HIGH ACCURACY BACK-RETREAT DIFFUSION-FUZZY CLUSTERING OF BREAST CANCER DATA FOR THE DETECTION OF MALIGNANCY

Ashutosh Patri
Department of Mining Engineering
National Institute of Technology, Rourkela
Rourkela, 769008, India
email: ashutoshpatri@gmail.com

Abhijit Nayak
Department of CSE
National Institute of Technology, Rourkela
Rourkela, 769008, India
email: abhijit623@gmail.com

Anup Anurag
Department of Information Technology and Electrical Engineering
ETH, Zurich
Zurich, Switzerland
email: anuraga@student.ethz.ch

ABSTRACT
A novel fuzzy clustering method has been proposed here for separating the breast cancer data, which operates with reasonable accuracy, allows flexibility in dataset and is modestly time consuming. This method can be applied to any type of cancer data set with some initial labels to obtain high accuracy result in the classification of unlabeled samples. Further, the curse of dimensionality is not an issue for the proposed scheme as it can be applied to data having any number of dimensions or attributes. The DifFUZZY unsupervised clustering algorithm is applied at the initial stage, giving an accuracy of 96.28% over Wisconsin Breast Cancer Dataset (WBCD); the result is further improved to 98.14% by using the proposed Back-Retreat algorithm. The formed clusters are estimated using three internal cluster validation indices and the performance of the method is evaluated using receiver operating characteristic (ROC) curves. The clustering algorithm is compared with Fuzzy C-Means (FCM) algorithm and the results are compared with different classifiers and clustering techniques.

KEY WORDS
DifFUZZY; Wisconsin Breast Cancer Data; Fuzzy C-Means; Breast Cancer; Fuzzy Clustering.

1 Introduction

Medical science is saving people from dangerous diseases, but still it is inefficient in curing or battling cancer. No preventive solution has been invented yet. However, lives can be saved by treating cancer using surgical approach if diagnosed at an earlier stage. Breast cancer is one type of non skin cancer that can wreck havoc in women. The scientific study says it's 100 times more likely in a woman body as compared to a man [1]. In 1975, 105 new cases were diagnosed for every 100,000 woman in USA, whereas in 2007 it increased to 125 [2]. But the overall mortality rate was decreased from 31 to 23 in the respective years. In United States woman breast cancer is the second most important reason of cancer related death. Research study says that in India every 1 in 28 women is likely to have breast cancer [3]. In 2010, around 1.5 million women were diagnosed with breast cancer. There has been more emphasis on diagnostic techniques in recent past that led to effective treatments causing considerable decrease in overall mortality rate.

Breast Cancer diagnosis includes three steps i.e. clinical examination, radiological investigation and pathological correlation. Pathological test is done using Fine Needle Aspiration (FNA), Surgical excision, tru-cut, percutaneous breast biopsy and core biopsy for confirmation of the malignancy in breast tissue [2]. Though the Surgical Biopsy methods give high accuracy in malignancy detection these are very costly and have high negative impact on patient psychology, whereas FNA is a relatively non-invasive, inexpensive, less painful and quicker method when compared to other methods of tissue sampling. Moreover FNA has been used as a diagnostic tool for breast lesions, with high sensitivity and specificity for many years and continues as an acceptable and reliable procedure of preoperative diagnosis, particularly in developing countries [4].

From medical perspective, a quick diagnosis means early detection of cancer that allows for more treatment options and eliminates the need for chemotherapy or other very expensive targeted therapy. FNA biopsies can give results within a shorter time period but some expertise is needed in order to achieve accurate result. In order to avoid human error some computer based classification system is needed. So for an accurate and successful diagnosis of new cases an efficient classifier or clustering technique is very much essential.

There have been many researches with WBCD in literature that include different clustering techniques or classifiers for the diagnosis of breast cancer. Earlier, the diagnosis and prognosis of breast cancer data was done using linear programming based Machine learning initiated by W.H. Wolberg, O.L. Mangasarian and W.N. Street at Wis-

