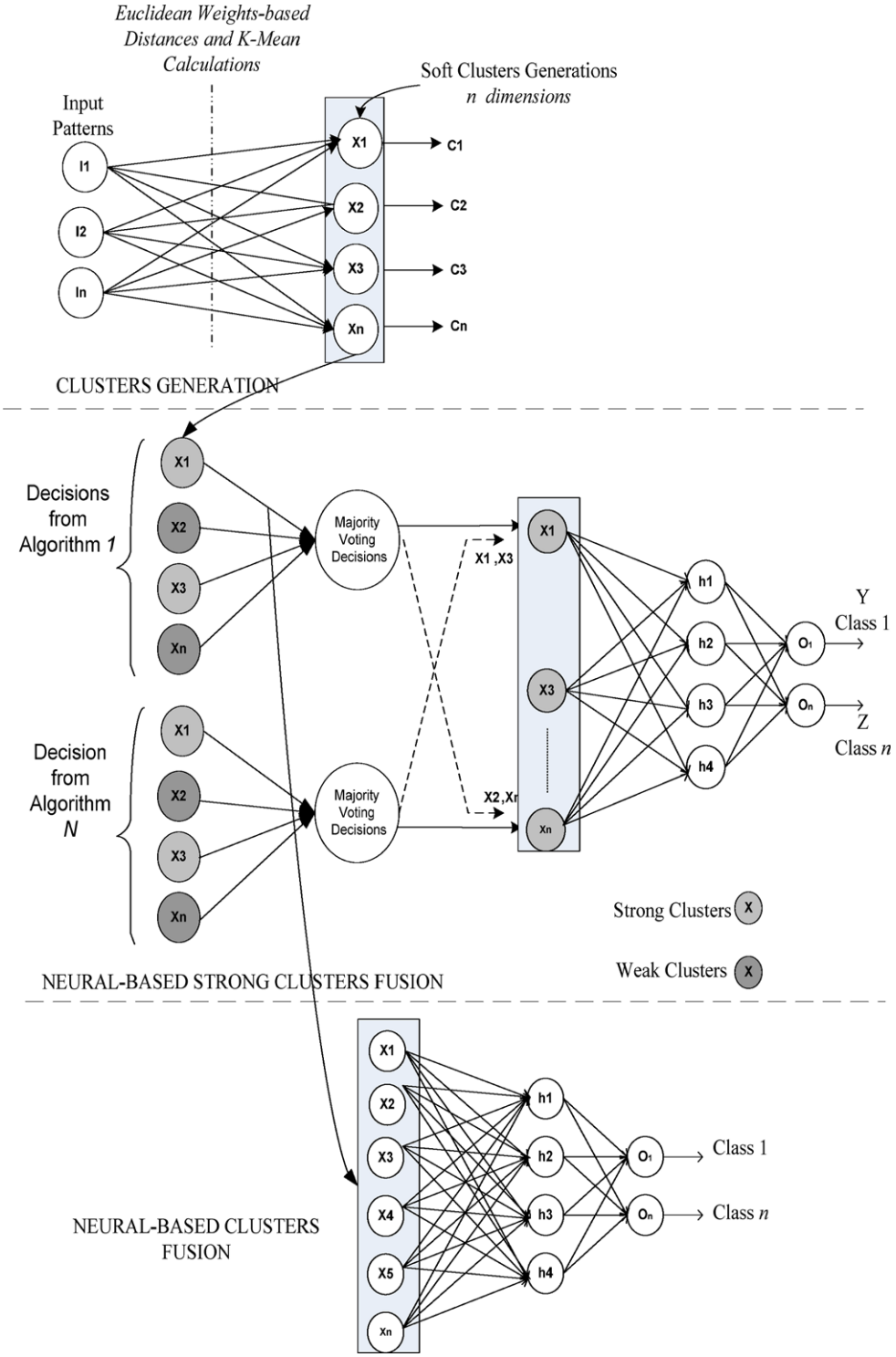

Fig. 3 Proposed data clustering strategy

proposed approach calculates the Euclidean distance and k-Means values in the input cases and generates various soft clusters.

Suppose given the input patterns: $P = \{I_1, I_2, \ldots, I_p\}$, where p is the number of input patterns, and n number of clustering algorithms A_i $(i = 1, \ldots, n)$ such that each algorithm A_i returns output clusters O_i of P which maximizes the confidence function associated with each individual cluster contained in cluster decisions. Formally, $O_i = \{X_1, \ldots, X_k, X_1, \ldots, X_k\}$

$$f_c(O_i(P)) = \max\{O_i(P)\} \quad (1)$$

where O_i represents collective output clusters or soft clusters generated by A_i clustering algorithms. X_k represents the value of kth cluster. Soft clusters generated by Algorithm 1

$$O_1 = \{X_1^1, \ldots, X_{k,}^1\}$$

Soft clusters generated by nth algorithm

$$O_n = \{X_1^n, \ldots, X_{k,}^n\}$$

The $f_c\{O_i(P)\}$ provides a maximum confidence given to an individual cluster, on the basis of clusters closest similarity with the input attributes, presented in each algorithm.

This research proposed the confidence function based on two clustering criterions: Euclidean Distance and k-Means, as shown in (2) and (3) respectively:

$$f_c(O_i(P)) = \max(\|P - W\|) \tag{2}$$

where W is the weight values assigned to the output units. Equation (2) can be further refined as,

$$f_c(O_i(P)) = \max((P_1 - W_1)^1 + (P_2 - W_2)^2 + \cdots + (P_n - W_n)^n) \tag{2.1}$$

The output unit which has the least Euclidean distance is considered as 'image' unit for the input pattern. Similarly, for the k-Mean criterion the confidence function is defined as

$$f_c(O_i(P)) = \max\left(\sum_{i=1}^{k} P - \mu_i\right) \tag{3}$$

where P is the vector space with k clusters of O_i, $i = 1, 2, \ldots, k$ and μ_i is the centroid or mean point of P.

If $T = \{T_i, \ldots, T_k\}$ is the target class (desired output) coupled with the input patterns, then the confidence function can be distinct as

$$f_c(O_i(P), T_{i-k}) \tag{4}$$

Functionally, to make decisions from a number of soft clusters using SOM criterion, the output units are designed almost twice the dimensions of input features spaces. The SOM was created which is consisted of 16 neurons partitioned in a single layer in a 2-D grid of 4×4 neurons. The random reference input vectors (neuron weights) are construed and assigned to each partition. For each input, the Euclidean distance between the input and each neuron was calculated. The reference vector with minimum distance is identified. After the most similar case is determined, all the neighborhood neurons, connected with the same link, adjust their weight with respect to the reference vector to form a group in two dimensional grids. The whole process is repeated several times, decreasing the amount of learning rate to increase the reference vector, until the convergence is achieved.

In k-Means criterion, as presented in (3), it randomly partitioned the input data into k-cluster centers along with its all closest features. With each input feature, it calculates the mean point of each feature and constructs a new partition by associating data-entities to one of the k clusters. Cluster features are moved iteratively between k clusters and intra-and-inter-cluster similarity. Distances are measured at each move. Features remained in the same cluster if they were closer to it otherwise move into new cluster. The centers for each cluster are recalculated after every move. The convergence achieved when moving object increased intra-cluster distances and decreases inter-cluster dissimilarity.

Let $x_1, \ldots, x_m$ are the soft clusters which are generated by applying proposed clustering criterions as depicted in (2) and (3). The clustering decisions made by SOM and k-Means criterions can be demonstrated by using decision matrix:

$$\begin{matrix} X_1 & \ldots & X_m \end{matrix}$$
$$\text{DecisionMatrix} = \begin{bmatrix} Y_{11} & \ldots & Y_{1m} \\ & \vdots & \\ Z_{n1} & \ldots & Z_{nm} \end{bmatrix} \tag{5}$$

where Y_{11}–Y_{1m} represents number of clusters formed for Class 1 and Z_{n1}–Z_{nm} represents number of clusters formed for Class m. In other words, a decision matrix represents the co-association between the classified patterns to the particular soft clusters. For instance, Y_{11}–Y_{1m} are the output clusters that represent cases which are classified as a Class 1, whereas Z_{n1}–Z_{nm} clusters represent cases which may belong to Class m. To identify the strong clusters in a decision matrix, which is generated from individual clustering algorithm (e.g. SOM), each column [e.g. Y_{11}, Z_{1m}] of the matrix is analyzed and maximum value and its associated class based on majority voting are marked. The cluster associated with analyzed column is called a strong cluster for associated class. The process is repeated for all columns in the decision matrix.

