Performance Analysis of Fuzzy K-Means and Fuzzy C-Means Algorithm on Breast Cancer Dataset

C.Sathya[1], V.Priya[2]
[1, 2] Department of Computer Science
[1] Vellalar College for Women(Autonomous),Erode-12
[2]Assistant Professor,Vellalar College for Women(Autonomous), Erode-12

Abstract-*A fuzzy k means clustering to deal with the problem where the points are somewhat in between centers or otherwise ambiguous by replacing distance with probability relative to the inverse of the distance. Fuzzy k-means uses a weighted centroid based on those probabilities. The resulting clusters are best analyzed as probabilistic distributions rather than a hard assignment of labels. The main focus of this paper is to analyze data mining techniques required for breast cancer data especially to discover the patient details like Id number, Clump thickness, Uniformity cell size, Uniformity cell shape, Marginal adhesion,Single Epithelial cell size, Bare Nuclei country, Bland Chromatin, Normal Nucleoli, and Mitoses. In this work, the fuzzy k-means and fuzzy c-means algorithm is used for clustering the given data. Here, breast cancer data set is used, it contains patient details grouping the cluster of cancer and non cancer dataset. The fuzzy k-means and fuzzy c-means algorithm mines the data based on input details can be cluster the value. In experimental result fuzzy k-means algorithm means occurs better accuracy than fuzzy c-means algorithm.*

Keywords-fuzzy k-means algorithm, fuzzy c-means algorithm, data set, clustering

.

I. INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve.

Clustering analysis plays an important role in the data mining field, it is a method of clustering objects or patterns into several groups. It attempts to organize unlabeled input objects into clusters or “natural groups” such that data points within a cluster are more similar to each other than those belonging to different clusters, i.e., to maximize the intra-cluster similarity while minimizing the inter-cluster similarity.

An unavoidable fact of data mining is that the subsets of data being analyzed may not be representative of the whole domain, and therefore may not contain examples of certain critical relationships and behaviours that exist across other parts of the domain. To address this sort of issue, the analysis may be augmented using experiment-based and other approaches, such as Choice Modelling for human-generated data. In these situations, inherent correlations can be either controlled for, or removed altogether, during the construction of the experimental design.

Breast cancer has become the leading cause of death in women in developed countries. The most effective way to reduce breast cancer deaths is detect it earlier. Early diagnosis requires an accurate and reliable diagnosis procedure that allows physicians to distinguish benign breast tumors from malignant ones without going for surgical biopsy. The objective of these predictions is to assign patients to either a ”benign” group that is noncancerous or a ”malignant” group that is cancerous. The prognosis problem is the long-term outlook for the disease for patients whose cancer has been surgically removed. In this problem a patient is classified as a ‘recur’ if the disease is observed at some subsequent time to tumor excision and a patient for whom cancer has not recurred and may never recur. The objective of these predictions is to handle cases for which cancer has not recurred (censored data) as well as case for which cancer has recurred at a specific time.

II. RELATED WORKS

The classifier accuracy has been surely enhanced by the use of any of Feature selection method than the classifier accuracy without feature selection. The performance of Decision tree classifier-CART with and without feature selection in terms of accuracy, time to build a model and size of the tree on various Breast Cancer Datasets are observed. From the results it is clear that, though we considered only breast cancer datasets, a specific feature selection may not lead to the best accuracy for all Breast Cancer Datasets.

The analyse the breast Cancer data available from the Wisconsin dataset from UCI machine learning with the aim of developing accurate prediction models for breast cancer using data mining techniques. In this experiment, we compare three classification techniques in Weka software and comparison results show that Sequential Minimal Optimization (SMO) has higher prediction accuracy i.e. 96.2% than IBK and BF Tree methods.SMO classifier is suggested for diagnosis of Breast Cancer disease based classification to get better results with accuracy, low error rate and performance.The gathered data is preprocessed, fed into the database and classified to yield significant patterns using decision tree algorithm. Then the data is clustered using Kmeans clustering algorithm to separate cancer and non cancer patient data. Further the cancer cluster is subdivided into six clusters. a diagnosis system for detecting breast cancer based on RepTree, RBF Network and Simple Logistic. In test stage, 10-fold cross validation method was applied to the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia database to evaluate the proposed system performances.