DOI: 10.2316/J.2014.216.818-0018

consin Hospital [5].In [6], Artificial Immune System (AIS) is compared with FCM and the analysis is done by implementing both algorithms on breast cancer data set. ANNs were trained many times to determine the optimum parameters in AIS for achieving highest classification accuracy of 97.8%. In [7], Kernel Method Clustering (KMC) algorithm is compared with SOM, K-means and neural gas algorithms for the clustering of WBCD. KMC is a batch clustering algorithm and it remains unaffected by the pattern ordering in the training set giving an accuracy of 97.0%. A generalized hybrid unsupervised learning algorithm, termed as rough-fuzzy probabilistic c-means (RFPCM) is proposed in [8] giving an accuracy of 91.92% which precisely integrates the principles of rough and fuzzy set while the Probabilistic Latent Variables (PLV) proposed in [9] replaces the binary latent variable with the fuzzy latent variable, indicating the belongingness of an object to a certain possible cluster, rather crisply assigning the object to that particular cluster with an accuracy of 96.05%. In [10] a RF-ANN structure is applied to the Wisconsin breast cancer data set where artificial neural network (ANN) is used as the base classifier and Rotation Forest (RF) algorithm is used as ensemble classifier and the obtained results are compared with the results of neural network optimized particle swarm optimization (PSO-ANN) giving an accuracy of 98.05% and 97.36% respectively. The Axiomatic Fuzzy Sets (AFS) fuzzy logic clustering algorithm has been studied further in [11], by improving the algorithm in the fuzzy description of each objects, each cluster and the final clustering criteria gives an accuracy of 95.9% when applied to Wisconsin breast cancer data. This improved algorithm can be applied to the data set with various data types such as numerical values, boole values, partial order relations, even human intuition descriptions. In [12], three different methods, optimized learning vector quantization (LVQ), big LVQ, and artificial immune recognition system (AIRS), are applied giving accuracies of 96.7%, 96.8%, and 97.2%, respectively. A new classifier based on Multi-Attributed Lens Recursive Partitioning Algorithm is implemented in [13] which gives an accuracy of 96.18% proving that the performances of this algorithm is better than C4.5 algorithm. [14] aims at finding the most suitable multi-classifier for breast cancer data set which compares the accuracies of the five classifiers such as Naïve Bayes (NB), Decision tree (J48), Multilayer Perception (MLP), K-nearest neighbor (IBK) and Self Organizing Map (SOM) based on 10-fold cross validation and finds SMO as the best with accuracy of 96.9957%. Following this a fusion at classification level between these classifiers is done. The fusion between SMO and MLP, SMO and IBK, SMO and NB all gives the same accuracy of 96.9957%. Fusion between three classifiers SMO, IBK and NB give an accuracy of 97.1388% while four including J48 gives an accuracy of 97.2818%. In [15], a medical decision making system based on SVM combined with feature selection has been applied on the task of diagnosing breast cancer. The importance of each feature is measured by F-score and the SVM parameters are optimized by grid search. The aim is to limit the number of input features in a classifier in order to have a good predictive and less computationally intensive model. Among nine different models constructed based on different features, model no. 5 achieved the highest classification accuracy; 98.53% for the 50-50% training-test partition, 99.02% for the 70-30% training-test partition, and 99.51% for the 80-20% training-test partition.

In this study a new clustering technique termed BR-DifFUZZY has been proposed for the diagnosis of breast cancer data. According to the proposed scheme general DifFUZZY algorithm is applied over the dataset in the first phase for assigning different data points into different clusters. The cluster members are further classified as core points and soft points within the cluster with only core points having full membership. In the second phase, the reformation of the core of a cluster is done by an iterative process termed as Back-Retreat algorithm for eliminating the partial memberships of the data points thus minimizing error. The proposed method is flexible enough to be applied over a dataset having any number of attributes and data points. The proposed method has been applied on Wisconsin Breast Cancer Dataset (WBCD) and is compared with basic fuzzy C-means clustering technique giving an accuracy of 98.14%. Also, other measures such as precision, recall, rate of positive predictions, lift values and ROC curves are used to show the performance of BR-DifFUZZY.

The paper can be organized as follows; second section describes general fuzzy C-means and DifFUZZY clustering algorithm and the proposed Back-Retreat Algorithm. Third section gives a brief idea about the chosen dataset. Section four covers simulation result analysis and discussion part which includes some comparative study with other clustering and classification techniques. Finally section five concludes the paper.

2 Fuzzy Clustering and Back-Retreat Algorithm

2.1 C-Means Fuzzy Clustering

The fuzzy C-Means Functional [16] was developed by Dunn and improved by Bezdek which is given by

$$J(Z;U,V)=\sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ik})^m\|Z_k-V_i\|_A^2 \qquad (1)$$

where $m \in [1,\infty)$ is the parameter that determines the fuzziness of the corresponding cluster.
$Z_k=[z_{1k},z_{2k},...z_{nk}]$ and $Z_k \in R^n$ represents 'N'data points with 'n'dimensions and $U=[\mu_{ik}] \in M_{fc}=\{U \in R^{c*N} | \mu_{ik} \in [0,1]\ , \forall i,k;\ \sum_{i=1}^{c}\mu_{ik}=1, \forall\, k; 0<\sum_{k=1}^{n}\mu_{ik}<N, \forall\, i\}$ is the fuzzy partition matrix of Z. μ_{ik} is the membership of the 'i^{th}'data point at 'k^{th}'cluster and 'c'is the number of clusters formed.

$$V=[V_1,V_2,V_3,...,V_c],\ V_i \in R^n \qquad (2)$$

Eqn. (2) represents the vector of centers and V_i is defined by (3). This is determined by (4), a squared inner product distance norm. The given functional in (1) is minimized by nonlinear optimization that includes iterative steps.

$$V_i = \frac{\sum_{k=1}^{N} (\mu_{ik})^m z_k}{\sum_{k=1}^{N} (\mu_{ik})^m};\ 1 \leq i \leq c \tag{3}$$

$$D_{ikA}^2 = \|Z_k - V_i\|_A^2 = (Z_k - V_i)^T A (Z_k - V_i) \tag{4}$$

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} \frac{D_{ikA}}{D_{jkA}}^{\frac{2}{(m-1)}}},\ 1 \leq i \leq c,\ 1 \leq k \leq N \tag{5}$$

"By sequential iteration steps the membership given by (5) as well as the cluster center given by (3), both are updated. The iteration process stops when $\|U^{k+1} - U^k\| < \epsilon$, where ϵ is a termination tolerance.

2.2 DifFUZZY Clustering

This algorithm [17] is developed by O. Cominetti and his research group, utilizing the diffusion process in Graph theory and fuzzy clustering for handling high dimensional microarray data. For capturing information of higher order neighborhood structure the concept of Random-Walk is utilized in this diffusion model. DifFUZZY identifies core clusters by constructing a hierarchy of (Euclidean) neighborhood graphs and solving a discrete optimization problem and then assigns membership values to the data set by using a diffusion distance.