Once the strong clusters are identified, the outputs for individual clustering algorithms are generated and passed to neural network for fusion.

Lets C_n^i be the output of i clusters from n algorithms and it can be written as follows.

$C_1^i = \{x_1^i, x_1^i, \ldots, x_1^i\}$ the output from strong clusters generated by Algorithm 1

$C_n^i = \{x_n^i, x_n^i, \ldots, x_n^i\}$ the output from strong clusters generated by Algorithm n

The input for neural network is $I = (C_1^i, \ldots, C_n^i)$ and the output is target value.

A single layer multi-layer perceptron neural network with a back propagation training algorithm was designed and used for the clusters fusion.

The system produces the confidence value for all classes. If there are data with two classes such as disease and no disease then after feeding input, the outputs for disease and no disease classes are calculated and the system predicts the presence of disease or not based on output confidence value.

4 Experimental results

The performance of proposed hybrid ensemble approach was experimented and evaluated with various benchmark databases. The databases include the Digital Database for Screening Mammograms (DDSM) [27], Wisconsin Breast Cancer Database (WBCD) [28], and Pima Indians Diabetes Database (PIDD) [28]. The benchmark databases are obtained from the USF and UCI online Machine Learning Repositories.

4.1 Digital database for screening mammograms

Mammography can be used as a screening or diagnostic method for the early detection of breast cancer [29]. Breast cancers are visible on a mammogram as a mass, calcification or combination of both. A mammogram is a test that is done to look for any abnormalities in a woman's breasts [29]. The test uses an X-ray machine to take pictures of both the breasts. The abnormalities that a woman or a healthcare provider cannot feel during a physical examination can be found with mammograms. Breast lumps can be benign (non cancerous) or malignant (cancerous). Two types of mammography exams are in practice today: Screening and Diagnostic. The former technique is performed to detect breast cancer when it is too small to be felt by a physician or a patient. It is performed on women with no complaints or symptoms of breast cancer. The procedure involves taking X-ray images of breast. The later technique is performed on a patient who has been evaluated as symptomatic by a physical exam or screening mammography.

The database of digital mammograms is adopted from Digital Database for Screening Mammograms (DDSM) established by University of South Florida. DDSM is a collaborative effort of the Massachusetts General Hospital, the University of South Florida, and Sandia National Laboratories. The primary purpose of the database is to facilitate the research in the area of computer aided diagnosis of breast cancer [29]. This is the largest publicly available database and can be downloaded freely from USF's online digital mammograms database website: http://marathon.csee.usf.edu/Mammography/Database.html.

The DDSM database contains approximately 2,500 case studies, whereby each study includes two images of each breast, along with some associated patient information (age at time of study, breast density rating, subtlety rating for abnormalities, keyword description of abnormalities) and image information (scanner, spatial resolution etc). The experimental dataset of digital mammograms was formed using the cases of the DDSM benchmark database. The dataset comprised of mass types of breast abnormalities. Each mammogram contains more than one suspicious area. The dataset has a total of 200 suspicious areas (masses). The dataset was divided into a groups of total of 100 (50 malignant, and 50 benign) for training and 100 for testing. Table 1 shows the distribution of mass dataset used for experiment.

Table 1 DDSM for mass database

Feature vector #	Type of features used to form feature vector	Classes representation
1	Density	Malignant = 1
2	Mass shape	
3	Mass margin	Benign = 2
4	Abnormality assessment rank	
5	Patient age	
6	Subtlety value	

More specifically, a set of six features (measurements) has been utilized in this research. The features include: Density, Mass Shape, Mass Margin, Abnormality Assessment Rank, Patient Age, and Subtlety Value.

4.2 Wisconsin breast cancer database

Breast cancer is a malignant tumor that resides in the cells of the breast [30]. Researchers acknowledge a number of the threat factors which include gender, genetic risk factors, family history, menstrual periods, not having children, ageing, obesity that causes a possibility of a woman developing breast cancer. However, how some of these threat factors develop in cells and what exactly the causes of breast cancers are still an arguable debate. Research is in progress to understand this problem domain study more and significant results have been reported by many bioinformatics scientists. They have also investigated that how some changes in DNA can be capable of normal breast cells to become cancerous [31]. In recent years, the breast cancers cases have exponentially increased and reported in hospitals all over the globe. It was reported in [31], breast cancer was the second most diagnosed type of cancer. This challenging problem has managed to draw attention of many healthcare practitioners from doctors to data managers. There have number of intelligent and autonomous techniques have been proposed for diagnosis and improvements in treatment methods and as a result breast cancer results have enhanced throughout the last decade. The main factor in this movement is the early discovery and precise diagnosis of this disease.

Table 2 Wisconsin breast cancer database

Attribute number	Attribute description	Range	Mean
1	Clump thickness	1∼10	4.44
2	Uniformity of cell size	1∼10	3.15
3	Uniformity of cell shape	1∼10	3.22
4	Marginal adhesion	1∼10	2.83
5	Single epithelial cell size	1∼10	3.23
6	Bare nuclei	1∼10	3.54
7	Bland chromatin	1∼10	3.45
8	Normal nucleoli	1∼10	2.87
9	Mitoses	1∼10	1.6

$N = 683$ observations, 239 malignant and 444 benign

Table 3 Pima Indians Diabetes database

Attribute number	Attribute description	Mean
1	Number of times pregnant	3.8
2	Plasma glucose concentration test	12.9
3	Diastolic blood pressure (mmHG)	69.1
4	Triceps skin fold thickness (mm)	20.5
5	2-hour serum insulin (U/ml)	79.8
6	Body mass index (weight (kg)/height m)	32
7	Diabetes pedigree function	0.5
8	Patient age (years)	33.2

$N = 768$ observations, 268 diabetics and 500 non-diabetics

In this study, the Wisconsin breast cancer database was adopted for evolution purpose. This is available via UC1 machine learning repository. The database was attained from the University of Wisconsin Hospitals, originally provided by Madison from Dr. William H. Wolberg. The data consist of 683 records taken from patients' breasts. Each record in the database has 9 attributes. The 9 attributes detailed in Table 2 are graded on an interval scale from a normal state of 1 to 9 (most abnormal state).

These attributes measure the external appearance and internal chromosome changes in 9 different scales. There are two class variables of breast cancer: malignant (cancerous) and benign (non-cancerous), which is represented numerically by 1 and 2 respectively. There are 239 malignant cases and 444 benign cases. The objective is to classify between malignant and benign cases.

4.3 Pima indians diabetes database

Diabetes is a disease that involves problems with shortage or absence of hormone insulin secretion by the pancreas [32]. There are two kinds of diabetes depending on whether or not you need insulin, they are called: Type I Diabetes also known as insulin-dependent diabetes, and Type II Diabetes known as non-insulin-dependent diabetes [33]. Studies of type I diabetes have been limited to children and adults under the age of 30. There is little difference in age allocation between females and males which have evidently been verified in surveys in children and young adults. It affects both genders, nevertheless in several communities the majority with type II diabetes are female. Type II diabetes is more common in middle age, elderly and overweight people. Diabetes can be measured as a disorder of the metabolic disposal of food. The relation of food and diabetic state should be assessed from two aspects: first, whether food precipitates the diabetic condition and, second, the type of food that is suitable for the person with established diabetes, whether it is insulin dependent or non-insulin dependent. Family history of diabetes put family members at risk that can result from inheriting the disease or the sharing of a common environment by members of the same family. It is vital to find out if diabetes exists in families. Gestational diabetes can develop during pregnancy; it is associated with increased risk for developing diabetes in following years. Gestational diabetes usually goes away after pregnancy but, once women experiences gestational diabetes, the chances are high that it will return in the future and increases their chances of developing type II diabetes [33].

In this study, the Pima Indians Diabetes database was adopted for evaluation purposes. This is available from UC1 machine learning repository, originally owned by National Institute of Diabetes and Digestive and Kidney Diseases. The database consists of 768 instances and according to the examination results 268 of them are diabetics and the rest are non-diabetics. Each record has eight attributes and these are detailed in Table 3.