The input dataset is given by $Z_1, Z_2, Z_3, \ldots, Z_N \in R^n$ where 'N 'and 'n 'defines the number of elements and the dimensions respectively. The data-points are divided into two groups i.e. Core points and Soft points. The clustering process is initiated first by defining the core points. An auxiliary function $F(\sigma) : (0, \infty) \to \mathbb{N}$ where σ is the Euclidian norm defined as $\|Z_i - Z_j\| < \sigma$ and $F(\sigma)$ is the number of components of σ neighborhood graph that contains at-least M vertices. As defined by the algorithm M is an external parameter that should be suitably chosen. The number of cluster is defined as the maximum of this function $F(\sigma)$.

Then the value of σ^* is computed which is the minimum value σ for which maximum number of clusters are made. After the computation of σ^* , the corresponding neighborhood graph is constructed and the components of this graph that contains at-least M vertices are defined as the Core points of the clusters. The membership value of the Core point in its own cluster is 1 whereas in other clusters its value is 0.

After the identification of Core points the other data points in the dataset are treated as the Soft points and their membership values are calculated for each formed clusters. The membership function is defined as given in (6).

$$\mu_c(Z_k) = \frac{dist(Z_k, C)^{-1}}{\sum_{l=1}^{C} dist(Z_k, l)^{-1}} \tag{6}$$

where

$$dist(Z_k, C) = \|P^\alpha e(j) - \bar{P}^\alpha e(j)\| \tag{7}$$

is the diffusion distance of data point Z_k from the C^{th} cluster. The value of $e(j) = 1$ for $j = k$ and $e(j) = 0$ in other cases. The membership $\mu_c(Z_k)$ is evaluated using the Euclidian distance of Z_k from the closest Core point of the corresponding cluster. The matrix P and $\bar{P}$ is defined by these following set of equations.

$$P = I + [W - D]\frac{\gamma_2}{maxD_{i,i}} \tag{8}$$

where,$I \in R^{(N \times N)}$ is an identity matrix. The auxiliary matrix W is defined as $W = \hat{W}(\beta^*)$. $\hat{W}(\beta^*)$ is the matrix with entries given by (9).

$$\hat{W}_{i,j}(\beta) = \begin{cases} 1, & i\ and\ j\ are\ Core\ points \\ & in\ same\ cluster \\ exp(\frac{-\|z_i - z_j\|^2}{\beta}), & otherwise \end{cases} \tag{9}$$

The parameter β is defined by following equations. A function $L(\beta) : (0, \infty) \to (0, \infty)$ is defined as given by (10).

$$L(\beta) = \sum_{i=1}^{N} \sum_{j=1}^{N} \hat{W}_{i,j}(\beta) \tag{10}$$

To find the desired value of β i.e. β^*, the (11) is evaluated

$$L(\beta^*) = (1 - \gamma_1)(N + \sum_{i=1}^{c} n_i(n_i - 1)) + \gamma_1 N^2 \tag{11}$$

where, n_i is the number of Core points in i^{th} cluster and γ_1 is an internal parameter whose value lies between 0 to 1 and here the default value (0.3) is taken. This parameter is associated with the time scale of Random-Walk. The parameter evolved from the concept of diffusion of data points in graph theory. When the value of γ_1 approaches 1, it signifies that the data points are highly connected and for $\gamma_1 \sim 0$ there will be no diffusion between the cluster cores. The matrix D is defined as the diagonal matrix with entries defined by (12).

$$D_{i,i} = \sum_{j=1}^{N} W_{i,j}\ ,\ i = 1, 2, 3, ..., N \tag{12}$$

The value of $\gamma_2 \in (0, 1)$ in (6) is another internal parameter and its default value of 0.1 is taken. For $\gamma_2 \sim 0$, from (8) it can be found that $P \sim I$.This parameter also ensures the entries for the matrix P is always non-negative.

The matrix $\bar{P}$ and $\bar{D}$ is evaluated using (6) and (10) using $\bar{W}$ in which the 'z^{th}'row and 'z^{th}'column of matrix W is replaced by the 'n^{th}'row and 'n^{th}'column respectively. The value of α that determines the diffusion distance in (7) is defined by (13), where $\gamma_3 \in (0, \infty)$ is an internal parameter with a default value of 1.0 and it defines

the number of time steps taken in the Random-Walk. λ_2 is the second largest eigenvalue of matrix P.

$$\alpha = \lfloor \frac{\gamma_3}{|log\lambda_2|} \rfloor \quad (13)$$

The matrix P can be thought of as a transition matrix whose all rows are summed up to 1, and whose entry $P_{i,j}$ corresponds to the probability of jumping from the node (data point) i to the node j in one time step. The j-th component of the vector $P^{\alpha}e$, which is used in (7), is the probability of a random walk ending up in the j-th node, $j = 1, 2, ..., N$ after α time steps, provided that it starts in the z^{th} node.

This matrix is also used to represent a different diffusion process, which is equivalent to the first random walk, but over a new graph, where the position of the data point Xs has been altered with the position of the data point Xn. This matrix then acts like the transition matrix for this auxiliary graph. In this context, the transition matrix P used in (7), which can be written as $P = I + (W - D)\Delta t$, is essentially a first-order approximation to the heat kernel of the graph associated with $L = D - W$. In particular, for every $\Delta t \geq 0$, the heat kernel $H_{\Delta t}$ of a graph G with graph Laplacian L is defined to be the matrix $H_{\Delta t} = e^{-\Delta t L} = I - \Delta t L + \frac{\Delta t^2 L^2}{2} - \ldots$. The importance of $H_{\Delta t}$ is that it defines an operator semi-group, describing fundamental solutions of the spatially discretized heat equation $u_t = (W - D)u$. Heat kernels are powerful tools for defining and investigating random walks on graphs, and they provide a connection between the structure of the graph, as encoded in the graph Laplacian, and the asymptotic behavior of the corresponding random walk.

2.3 Back-Retreat DifFUZZY Algorithm

Here, the Back-Retreat (BR) algorithm is amalgamated with the previously described DifFUZZY algorithm for increasing the efficiency in clustering of the available labeled dataset so that it will increase the sensitivity of this hybrid clustering method to classify the unlabeled new sample correctly into its category i.e. benign or malignant. The BR-DifFUZZY algorithm is described below step by step. Let Z be a set of N independently distributed data points having identical dimensions such that $Z_1, Z_2, Z_3, \ldots, Z_N \in Z$ and Y be the set of corresponding labels such that $L_1, L_2, L_3, \ldots, L_n \in Y$ and C be the set of clusters formed such that $C_1, C_2, C_3, \ldots, C_p \in C$

- **Step 1:** Calculate Membership matrix, M_mat by DifFUZZY clustering algorithm
  M_mat = DifFUZZY $(Z, M, [\gamma_1, \gamma_2, \gamma_3])$
  where,
  M: external parameter
  $\gamma_1, \gamma_2, \gamma_3$: internal parameter.

- **Step 2:** In M_mat,
  For each Z_i with $\mu_{Cj} = 1$
  Find L_{Zi} and compare it with C_j.
  If $(L_{Zi} == C_j)$
  Z_i is denoted as a Core point (CP).
  Else
  Core error found.
  Do nothing.
  End
  End
  μ_{Cj} (Z_i) : The membership value of data point Z_i in the cluster C_j.
  Set the flag.

- **Step 3:** In M_mat,
  Z_is having $0 < \mu_{Cj} < 1$ are denoted as Soft points(SPs).
  For each SP $\in$ Z
  Determine C_j for which μ_{Cj}(SP) is maximum.
  Find L_{sp} and compare with C_j.
  If $(L_{sp} == Cj)$
  Then make μ_{C_j} (SP) =1.
  i.e treat SP as a CP.
  Reset the flag.
  Else
  Membership error in cluster C_j.
  Do nothing.
  End
  End

- **Step 4:**
  If (flag is reset)
  Update CP values.
  Find M_mat = DifFUZZY (Z).
  With core formed by the CPs.
  Repeat from STEP 2.
  Else
  Results obtained.
  i.e the initial core structure of the clusters.
  End
  Return

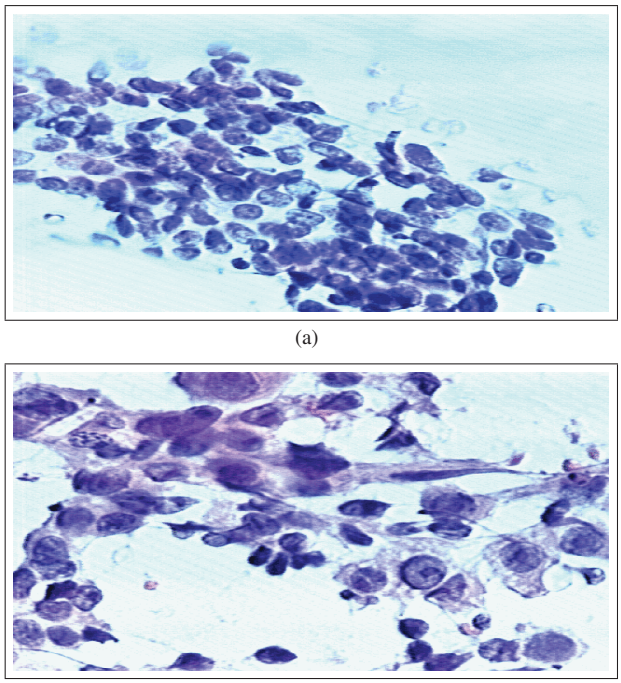
Every new unlabeled data should be included in its corresponding class of existing labeled dataset with its label if it is proved to be predicted perfectly by this clustering process otherwise it should be treated as an error. So that the clustering process will be more sensitive and the accuracy in diagnosing new cases will increase considerably.

3 Description of the Dataset

3.1 Wisconsin Breast Cancer Dataset

As mentioned in the introduction part the Fine Needle Aspiration (FNA) method is used for the diagnosis purpose and image analysis is done to determine various features of the fluid taken from the tumour portion of the breast. The tumour is categorized into two types i.e. Malignant and Benign. Fig. 1(a) and fig. 1(b) show the benign and malignant breast mass image respectively. Benign tumours are not cancerous and it can be removed; as in this case, affected

cells dont attack other tissue. Whereas, malignant tumour is cancerous and the affected cells break away and enters the lymphatic system forming secondary tumours [18, 19]. The novel work of making Wisconsin breast cancer dataset



(a)

(b)

Figure 1. (a) Benign Breast mass[20] (b) Malignant Breast mass [20]

was initiated by Dr. William H. Wolberg and Prof. Olvi L. Mangasarian of Wisconsin-Madison University, USA. The collected data are properly digitized (an integer value ranging from 1 to 10) into 9 attributes i.e. clump thickness(CT), Uniformity of Cell Size(UC), Uniformity of Cell Shape(UCS), Marginal Adhesion(MA), Single Epithelial Cell Size(SECS), Bare Nuclei(BN), Bland Chromatin(BC), Normal Nuclei(NN), and Mitoses(MI). The dataset contains total 699 data points in which 458 data are benign and 241 data are malignant. There are 16 missing values in this dataset which are replaced by the average value of that feature for experimental evaluation. The digitized values assigned for benign and malignant are 2 and 4 respectively.