5 Experimental setup

The proposed Hybrid Data Mining Ensemble (HDME) approach, including all the serial and parallel hybrid combinations, was implemented by using the C++/Java languages and the MATLAB software package (version 7.0 with neural networks toolbox) on Windows platform. Separate programs that were written for individual tasks of the proposed technique are described below.

- Data collection and feature extraction process.
- Hybrid data mining ensemble for soft clusters and decision matrix generation.
- SOM clustering criterion to cluster data into various groups (soft clusters).

Table 4 Network parameters setup

Parameter	Range for SOM	Range for k-Means	Range for MLP
Learning rate	0.05∼1	0.05∼1	0.05 ∼ 1
Momentum	NA	NA	0 ∼ 1
Training iteration	500∼10000	500∼10000	500 ∼ 10000
Transfer function	Gaussian function	Gaussian function	Sigmoid (logistic)
Bias input	NA	NA	0.5 ∼ 1
Hidden layer	NA	NA	0.5 ∼ 1
Learning function	Linear decay	Linear decay	Static
Map height	6∼10	6∼10	NA
Map width	8∼10	8∼10	NA
Neighborhood function	Gaussian	Gaussian	NA
Neighborhood size	5∼10	5∼10	NA
Seed	1∼5	1∼5	NA
Topology	Hexagonal	Hexagonal	NA
Number of clusters	1∼100	1∼100	NA

- k-Means clustering criterion to cluster data into various groups (soft clusters).
- Confidence function for majority voting by assigning maximum confidence to the strong clusters.
- Multilayer perceptron (MLP) based neural fusion of clustered patterns obtained from SOM and k-Means.

All these programs were linked together using scripts to perform step-by-step execution. The datasets were formed and kept in Text and CSV files. The attributes of the benchmark databases; DDSM, WBCD, and PIDD are detailed in Tables 1, 2, and 3 respectively. These attributes were used as the inputs of the clustering algorithms. The key design decisions for the algorithms used in HDME are the architecture to produce soft clusters and generate decision matrix that reflects the decision of individual clustering algorithm on the given dataset. The adequate functioning of clustering algorithms depends on the sizes of the training set and test set. To comparatively evaluate the performance of the algorithms, all the algorithms presented in this study were trained by the same training data set and tested with the evaluation data set, 50% data used for training purposes and 50% for testing.

5.1 Network parameters

The network parameters which were used to form proposed ensemble are discussed below and summarized in Table 4.

- Bias Input: Used 1 as the bias constant input and 0 for no bias constant input for MLP.
- Hidden Layer: One hidden layer was used in MLP (back-propagation architecture) for learning purposes.
- Learning Rate: Learning rate was set between 0.05 and 1 (recommend 0.1).
- Momentum: Momentum factor between 0 to 1 was used to speed up calculations.
- Number of Clusters: Set the number of clusters K before data partitioning.
- Neighborhood Function: Gaussian Neighborhood function was used to achieve convergence.
- Neighborhood Size: Initial neighborhood size set to be the maps largest dimension.
- Seed: Random number generator seed in the range of 1 to 5 (whole numbers).
- Topology: Hexagonal map topology used to define a distance between the map units.
- Map Height: n dimensions of the map.
- Map Width: Map width set larger than the height.
- Training Iterations: Number of training iterations was set between few hundred to a few thousands.
- Transfer Function: Gaussian function for SOM and k-Means and Sigmoidal function were used in MLP to achieve convergence.

6 Performance evaluation measures

The choice of standard performance assessment measures enables appropriate evaluation of proposed technique with other existing techniques. The performance of the data mining algorithms can be determined by the computation of total classification accuracy, sensitivity, specificity, confusion matrix, and ROC curves [34]. Four standard evaluation methods such as measuring classification accuracy, sensitivity, specificity, and confusion matrix are used to measure the performance of the proposed HDME ensemble. The evaluation methods are defined as follow.

6.1 Classification accuracy

Classification accuracy is the most common evaluation technique to measure the performance of the classifiers. Most of the researchers used it as the main parameter of criteria to evaluate the performance of data mining system for classifying the medical databases. The higher the classification accuracy, the better the system is performing. The advantage of this measure lies in its simplicity; the disadvantage is that it can be deceptive. In the proposed technique the classification accuracy was calculated by using the following formula.

$$\text{Classification Accuracy} = \left[\frac{\text{Correct classified patterns}}{\text{Total number of patterns}}\right]$$

6.2 Sensitivity

The sensitivity measure approach was first introduced in medical domain to ensure the test ability of the classifier. It was calculated in the same way as the classification accuracy. However sensitivity regards only positive cases, for instance, it can be used to find patients with observed final diagnosis. In this study, sensitivity was computed as a number of true positive decisions over a number of actual positive cases. It can be represented as follow.

$$\text{Sensitivity} = \left[\frac{\text{TP}}{\text{TP} + \text{FN}}\ (\%)\right]$$

where, TP = True Positive cases and FN = False Negative cases.

6.3 Specificity

The specificity measure approach was also established in medical domain and computed in the same fashion as sensitivity. The difference is that it deals only with negative cases, for example patients without observed final diagnosis. The specificity can be calculated as the number of true negative decisions over number of actual negative cases. This is represented as follow.

$$\text{Specificity} = \left[\frac{\text{TN}}{\text{FP} + \text{TN}}\ (\%)\right]$$

where, TN = True Negative cases and FP = False Positive cases.

6.4 Confusion matrix

A confusion matrix represents information about actual and classified cases produced by a classification system. Performance of such a system is commonly evaluated by demonstrating the correct and incorrect patterns classification.

Table 5 Representation of confusion matrix

Actual	Predicted	
	Positive	Negative
Positive	X1	X2
Negative	Y1	Y2

The typical construction of the confusion matrix for the two classes is represented in Table 5. Where row (X1 and X2) represents the actual patterns and column (Y1 and Y2) represents the classified patterns for a class particular class. The difference between the actual patterns and the classified patterns is used to determine the performance of the proposed techniques.

7 Experiment results

In this section, the performance of individual clustering algorithm and the proposed Hybrid Ensemble techniques; PNCF and PNSCF, are evaluated on the above-discussed benchmark databases. The experimental results for DDSM are presented in Tables 6 and 7. The experimental results for WBCD are presented in Tables 8 and 9. The experimental results for PIDD are presented in Tables 10 and 11.

7.1 Confusion matrix based analysis

For the comparative analysis, this study draws the confusion matrix for the both investigated individual clustering algorithms and the proposed hybrid ensemble approach for all benchmark databases used in this study. The confusion matrices showing the classification results of the investigated approaches implemented for the task to classify between malignant and benign cases in the digital database for the screening of mammograms (DDSM) database is given in Table 12. The confusion matrices present in Table 13, showing the classification results of the investigated approaches implemented for the detection of breast cancer in the Wisconsin breast cancer database (WBCD). Table 14, represents the confusion matrices for the investigated approaches implemented for the classification of diabetes and non-diabetes patient in Pima Indians Diabetes database (PIDD).