4 Results, Analysis and Discussion

For the simulation work of the proposed model, MATLAB Version 7.6.0.324 r2008a is used. As mentioned earlier, Wisconsin Breast Cancer Data (WBCD) is considered as the experimental dataset and DifFUZZY algorithm is implemented in MATLAB as described in section II. For fuzzy C-means and K-means clustering, inbuilt functions of MATLAB are used for experimental evaluation. In unsupervised clustering process there are no predefined groups and the grouping process for various algorithms is different depending on the interaction of algorithm with the dataset and the initial assumptions made. So for sensitive dataset clusters validity indices are used to have an in-depth idea about the number of clusters and the clustering process.

There are three fundamental clustering evaluation criteria i.e. external, internal and relative criteria. The evaluation using external criteria is regarding the comparisons between the pre-specified structures, in which the evaluation is associated with the number of data points that are not used in clustering and this criterion reflects the intuition about the clustering structure of the given dataset. The evaluation using relative criterion is by comparing the algorithm with other clustering algorithms and then by varying the input parameters of the same. But the internal criteria are the most important criteria for deciding the number of cluster that is to be made. In case of internal criteria, the clustering result analyses are assessed in terms of those parameters that involve the vectors or the attributes of the dataset itself. So for a perfect evaluation of the clustering, the benign or malignant attribute of each data point is assumed to be unknown and considering three internal indices the number of cluster is decided. In FCM clustering the number of clusters i.e. 2 is given as an input before starting the clustering process in MATLAB simulation using fcm(data,cluster_n) inbuilt function. But for an unlabeled dataset other than WBCD the number of clusters is unknown for which cluster estimation using internal indices is indispensable. Here three basic parameters i.e. Davies-Bouldin Index (DB) [21], Dunn's Index [22] and Calinski Harabasz (CH) index [23] are used to estimate the number of clusters in WBC dataset. The overlap of benign and malignant classes in WBC dataset is vey less, so by optimizing the objective functions for each of the above three indices the same value '2'is returned as the number of clusters. So, optimizing the objective functions for each of the above three indices returns the same value '2'as the number of clusters.

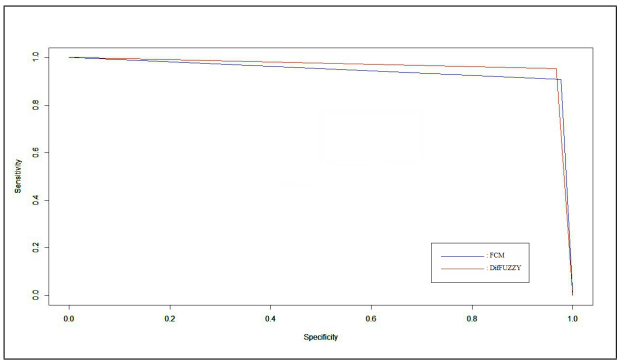
After estimating the number of cluster the external index M in DifFUZZY clustering is fixed to a suitable value. Here this value is taken 35 but it can be varied according to the dataset and validity indices. The internal indices are reserved to its default values. After DifFUZZY clustering 534 Core points and 165 Soft points are found. For 165 soft points its maximum membership value is considered to belong the corresponding class i.e. benign or malignant. The efficiency is found to be 96.28 %. The clustering results are given in table 1.

Table 1. Cluster Prediction Using DifFUZZY

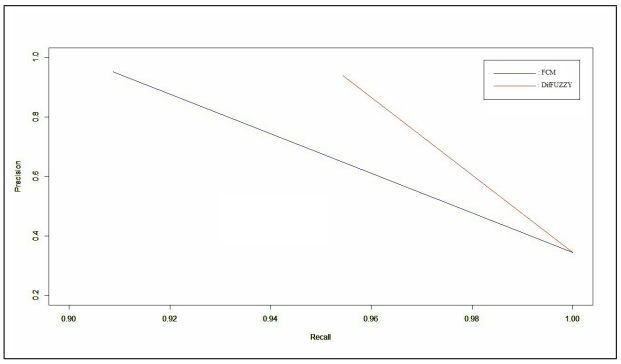
Clusters	DifFUZZY	
	Wrongly Predicted	Correctly Predicted
Cluster 1 (Benign)	11	443
Cluster 2 (Malignant)	15	230