To understand the concept of confusion matrix in this study, the Table 12 is further evaluated. The SOM algorithm correctly classified 88 cases out of all the cases presented to it from DDSM database. The row values (48, 2) are the actual cases for the class malignant, and row values (10, 40) represent the actual class benign. However, the classified outputs are represented by column (48, 10) and

Table 6 Performance on database for digital screening mammograms

Approaches	Classification error (%)	Classification accuracy (%)	Sensitivity	Specificity
SOM	12.00	88.00	97.00	90.00
k-Means	16.00	84.00	89.00	82.00
MLP	12.00	88.00	99.00	84.00
Proposed PNSCF	**0.00**	**100.00**	**100.00**	**100.00**
Proposed PNCF	6.00	94.00	98.00	95.00

Table 7 Detailed accuracy by classes: TP = True Positive, FP = False Positive, and F-Measure = Frequency Measure over Class Accuracy

Approaches	Classes	TP rate	FP rate	F-Measure
SOM	Benign (SOM)	0.81	0.04	0.87
	Malignant (SOM)	0.96	0.24	0.88
k-Means	Benign (k-Means)	0.94	0.14	0.84
	Malignant (k-Means)	0.86	0.06	0.83
MLP	Benign (MLP)	0.82	0.06	0.87
	Malignant (MLP)	0.94	0.18	0.88
Proposed PNSCF	Benign	0.00	0.00	1.00
	Malignant	0.00	0.00	1.00
Proposed PNCF	Benign	0.95	0.03	0.94
	Malignant	0.97	0.04	0.94

Table 8 Performance on Wisconsin breast cancer database

Approaches	Classification error (%)	Classification accuracy (%)	Sensitivity	Specificity
SOM	13.50	86.50	81.30	89.40
k-Means	10.10	89.90	94.20	85.50
MLP	12.50	87.50	91.75	95.00
Proposed PNSCF	**2.10**	**97.90**	**98.50**	**94.30**
Proposed PNCF	4.00	96.00	99.00	90.00

Table 9 Detailed accuracy by classes: TP = True Positive, FP = False Positive, and F-Measure = Frequency Measure over Class Accuracy

Approaches	Classes	TP rate	FP rate	F-Measure
SOM	Benign (SOM)	0.85	0.15	0.87
	Malignant (SOM)	0.87	0.26	0.84
k-Means	Benign (k-Means)	0.88	0.1	0.88
	Malignant (k-Means)	0.83	0.9	0.89
MLP	Benign (MLP)	0.87	0.21	0.86
	Malignant (MLP)	0.91	0.15	0.88
Proposed PNSCF	Benign	0.98	0.11	0.96
	Malignant	0.96	0.05	0.97
Proposed PNCF	Benign	0.99	0.02	0.96
	Malignant	0.98	0.04	0.94

column $(2, 40)$ for the classes malignant and benign respectively. The comparison of these rows and columns, between actual pattern and classified patterns, can provide interesting insights. For instance: for the malignant class accuracy, it is noticed that the original malignant patterns were $(48, 2)$ and the classifier indicates $(48, 10)$. Thus, it classified 48 cases correctly as a malignant class and misclassified 2 cases. It is also noticeable that those two patients

Table 10 Performance on Pima Indians Diabetes database

Approaches	Classification error (%)	Classification accuracy (%)	Sensitivity	Specificity
SOM	20.70	79.30	86.40	75.30
k-Means	22.42	77.58	81.70	78.90
MLP	14.70	85.30	79.50	83.20
Proposed PNSCF	**10.50**	**89.50**	**90.35**	**87.30**
Proposed PNCF	13.55	86.45	91.45	92.90

Table 11 Detailed accuracy by classes: TP = True Positive, FP = False Positive, and F-Measure = Frequency Measure over Class Accuracy

Approaches	Classes	TP rate	FP rate	F-Measure
SOM	Diabetes (SOM)	0.81	0.35	0.79
	Non-diabetes (SOM)	0.86	0.34	0.78
k-Means	Diabetes (k-Means)	0.8	0.41	0.76
	Non-diabetes (k-Means)	0.81	0.39	0.77
MLP	Diabetes (MLP)	0.85	0.2	0.85
	Non-diabetes (MLP)	0.91	0.24	0.84
Proposed PNSCF	Diabetes	0.93	0.15	0.90
	Non-diabetes	0.92	0.12	0.91
Proposed PNCF	Diabetes	0.89	0.19	0.85
	Non-diabetes	0.94	0.21	0.84

Table 12 Confusion matrices of the investigated approaches used for the classification of DDSM

Approaches	Desired result	Output result	
		Malignant	Benign
SOM	Malignant	48	2
	Benign	10	40
k-Means	Malignant	38	12
	Benign	4	46
MLP	Malignant	47	3
	Benign	9	41
PNSCF	Malignant	50	0
	Benign	0	50
PNCF	Malignant	48	2
	Benign	4	46

Table 13 Confusion matrices of the investigated approaches used for the classification of WBCD

Approaches	Desired result	Output result	
		Malignant	Benign
SOM	Malignant	100	20
	Benign	25	196
k-Means	Malignant	104	16
	Benign	18	203
MLP	Malignant	93	27
	Benign	16	205
PNSCF	Malignant	117	3
	Benign	4	217
PNCF	Malignant	111	9
	Benign	5	216

will be given clear when they were supposed to be treated like a cancer patients. Similarly, for the benign class accuracy, the actual cases are $(10, 40)$ and whereas the classifier indicates $(2, 40)$. The 40 cases were classified correctly as a class benign and 10 cases were misclassified. In this scenario, those 10 patients who are not the victim of cancers will be treated like a cancer patient despite it being the opposite scenario. However, the overall outcome is much more favourable: 48 classified correctly as a malignant class and 40 classified correctly as a benign class. Similarly, by constructing the confusion matrix method on the K-Means classifiers, it can be noticed that the 38 cases were classified correctly as a class malignant (12 cases were misclassified) and 46 cases classified correctly as a class benign (misclassified 4 cases), overall 84 cases were classified correctly. Individual MLP classified 47 and 41 cases correctly as a class malignant and benign respectively, with the ratio of 3 misclassified cases of a class malignant and 9 cases for a class benign, overall classified 88 cases correctly.

8 Comparative analysis

This section presents a comprehensive analysis of the results attained using the individual clustering algorithms and proposed hybrid ensemble techniques, PNSCF and PNCF, on benchmark databases such as DDSM, WBCD, and PIDD. A comparison of the best results achieved in this research with previously published results is made to evaluate the performance of the proposed approach. Finally a summary of the analysis and comparison is presented.

Table 14 Confusion matrices of the investigated approaches used for the classification of PIDD

Approaches	Desired result	Output result	
		Diabetes	Non-diabetes
SOM	Diabetes	99	35
	Non-diabetes	44	206
k-Means	Diabetes	89	45
	Non-diabetes	41	209
MLP	Diabetes	102	32
	Non-diabetes	24	226
PNSCF	Diabetes	116	18
	Non-diabetes	22	228
PNCF	Diabetes	110	24
	Non-diabetes	28	222

8.1 Comparative analysis with digital database for screening mammograms database

Performing comparative analysis with other existing approaches for digital mammograms classification was a challenging task. Most of the proposed approaches found in the literature were generally tested on digital mammograms but adopted different databases or same databases with different features compare to what used in this study. However, there were some approaches which were quite relevant to this research work demonstrated some interesting facts.

Panchal et al. [35] utilize an auto-associator-MLP based hybrid classifier on DDSM databases and they stated 90.9% accuracy on test dataset; however the training of auto-associator and MLP took much longer time than the proposed methodology. Verma [36] proposed a neural algorithm and tested on the DDSM mammograms and achieved 94% accuracy on test set, however more iterations were used. In [37] Mahmoud et al. proposed the approach for the classification of tumors (masses) in mammograms using two segments approach. In the first stage, they extracted mammography features by using a combination of morphological operations and a region growing technique. In the second phase, segmented regions are classified by using a NN as normal, benign, or malignant tissues based on different measurements (shape, intensity variation, spread pattern etc.). Experiments were performed on mammogram images of the DDSM database and 82.9% classification accuracy was claimed, as show in Fig. 4.