Figure 2 (a) - (d) show various representation for the clustering model evaluation. Performance of both FCM and DifFUZZY algorithm on the WBC dataset is shown using these plots. For these plots ROCR package of R version 3.0.2 is used. Parameters like sensitivity, specificity, precision, recall, true positive rate, false positive rate,lift value and rate of positive prediction are taken into consideration for the model evaluation [24]. Fig.3 and fig.4 depicts FCM and DifFUZZY clustering of WBC dataset. The membership value plot of the DifFUZZY clustering is shown in fig.5 (a) and (b) which clearly shows that the clustering technique is highly sensitive. Then the proposed BR algorithm is applied to increase the efficiency in DifFUZZY clustering and it is found to be 98.14%. Fig. 6 depicts the clustering of WBCD using BR-DifFUZZY. Table 2 shows the comparative study among various clustering techniques and classifiers. The Kwoks Support Vector Machine using both Gaussian and polynomial kernel, K-nearest neighbor (IBK), C4.5, Multi Scale classifier (MSC), Fuzzy Entropy Based Fuzzy Classifier (FEBFC), Decision tree (J48), Multilayer Perception (MLP), Axiomatic Fuzzy Set (AFS), Naïve Bayes (NB) classifier are applied on WBCD dataset and the efficiencies are summarized [11,14] in the table 2. The classifiers results are shown after 10 fold cross validation. As mentioned earlier, in principle the DifFUZZY clustering process is divided into two parts i.e. selection of core points and assignment of membership values to the soft points. The back retreat algorithm converts the partial membership of the soft points into total membership thus making them as part of the core in the corresponding clusters. So the BR-DifFUZZY algorithm is considered as a variation over the general DifFUZZY algorithm where the reformation of the core of a cluster is done in each iteration based on the membership values of the soft points. In BR algorithm first the membership values of the soft points get modified. If this modification leads them to be the member of the clusters they actually belong to then in the next iteration they will be treated as the core members of the corresponding clusters.

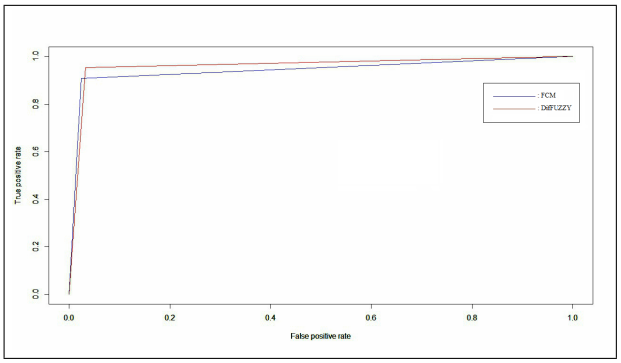
The membership value of some of the soft points which were wrongly predicted before may get modified enough by the newly formed core so that they can be included in their belonging cluster from the following iterations. Nevertheless some of the outliers are wrongly predicted and the clustering technique failed to separate them accordingly. In this paper 16 data points are used which have some missing attributes and those values are replaced by the average of the column they belong to and three of these are wrongly predicted. According to the research work done by Heiko Timm et al, the missing values have greater impact on the error in clustering and prediction [25]. This might have improved the efficiency.



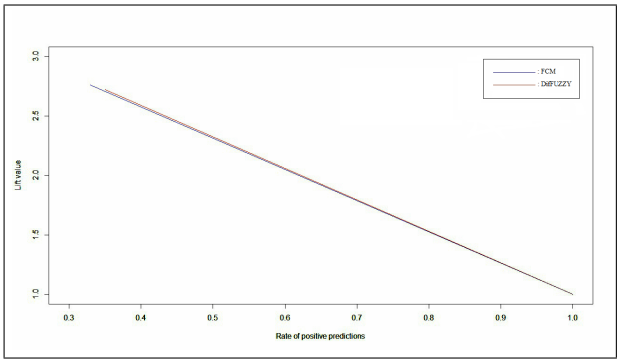
(a)



(b)



(c)



(d)

Figure 2. (a) Sensitivity Vs. Specificity (b) Precision Vs. Recall (c) TPR Vs. FPR (d) Lift Value Vs. Rate of Positive Predictions

5 Conclusion

Literature review shows that there have been many researches that try to diagnosis breast cancer with the help

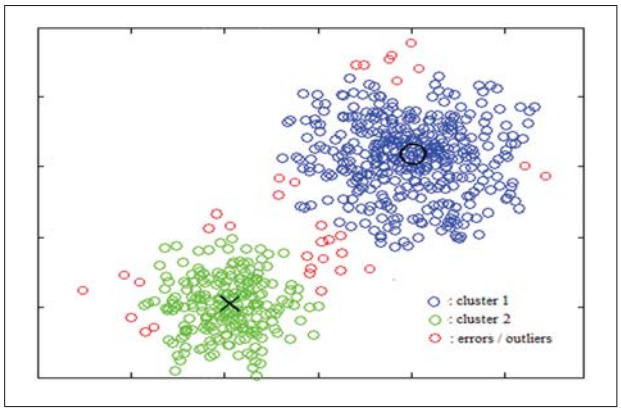


Figure 3. FCM clustering of WBCD

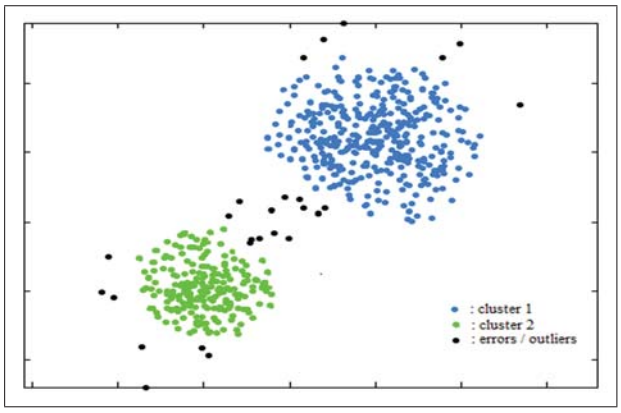
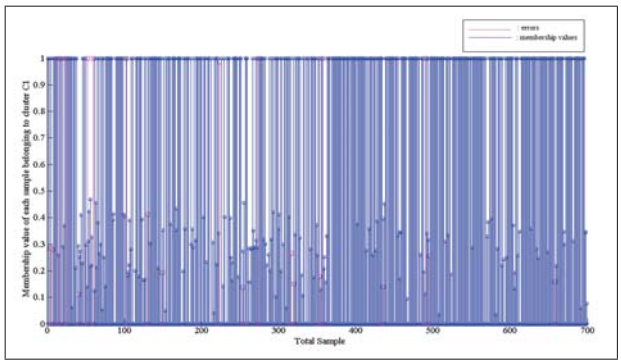


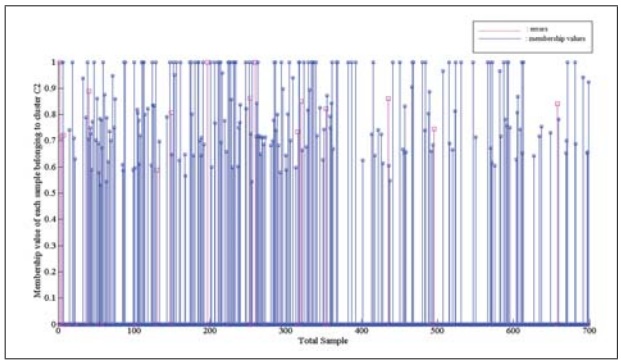
Figure 4. DifFUZZY clustering of WBCD

Table 2. Comparison between different Classifier and Clustering Techniques

Clustering & Classification Methods	*Accuracy (In %)*
Kwok'SVM Gaussian	91.6
Kwok'SVM Polynomial	93.6
IBK	94.5
C4.5	94.7
MSC	94.9
FEBFC	95.1
J48	95.13
MLP	95.27
FCM	95.28
Self Organizing Map	95.32
K-Means	95.7
AFS Fuzzy	95.9
NB	95.9
BR-DifFUZZY	98.14



(a)



(b)

Figure 5. (a) Membership in Cluster 1 (benign) (b)Membership in Cluster 2 (malignant)

of different clustering techniques and classifiers. None of those clustering techniques have reported a significant accuracy whereas some supervised learning techniques like SVM and Multi-classifier show superior results. However they all claim their accuracy over WBCD which has 699 data points, each having nine attributes mainly based on the clinical features of lesion such as size, shape and texture, but [26] shows that more attributes can be included in a dataset from FNA biopsy such as side of lesion, its distance from the nipple .This may gives rise to datasets having discrete values. Even [27] includes many examples where analyses are done over datasets having random number of attributes and data points.

But it has been observed that SVM does not perform well enough for random and discrete data. For classification problem involving more than 2 classes SVM loses any advantage it has over other classifiers. It also fails under an evolving or adaptive learning context. Moreover according to [28] Wisconsin Breast Cancer dataset should not be used as a benchmark for classification algorithms since any linear distance classifier will probably perform with an accuracy of over 90% and the non-linear classifiers accuracies will not show notable differences as these differences will derive only from the sparse, potentially erroneous, remaining non-linearity present in the data.

Due to reasons mentioned above, in this paper, we proposed a new clustering method termed BR-DifFUZZY

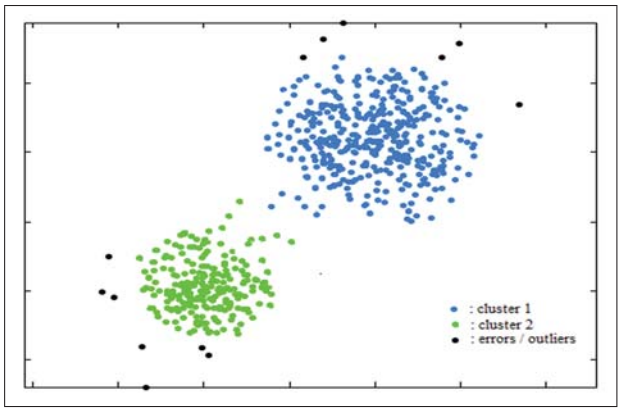


Figure 6. BR-DifFUZZY clustering of WBCD

that can differentiate between the benign and malignant tumors with reasonable accuracy over any dataset, as this algorithm separates the dataset into different clusters based on the underlying structure irrespective of type, dimension and number of data points. The proposed method was applied over WBCD due to the dearth of online digitized dataset and an accuracy of 98.14% was obtained, which is more than any other clustering methods; Hence on the basis of performance, the proposed Br-DifFUZZY technique was proved to be on level with other-state-of the art algorithms, which have been applied to WBCD earlier, making it an interesting alternative. The proposed scheme can be used as a tool which will aid the physicians in making an accurate decision on their patients. As the applications of our technique become varied, interesting results can be anticipated in future.

References

[1] What Are The Risk Factors For Breast Cancer? Available: http://www.cancer.org/cancer/breastcancer /detailedguide/breast-cancer-risk-factors

[2] National Cancer Institute Cancer Advances In Focus. Available: http://www.cancer.gov/cancertopics/factsheet/cancer-advances-in-focus/breast

[3] Cancer Information and Type of Cancer. Available: https://tmc.gov.in/cancerinfo/breast/breast.html

[4] B. Chaiwun and P. Thomer, Fine needle aspiration for evaluation of breast masses, *International Journal of Current Opinion in Obstetrics and Gyncology*, 19(1), 2007, 48-55

[5] O. L. Mangasarian et al. (1994). *Breast Cancer Diagnosis And Prognosis Via Linear Programming [Online]*. Available FTP: ftp.cs.wisc.edu Directory: mathprog/tech-reports File: 94-10.pdf.