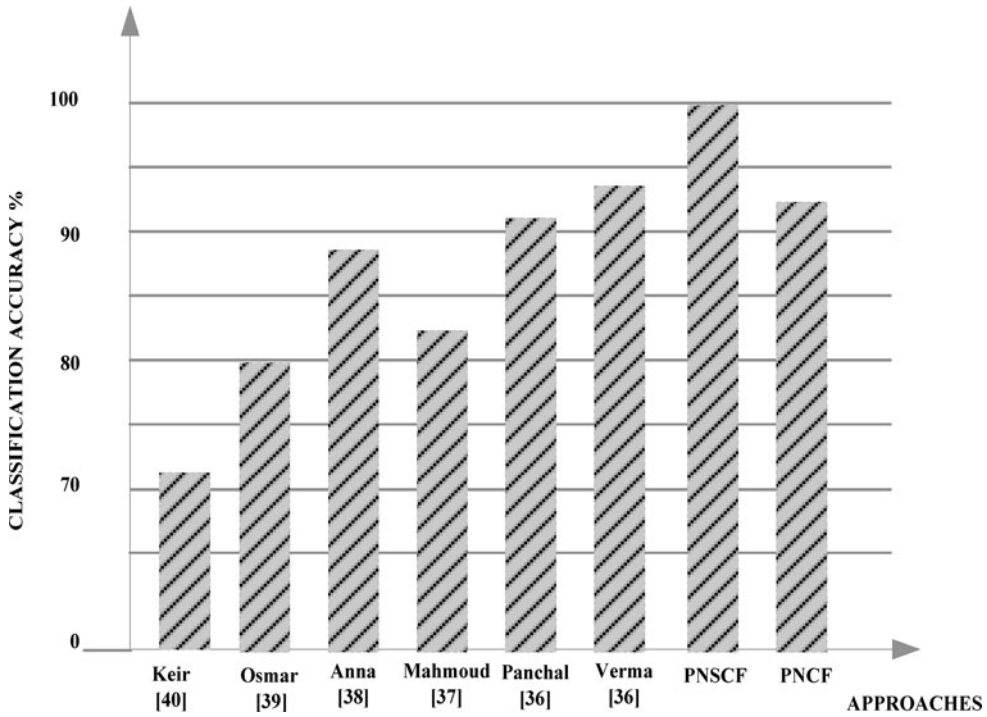

Fig. 4 Performance comparison of approaches with DDSM database

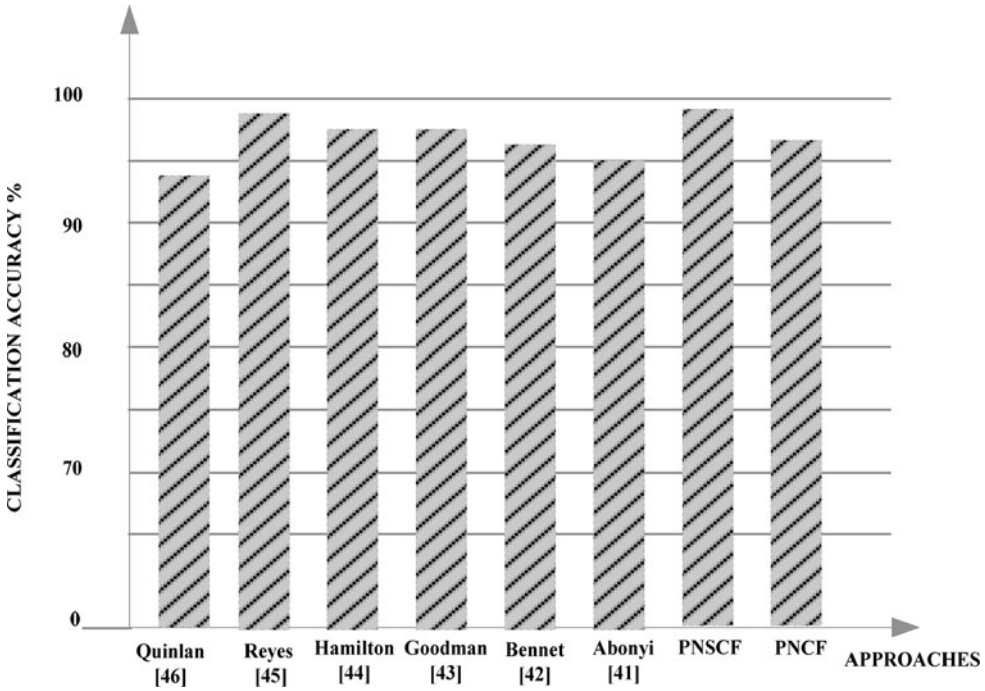

Fig. 5 Performance comparison of approaches with WBCD database

Anna et al. [38] investigated the texture properties of the tissue surrounding microcalcification and their contribution towards breast cancer diagnosis. They used K-NN approach to discriminate benign and malignant classes in digital mammograms. The Digital Database for Screening Mammograms (DDSM) was used, which consisted of 100 mammography's images. The overall classification accuracy demonstrated was 89%, as shown in Fig. 4. Osmar et al. [39] deployed an association rule-based classifier for mammography classification and managed to attain over 80% in accuracy. Keir et al. [40], proposed bootstrap aggregation (bagging) technique to extract features and used feed forward neural network to classify the mammography images, obtained by DDSM. The overall classification accuracy reported on four-classes problem was 71.4%, as shown in Fig. 4.

8.2 Comparative analysis with Wisconsin breast cancer database

When literature related to the classification of Wisconsin breast cancer database are examined, it can be seen that a great variety of approaches have been used and obtained high classification accuracies. Among these, Abonyi et al. [41] proposed a new fuzzy model structure where each rule can represent more than one classes with different probabilities. A supervised clustering algorithm was used identify the fuzzy model. The relevant input variables of the fuzzy classifier were selected based on the analysis of the clusters by Fisher's interclass separability criteria. This approach was applied to the Wisconsin breast cancer classification problems and obtained 95.57% accuracy, as shown in Fig. 5.

Bennet et al. [42] used support vector machine (SVM) to generalize the parameters in the decision tree (DT). By varying the kernel function used, a logically simple decision trees was created with multivariate linear, nonlinear or linear decisions. The preliminary results on WBCD database indicated that the hybrid combination of SVM and DT methods showed better performance by attaining 97.2% classification accuracy. Goodman et al. [43] used three diverse methods to the WBCD database classification problem which resulted in the following accuracies: the optimized LVQ method's performance was 96.7%, the big LVQ method reached 96.8% and the last method, AIRS, which was the modified version of the artificial immune system (replacing and maintaining of its memory cell population) reached 97.2% classification accuracy. Hamilton et al. [44] presented the RIAC (Rule Induction through Approximate Classification) algorithm, designed based on the theory of rough sets approximation technique. The purpose was to induce the rules from examples. Imprecise data are generalized using a rough-sets based. RIAC approach treats each generated classification rule as a piece of uncertain evidence, which by itself is of little value with respect to classification. Experimental results specify that RIAC method achieved 96% classification accuracy on WBCD database.

Pena-Reyes et al. [45] focuses on the combining two clustering algorithms: fuzzy systems and evolutionary algorithms, so as to automatically produce diagnostic systems. This fuzzy-genetic combination managed to achieve 97.51% classification accuracy on the WBCD database. Quinlan [46]

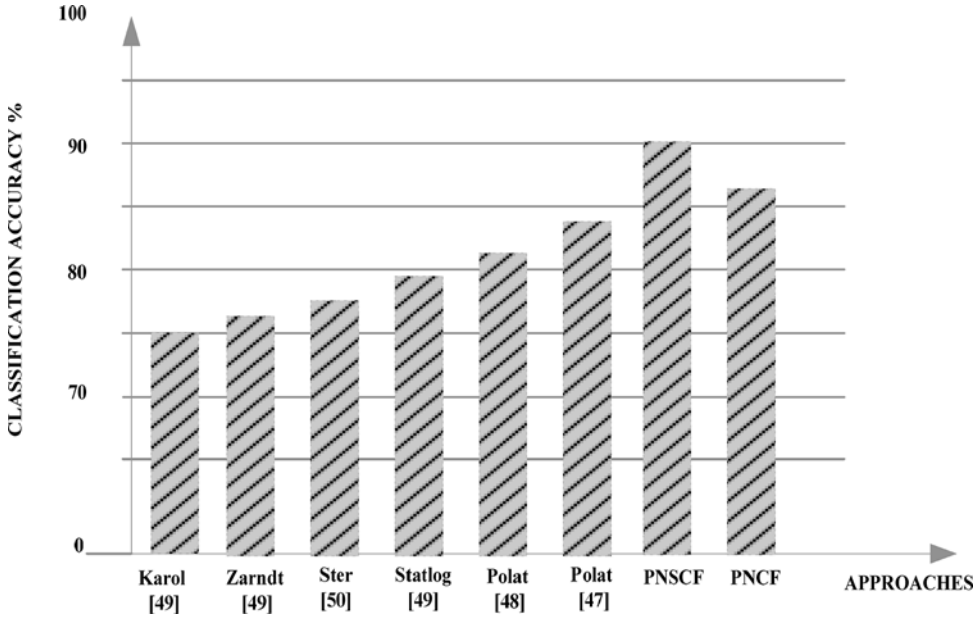

Fig. 6 Performance comparison of approaches with Pima Indian Diabetes database

proposed the modified version of C4.5 by modifying the formation and evaluation of tests on continuous attributes. They justified their approach by applying it to the WBCD classification problem and reported 94.74% classification accuracy.