[6] S. Ozsen and R. Ceylan, Comparison of AIS and fuzzy c-means clustering methods on the classification of breast cancer and diabetes datasets, *Turkish Journal of Electrical Engineering & Computer Science*, 2014, doi: 10.3906/elk-1210-62

[7] F. Camastra and A. Verri, A Novel Kernel Method for Clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5), 2005, 801-805

[8] P. Maji and S. K. Pal, Rough Set Based Generalized Fuzzy C-Means Algorithm and Quantitative Indices, *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 37(6), 2007, 1529 - 1540

[9] Z. X. Yin and J. H. Chiang, Patterns Discovery on Complex Diagnosis and Biological Data Using Fuzzy Latent Variables, *Proc. 23rd IEEE Int. Conf. on Data Engineering*, Istanbul, Turkey, 2007, 576-585.

[10] H. Koyuncu and R. Ceylan, Artificial Neural Network Based on Rotation Forest for Biomedical Pattern Classification , *Proc. 36th IEEE International Conf. on Telecommunication and Signal Processing*, Rome, Italy, 2013, 581 - 585.

[11] X. Wang, X. Liu, and L. Zhang, The Improved Fuzzy Clustering Algorithm Based on AFS Theory and Its Applications to Wisconsin Breast Cancer Data, *Proc. IEEE International Conf. on Intellectual Control and Information Processing*, Dalian, China, 2010, 374-378.

[12] D. E. Goodman, L. C. Boggess and A. B. Watkins, Articial immune system classication of multiple-class problems, *Proc. Articial neural networks in engineering*, St. Louis, Missouri, 2002, 179-183.

[13] C. Sirisomboonrat, and K. Sinapiromsaran, Breast Cancer Diagnosis using Multi-Attributed Lens Recursive Partitioning Algorithm, *Proc. 10th Int. Conf. ICT and Knowledge Engineering*, Bangkok, Thailand, 2012, 40 - 45.

[14] *G*. I. Salama, M. B. Abdelhalim, and M. A. Zeid, Experimental Comparison of Classifiers for Breast Cancer Diagnosis, *Proc 17th IEEE Int. Conf. on Computer Engineenig & Systems*, Cairo, Egypt, 2012, 180-185.

[15] M. F. Akay , Support vector machines combined with feature selection for breast cancer diagnosis , *International Journal of Expert Systems with Applications*, 36(2), 2009, 3240- 3247.

[16] J. C. Bezdek, R. Ehrlich and W. Full, FCM: The Fuzzy C-Means Clustering Algorithm, *Computers & Geosciences*, 10(2), 1984, 191-203.

[17] O. Cominetti, A. Matzavinos, S. Samarasinghe, D. Kulasiri, S. Liu, P. K. Maini, and R. Erban, DifFUZZY: A fuzzy clustering algorithm for complex data sets, *International Journal of Computational Intelligence in Bioinformatics and Systems Biology*, 1(4), 2010, 402-417.

[18] BreastCancer.Available:http://www.cancer.gov/cancertopics/types/breast

[19] Dr. William H. Wolberg. (1992). *Wisconsin Breast Cancer Database, UCI Machine Learning Repository*. Available: http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data

[20] W. N. Street et al. (1993). *Nuclear feature extraction for breast tumor diagnosis [Online]*. Available FTP: ftp.cs.wisc.edu Directory: math-prog/cpo-dataset/machine-learn/cancer/cancerimages File:$92\_7241.gif, 92\_5292.gif$.

[21] D.L. Davies and D.W. Bouldin, A Cluster Separation Measure, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2), 1979, 224-227.

[22] J.C. Dunn, A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, *Cybernetics and Systems*, 3(3), 1973, 32-57.

[23] T. Calinski and J. Harabasz, A Dendrite Method for Cluster Analysis, *Communication in Statistics*, 3(1), 1974, 1-27.

[24] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, ROCR: visualizing classifier performance in R, *Bioinformatics*, 21(20), 2005, 3940-3941.

[25] H. Timm, C. Dring, and R. Kruse , Fuzzy Cluster Analysis of Partially Missing Datasets, *Proc. 2nd European Symp. on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems*, Algarve, Portugal, 2002, 426-431.

[26] J. Bishop, M. Coleman, B. Cooke, R. Davies, F. Frost, J. Grace, L. Reeves, M. Rickard, N. Wetzig, and H. Zorbas, *The specimen: request, preparation and processing* (Breast fine needle aspiration cytology and core biopsy: a guide for practice,1st Ed., Camperdown, Australia, NBCC, 2004), ch. V, 27-34.

[27] Y.-H. Yu, W. Wei, and J.-L. Liu, Diagnostic value of fine-needle aspiration biopsy for breast mass: a systematic review and meta-analysis, *BMC Cancer*, 12, 2012, article 41.

[28] S. Pantazi, Y. Kagolovsky, and J. R. Moehr, Cluster Analysis of Wisconsin Breast Cancer Dataset Using Self-Organizing Maps, *Studies in Health Technology and Informatics, 90*, 2002, 431-436.