8.3 Comparative analysis with Pima Indians Diabetes Database

There is a lot of work related to classification of the Pima Indians diabetes data set in the literature. Many researchers have demonstrated their success by achieving high prediction accuracy.

Polat et al. [47] proposed the updated version of artificial immune recognition system (AIRS). They introduced a fuzzy logic in AIRS system (Fuzzy-AIRS) to deal with medical classification problems. The highest classification accuracies obtained by applying the AIRS and Fuzzy-AIRS algorithms was respectively 79.22% and 84.42% for classification of the Pima Indian diabetes database, as shown in Fig. 6.

Polat et al. [48] used Generalized Discriminant Analysis (GDA) and Least Square Support Vector Machine (LS-SVM) for diagnosis of diabetes disease. The introduced a two-stage cascade learning system based on Generalized Discriminant Analysis and Least Square Support Vector Machine. The Generalized Discriminant Analysis stage was used to discriminant feature variables between healthy and patient (diabetes) data as pre-processing process. The combination of LS-SVM was used in latter stage in order to classification of diabetes dataset. The LS-SVM obtained 78.21% classification accuracy, while GDA–LS-SVM obtained 82.05% classification accuracy.

Statlog [49] reported 77.7%, 77.6%, 76.8% and 75.7% classification accuracies on Pima Indians diabetes database using Logdisc, DIPOL92, SMART and RBF methods. Ster et al. [50] achieved 77.5%, 76.6%, 76.5%, 76.4%, 75.8% and 75.8% using linear discriminant analysis, ASI, Fisher discriminant analysis, MLP-BP, LVQ and LFC techniques to classify between diabetes and non-diabetes patient in PIDD database. While Zarndt et al. [49] reported 75.8% classification accuracy by forming the combination of multi-layer perceptron and back propagation neural network (MLP-BP) by way of 10-fold cross-validation to classify PIDD database. Karol et al. [49] accomplished 75.5% classification accuracy using the k-nearest neighbor (k-NN) method by way of 10-fold cross-validation.

9 Discussion and summary

This section provides a discussion of the performance of the proposed hybrid ensemble techniques; PNSCF and PNCF, to extract meaningful knowledge from medical databases. This study used three benchmark databases consisting of DDSM, WBCD, and PIDD, in order to test the performance of the proposed techniques. The proposed clustering approaches have been compared with other existing individual and hybrid approaches with respect to classification accuracy, sensitivity and specificity values, and confusion matrix. The relation between classification accuracies, sensitivity and specificity values, and confusion matrix in proposed techniques for the classification of the DDSM database is shown in Tables 6 and 7, while the relation for the classification of the WBCD database is shown in Tables 8 and 9. The relation for the classification of the PIDD database is shown in Tables 10 and 11. As it can be seen in these Tables that proposed two hybrid approaches PNSCF and PNCF performed far more superior than the original classifiers.

The proposed parallel neural-based strong clusters fusion (PNSCF) technique has produced the outstanding results on all the benchmark databases by achieving 100% accuracy for DDSM database, almost 98% for WBCD database, and 89% for PIDD database. The technique of generating number of soft clusters (n-dimensional data) from the input features and finding the strong clusters, by using confidence function, made a great impact on the performance of PNSCF approach. In this case, PNSCF approach, dealt with only strong clusters by neglecting the weak clusters based on the classifiers decisions. When dealing with both strong and weak soft clusters and relying on MLP to make a final classification decision, PNCF approach has also proved its authenticity by achieving 94% accuracy for DDSM database, 96% for WBCD database, and 86% for PIDD database. This also justify the choice of MLP as a data fusion approach, as its strengths lie in its ability to deal with large amount noisy data and can generalize quite well to similar unseen data.

When comparing proposed approaches with other existing approaches on the DDSM benchmark database, the proposed hybrid data mining approaches, PNSCF and PNCF have demonstrated better performance compared to existing approaches by achieving 100% and 94% classification accuracies, as shown in Fig. 4. The proposed hybrid data mining approaches, PNSCF and PNCF have also demonstrated better performance compared to existing approaches by achieving 98% and 96% classification accuracies on WBCD database, as shown in Fig. 5. The proposed approaches, PNSCF and PNCF outperformed the existing approaches by achieving 90% and 86% classification accuracies on Pima Indians Diabetes database, as shown in Fig. 6.

There are many reasons for consistent superior performance by the proposed ensemble approach. The first reason is that the different clustering algorithms have captured different characteristics of data. The second reason is forming of strong clusters which put together strongly associated groups. There are a number of groups within a class so clustering into several clusters and taking strong clusters for each class allows learning process to learn different characteristics for same class. Multiple strong cluster based features provide more correlated and accurate input feature space. Third and final reason which is most important one is neural fusion. The neural network uses only strong clusters to learn which means it has good knowledge base for generalization.

When comparing time complexity of the proposed approach with other approaches, the proposed hybrid ensemble approach takes more time than any individual algorithm because it uses multiple clustering algorithms and neural fusion algorithms which require time for clustering and training processes. The time complexity is reduced by running all clustering algorithms simultaneously and training of neural network using fast learning algorithm. The time taken by neural fusion is very minimal.

In addition to accuracy and time, when comparing other things such as memory usage, the proposed approach requires slightly more memory to run the ensemble approach than any individual algorithm, however it does not require huge memory storage for storing outputs of clustering algorithms and neural weights.

This study also used the confusion matrix to test the proposed approaches such as PNSCF and PNCF, and including the individual clustering algorithms, for the all benchmark databases DDSM, WBCD, and PIDD used in this research. The confusion matrices for all approaches using all three databases are presented in Tables 12, 13, and 14 respectively. The results in confusion matrices clearly demonstrate the significant improvement in classification accuracies by the proposed techniques.

The proposed techniques, PNSCF and PNCF showed the significance of hybrid ensemble approach in medical databases. It was shown that the proposed techniques are promising techniques and perform comparatively with other previously proposed techniques. The PNSCF and PNCF techniques attained very good classification results with all benchmark databases. This demonstrates that different hybrid combinations need to be taken into account or considered when dealing with different types of data. Table 15 summarizes the classification accuracy of proposed and existing approaches on benchmark databases.

In summary, we would like to go back to our aims set at the beginning of this research. Hybrid ensemble technique has been proposed and evaluated on some complex medical benchmark databases. The research results presented in this paper confirms that hybrid techniques produce significantly better results than individual techniques so it is fair to say that we should use hybrid ensemble techniques because they produce better results. The main characteristics are the use of different type of clustering to unlabelled data and parallel fusion. Neural networks have shown that they are good candidates for fusion in hybrid intelligent system as they have achieved better accuracy than non-neural fusion investigated in this research. A number of clustering algorithms were reviewed and investigated in this research. The k-means and SOM which are different clustering strategies have shown that they are very efficient in creating soft clusters and using them in proposed hybrid system.

The research in this study has focused on medical databases and it has evaluated the proposed approaches on three well known benchmark databases. However the proposed approaches are generic in nature and can be well fitted in any problem domain which has training data in form of feature values available. The approaches have been evaluated on small size of input features, however there is nothing in the approaches which can stop using large number of input features. The major insight this work brought forward is

Table 15 Summary of classification accuracies using proposed and existing approaches

Author	Method	Classification accuracy (%)		
		DDSM	WBCD	PIDD
Mahmoud et al. [37]	Two-Segments Approach	82.90	NA	NA
Anna et al. [38]	k-NN	89.00	NA	NA
Osmar et al. [39]	Association Rules	80.00	NA	NA
Keir et al. [40]	Bootstrap based MLP	71.40	NA	NA
Panchal et al. [35]	Auto-associator-MLP	91.00	NA	NA
Verma [36]	Neural Algorithm	94.00	NA	NA
Quionlon et al. [46]	Modified C4.5 DT	NA	94.74	NA
Reyes et al. [45]	Fuzzy-Genetic	NA	97.51	NA
Hamilton et al. [44]	RIAC Method	NA	96.00	NA
Goodman et al. [43]	AIRS	NA	97.20	NA
Bennet et al. [42]	SVM-DT	NA	97.20	NA
Abonyi et al. [41]	NN-Fuzzy	NA	95.57	NA
Karol et al. [49]	k-NN	NA	NA	75.50
Zandt et al. [49]	MLP-BP	NA	NA	75.80
Ster et al. [50]	LVQ	NA	NA	76.60
Statlog et al. [49]	Logdisc	NA	NA	77.70
Polat et al. [48]	GDA LS-SVM	NA	NA	82.05
Polat et al. [47]	Fuzzy-AIRS	NA	NA	84.42
Carpenter et al. [57]	Fuzzy ARTMAP	NA	NA	78.5
Proposed PNSCF	Majority-Voting	100.00	98.00	90.00
Proposed PNCF	Neural based Data Fusion	94.00	96.00	86.00

that by parallel fusion of different clustering technique can improve classification accuracy significantly. The fusion by neural networks can further improve the accuracy. The proposed research has proven that the hybrid combination of SOM and k-Means clustering with the classification strength of MLP is most appropriate way to develop hybrid data mining systems.

10 Conclusion and future work

The structure of medical data repositories, which consist of complex, large and unlabelled data samples, seemed to be a good candidate for unsupervised learning algorithms. This study identified that the unsupervised learning algorithms such as self-organizing map (SOM) and k-Means which have been reported in various medical classification/prediction literature ranges from feature selection, extraction, and data clustering to data visualization. Significance of neural networks such as BP based MLP classifiers, has also been extensively reported in the literature for the classification of different medical databases. What makes neural networks a promising method is their ability to generalize and reach near-optimum solutions from incomplete data and what's more the ability to combine data of a different nature in one system, such as data derived from medical reports and medical images.

The proposed research focused on the hybrid combination of SOM and k-Means clustering and the classification strength of MLP to form a novel hybrid ensemble approach. This paper investigated the proposed hybrid ensemble approach, which combined various clustering algorithms in parallel such as parallel neural-based strong clusters fusion (PNSCF) and parallel neural-based clusters fusion (PNCF). The approaches were evaluated on three benchmark medical databases such as Digital Database for Screening Mammograms (DDSM), Wisconsin Breast Cancer Database (WBCD), and Pima Indians Diabetes Database (PIDD), which are of great importance in medicine. The comparative analysis was also presented to compare the performance of the proposed approach with existing individual and hybrid data mining approaches. The PNSCF and PNCF approaches showed great performances.

According to the experimental results, the proposed approach showed a considerably higher performance with regard to classification accuracy for the DDSM, WBCD, and PIDD databases. The classification accuracies of the proposed approaches for the databases used were the highest among the existing approaches used for related prob-

Table 16 Summary of best classification accuracies using proposed and existing approaches

Author	Approach	Classification accuracy (%)		
		DDSM	WBCD	PIDD
Verma et al. [36]	Neural Algorithm	94.00	NA	NA
Reyes et al. [45]	Fuzzy-Genetic	NA	97.51	NA
Polat et al. [47]	Fuzzy-AIRS	NA	NA	84.42
Proposed PNSCF	Strong Clusters	100.00	98.00	90.00

lems, as shown in Table 16. When comparing proposed approaches with other existing approaches on benchmark medical databases, the proposed hybrid data mining approaches, PNSCF and PNCF have demonstrated better performance compared to existing approaches by achieving 100% and 94% classification accuracies on DDSM database, 98% and 96% classification accuracies on WBCD database, and 90% and 86% classification accuracies on PIDD database respectively. Note: Table 16 only shows the existing approaches which obtained higher accuracy in particular problem domain and then compared with proposed PNSCF approach which attained highest accuracies.

As shown in Table 16, the proposed approach improved the existing classification accuracy by 6% on DDSM, 0.49% on WBCD, and 5.68% on PIDD respectively. This shows the significance of the proposed hybrid approach and its contribution in the field of medical data classification. The results obtained in this paper confirmed the validity of the hybrid approach for medical decision-making.

The research presented in this paper leads to several interesting new research possibilities. A new strategy can be considered to assign a confidence value to the individual or group of clusters generated by different clustering algorithms in the proposed approach. For this reason, an algorithm can be designed which can automatically assign the weighted-values or set the threshold value which could provide further insight into the process of identifying and selecting strong clusters. This research used an MLP with BP training algorithm. It will be interesting to investigate and see the impact of fusion by replacing MLP-BP with other learning techniques such as RBF and SVM.

Acknowledgement This work was supported by ARC ISSNIP Research Network Grant.

References

1. Damien M, Graham JW, Jie C, Huidong J (2005) A delivery framework for health data mining and analytics. In: Proceedings of the twenty-eighth Australasian conference on computer science, Newcastle, Australia, pp 381–387
2. Gulbinat W (1997) What is the role of WHO as an intergovernmental organisation In: The coordination of telematics in healthcare? World Health Organisation. Geneva, Switzerland at http://www.hon.ch/libraray/papers/gulbinat.html
3. Handl J, Knowles J (2007) An evolutionary approach to multiobjective clustering. IEEE Trans Evol Comput 56–76
4. Korkmaz EE, Du J, Alhajj R, Barker K (2006) Combining advantages of new chromosome representation scheme and multiobjective genetic algorithms for better clustering. In: Proceedings of intelligent data analysis, pp 163–182
5. Boulis C, Ostendorf M (2004) Combining multiple clustering systems. In: Boulicaut J, Esposito F, Giannotti F, Pedreschi D (eds) 8th European conference on principles and practice of knowledge discovery in databases. Lecture notes in computer science, pp 63–74
6. Fred ALN, Jain AK (2005) Combining multiple clusterings using evidence accumulation. IEEE Trans Pattern Anal Mach Intell 835–850
7. Evgenia D, Andreas W, Kurt H (1999) Voting in clustering and finding the number of clusters. In: Bothe H, Oja E, Massad E, Haefke C (eds) Proceedings of the international symposium on advances in intelligent data analysis (AIDA 99). ICSC Academic Press, pp 291–296
8. Greene D, Tsymbal A, Bolshakova N, Cunningham P (2004) Ensemble clustering in medical diagnostics. In: Proceedings of the 17th IEEE symposium on computer-based medical systems. IEEE Comput Soc, Washington, pp 576–581
9. Lourenco A, Fred A (2005) Ensemble methods in the clustering of string patterns. In: Proceedings of the seventh IEEE workshops on application of computer vision. IEEE Comput Soc, Washington, pp 143–148
10. Greene D, Cunningham P (2006) Efficient ensemble methods for document clustering. Tech Rep TCD-CS-2006-48. Department of Computer Science, Trinity College Dublin
11. Chen D, Chang RF, Huang YL (2000) Breast cancer diagnosis using self-organizing map for sonography. Ultrasound Med Biol 405–411
12. West D, West V (2000) Model selection for a medical diagnostic decision support system: a breast cancer detection case. Artif Intell Med 183–204
13. Pattaraintakorn P, Cercone N, Naruedomkul K (2005) Hybrid intelligent systems: selecting attributes for soft-computing analysis. In: 29th annual international computer software and applications conference (COMPSAC), pp 319–325
14. Dietterich TG (2000) Ensemble methods in machine learning. In: First international workshop on multiple classifier systems. Lecture notes in computer science, pp 1–15
15. Hu X (2001) Using rough sets theory and database operations to construct a good ensemble of classifiers for data mining applications. In: IEEE ICDM, pp 233–240
16. Dudoit S, Fridlyand J (2003) Bagging to improve the accuracy of a clustering procedure. Bioinformatics 1090–1099
17. Fischer B, Buhmann JM (2003) Path-based clustering for grouping of smooth curves and texture segmentation. IEEE Trans Pattern Anal Mach Intell 513–518
18. Leisch F (1999) Bagged clustering. Working Papers. SFB adaptive information systems and modeling in economics and management science. Institut für Information, Abt. Produktionsmanagement, Wien, Wirtschaftsuniv

19. Fred ALN (2001) Finding consistent clusters in data partitions. In: Roli F, Kittler J (eds) Proc 3d Int workshop on multiple classifier systems. LNCS, vol 2364, pp 309–318
20. Fred ALN, Jain AK (2002) Data clustering using evidence accumulation. In: Proc of the 16th international conference on pattern recognition, pp 276–280
21. Kellam P, Liu X, Martin NJ, Orengo C, Swift S, Tucker A (2001) Comparing, contrasting and combining clusters in viral gene expression data. In: Proceedings of 6th workshop on intelligent data analysis in medicine and pharmocology, pp 56–62
22. Boulis C, Ostendorf M (2004) Combining multiple clustering systems. In: Boulicaut J, Esposito F, Giannotti F, Pedreschi D (eds) 8th European conference on principles and practice of knowledge discovery in databases. Lecture notes in computer science, pp 63–74
23. Martin HCL, Alexander PT, Anil KJ (2004) Multiobjective data clustering. In: IEEE computer society conference on computer vision and pattern recognition, pp 424–430
24. Evgenia D, Andreas W, Kurt H (1999) Voting in clustering and finding the number of clusters. In: Bothe H, Oja E, Massad E, Haefke C (eds) Proceedings of the international symposium on advances in intelligent data analysis (AIDA 99). ICSC Academic Press, pp 291–296
25. Greene D, Tsymbal A, Bolshakova N, Cunningham P (2004) Ensemble clustering in medical diagnostics. In: Proceedings of the 17th IEEE symposium on computer-based medical systems. IEEE Comput Soc, Washington, pp 576–581
26. Xiahua H, Illhoi Y (2004) Cluster ensemble and its applications in gene expression analysis. In: Proceedings of the second conference on Asia-Pacific bioinformatics. Dune din, New Zealand, vol 29, pp 297–302
27. Setiono R (2000) Generating concise and accurate classification rules for breast cancer diagnosis. Artif Intell Med 205–219
28. Blake CL, Merz CJ (1996) UCI repository of machine learning databases. Available from http://www.ics.uci.edu./~mlearn/MLReporsitory.html
29. Joachim D, Sabine B, Johann FD (1993) Segmentation of microcalcifications in mammograms. IEEE Trans Med Imag 12–18
30. Jerez-Aragones JM, Gomez-Ruiz JA, Ramos-Jimenez G, Munoz-Perez J, Alba-Conejo E (2003) A combined neural network and decision trees model for prognosis of breast cancer relapse. Artif Intell Med, pp 45–63
31. Kıyan T, Yıldırım T (2003) Breast cancer diagnosis using statistical neural networks. In: XII TAINN symposium proceedings, Çanakkale, Turkey 754–761
32. Kayaer K, Yıldırım T (2003) Medical diagnosis on Pima indian diabetes using general regression neural networks. In: Artificial neural networks and neural information processing (ICANN/ICONIP), Istanbul, Turkey, June 26–29, pp 181–184
33. Kemal P, Salih G, Ahmet A (2008) A cascade learning system for classification of diabetes disease: generalized discriminant analysis and least square support vector machine expert systems with applications, pp 482–487
34. Watkins AB (2005) Exploiting immunological metaphors in the development of serial, parallel, and distributed learning algorithms. PhD dissertation, University of Kent, Canterbury, March
35. Panchal R, Verma B (2006) Neural classification of mass abnormalities with different types of features in digital mammography. Int J Comput Intell Appl, pp 61–67
36. Verma B (2006) A neural learning algorithm for the diagnosis of breast cancer. IEEE international joint conference on neural networks, IJCNN'06, Canada. IEEE Press, New York, pp 10786–10791
37. Mahmoud RH, Yo-Sung H (2005) Automated detection of tumours in mammograms using two segments for classification, pp 910–921
38. Anna K, Ioannis B, Spyros S, Philippos S, Eleni L, George P, Lena C (2006) A texture analysis approach for characterizing microcalcifications on mammograms. In: International special topic conference on Information technology in bio medicine, pp 251–257
39. Osmar RZ, Maria-Luiza A, Alexandru C (2002) Mammography classification by an association rule-based classifier. In: Third international ACM SIGKDD workshop on multimedia data mining (MDM/KDD'2002) in conjunction with eighth ACM SIGKDD, Edmonton, Alberta, Canada, pp 62–69
40. Keir B, Sameer S (2002) Classification of mammographic breast density using a combined classifier paradigm. In: 4th international workshop on digital mammography, pp 177–180
41. Abonyi J, Szeifert F (2003) Supervised fuzzy clustering for the identification of fuzzy classifiers. Pattern Recogn Lett 2195–2207
42. Bennet KP, Blue JA (1997) A support vector machine approach to decision trees. Math Report, Rensselaer Polytechnic Institute, pp 97–100
43. Goodman DE, Boggess L, Watkins A (2003) An investigation into the source of power for AIRS, an artificial immune classification system. In: Proceedings of the international joint conference on neural networks (IJCNN '03). IEEE Press, New York, pp 1678–1683
44. Hamilton HJ, Shan N, Cercone N (1996) RIAC: a rule induction algorithm based on approximate classification. Technical Report CS 96-06, University of Regina
45. Pena-Reyes CA, Sipper M (1999) A fuzzy-genetic approach to breast cancer diagnosis. Artif Intell Med 131–155
46. Quinlan JR (1996) Improved use of continuous attributes in C4.5. J Artif Intell Res 77–90
47. Polat K, Gunes S, Tosun S (2006) Diagnosis of heart disease using artificial immune recognition system and fuzzy weighted preprocessing. Pattern Recogn 2186–2193
48. Polat K, Ahan SS, Gunes S (2006) A new method for medical diagnosis: artificial immune recognition system (AIRS) with fuzzy weighted preprocessing and application to ECG arrhythmia. Expert Syst Appl 264–269
49. Weiss SM, Kapouleas I (1990) An empirical comparison of pattern recognition, neural nets and machine learning classification methods. In: Shavlik JW, Dietterich TG (eds) Readings in machine learning. Morgan Kauffmann, San Mateo
50. Ster B, Dobnikar A (1996) Neural networks in medical diagnosis: comparison with other methods. In: Proceedings of the international conference on engineering applications of neural networks (EANN '96), pp 427–430
51. Mitra S, Banka H, Pedrycz W (2006) Rough-fuzzy collaborative clustering. IEEE Trans Syst Man Cybern, Part B 36(4):795–805
52. Wiering MA, van Hasselt H (2008) Ensemble algorithms in reinforcement learning. IEEE Trans Syst Man Cybern, Part B 38(4):930–936
53. Liu Y, Yao X (1999) Simultaneous training of negatively correlated neural networks in an ensemble. IEEE Trans Syst Man Cybern, Part B 29(6):716–725
54. Islam MM, Yao X, Shahriar SM, Islam MA, Murase K (2008) Bagging and boosting negatively correlated neural networks. IEEE Trans Syst Man Cybern, Part B 38(3):771–784
55. Parikh D, Polikar R (2007) An ensemble-based incremental learning approach to data fusion. IEEE Trans Syst Man Cybern, Part B 37(2):437–450
56. Hassan SZ, Verma B (2007) A hybrid data mining approach for knowledge extraction and classification in medical databases. In: 7th international conference on intelligent systems design and applications, Brazil, pp 503–510
57. Carpenter GA, Tan AH (1993) Rule extraction, fuzzy ARTMAP, and medical databases. In: Proceedings of world congress on neural networks, Portland, USA, vol I, pp 501–506

58. Carpenter GA (1997) Distributed learning, recognition, and prediction by ART and ARTMAP neural networks. Neural Netw 10(8):1473–1494
59. Carpenter GA, Grossberg S, Markuzon N, Reynolds J, Rosen D (1992) Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. IEEE Trans Neural Netw 3(5):698–713
60. Carpenter GA, Grossberg S, Reynolds J (1991) ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. Neural Netw 4(5):565–588

Brijesh Verma is a Professor of Information Technology in the School of Computing Sciences at CQUniversity, Australia. His research interests include pattern recognition and computational intelligence. He has published twelve books, five book chapters and over hundred papers and supervised twenty nine research students. He has served as a chief/co-chief investigator on twelve competitive research grants. He has served on the organising and program committees of many conferences and editorial boards of international journals. He has served as a Chair of IEEE Computational Intelligence Society's Queensland chapter and won the 2009 Outstanding Chapter Award.

Syed Zahid Hassan has received his PhD from CQUniversity Australia in 2009. After graduating from CQUniversity, he joined Mxi Technologies in Sydney, Australia. His research interests include intelligent systems, data mining and pattern recognition. He has published 1 book chapter and over 20 papers in conferences and journals. He was a recipient of ARC research network's postgraduate award in 2008.