The correct classification rate of proposed system is 74.5%. This research demonstrated that the Simple Logistic can be used for reducing the dimension of feature space and proposed Rep Tree and RBF Network model can be used to obtain fast automatic diagnostic systems for other diseases.The total time taken to build the model is at 0.62 seconds. These results suggest that among the machine learning algorithm tested, Simple logistic classifier has the potential to significantly improve the conventional classification methods used in the study.

The prediction of survivability rate of breast cancer patients using data mining techniques. The data used is the SEER Surveillance, Epidemiology, and End Results Public-Use Data. The preprocessed data set consists of 151,886 records, which have all the available 16 fields from the SEER database. The outcome of three data mining techniques: the Naïve Bayes, the back-propagated neural network, and the C4.5 decision tree algorithms.

III. FUZZY MEANS ALGORITHM

Clustering is an automatic learning technique aimed at grouping a set of objects into subsets or clusters. The goal is to create clusters that are coherent internally, but substantially different from each other. Automatic document clustering has played an important role in many fields like information retrieval, data mining, etc

.

Fuzzy K-Means Algorithm

The Fuzzy K-means algorithm uses Gaussian weighted feature vectors to represent the prototypes. The algorithm starts with large numbers of clusters and eliminates clusters by distance or by size so the prototypes are much more representative. The algorithm computes fuzzy weightequation (1).

The FKM algorithm does the following steps:

- Standardize the Q sample feature vectors and use a large K_{init}.
- Eliminate the prototypes that are closer to other prototype than a distance threshold D_{Thresh}, which can be set by a user.
- Apply k-means as the first step to get the prototypes.
- Eliminate small clusters.
- Loop in computing the fuzzy weights and MWFEV (modified weighted fuzzy expected value) for each cluster to obtain the prototype and then assign all of the feature vectors to clusters based on the minimum distance assignment. End the loop when the fuzzy centers do not change.
- Merge clusters.

Initial K. The FuzzyM algorithm uses a relatively large K_{init} to thin out the prototypes. The default K_{init} is calculated as:

$$K_{init} = \max\{6N + 12\log_2 Q,\ Q\}$$

Clustering. To obtain a more typical vector to represent a cluster, the algorithm uses modified weighted fuzzy expected value as the prototypical value:

$$\boldsymbol{\mu}^{(r+1)} = \sum_{\{p=1,\,P\}} \alpha_p^{(r)} \mathbf{x}_p$$

Where$\boldsymbol{\mu}^{(r+1)}$ is obtained by Picard iterations. The initial value $\mu^{(0)}$ is the arithmetic average of the set of real vectors in equation 1.

Where σ^2 is the mean-square error.the value $\mu = \mu^{\infty}$ to which this process converges is our modified weighted fuzzy expected value(MWFEV).

The σ^2 is the weighted fuzzy variance (WFV). $\mu^{(0)}$of a set real values $x_1, \ldots . x_p$ as the initial center value.

The process computes the fuzzy weights and the componentwisely for each cluster to obtain the cluster center. It also computes the mean-square error σ^2 for each cluster, and

then assigns each feature vector to a cluster by the minimum distance assignment principle.

Fuzzy C-Means Algorithm

Fuzzy clustering (also referred to as soft clustering), data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between the data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters.

The Fuzzy C-means algorithm is one of the most widely used fuzzy clustering algorithms. This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center.

Clearly, summation of membership of each data point should be equal to one. After each iteration membership and cluster centers are updated according to the formula in equation 2, 3, 4.

1.Initialization: $C_{opt} = c \leftarrow c_{max}$.
2. Apply FCM to the data set to update the clustercentersv*i* and the membership values μ_{ik}.
Do iteration and test for convergence; if not goto 2.

4.if$(c = c_{max})\{ \alpha = Dis(c_{max})$; indexValue =$V_{cwb}(c)\}$
else if ($V_{CWB}(c)$<indexValue). $\{C_{opt} \leftarrow c$; indexValue = $V_{CWB}(c)$; }

5. c ← $c - 1$, if $(c = c_{min} - 1)$stop else goto 2.

The explanation of fcm validation algorithm are,

- After step 3 the fuzzy partition is generated.
- This partition is validated by the Vcwb in step 4.
- The parameter “index value” stores the minimum value of Vcwb. After step 5 the optimal numberof clusters Copt will be found that corresponds to the minimum value of the index value Vcwb within the range of $[C_{\min}, C_{max}]$.

IV. METHODOLOGY AND RESULTS

The breast cancer is a common disease in women .Current work is based on detecting breast cancer using data mining technique. The Fuzzy c means and Fuzzy k means algorithm is used to detect breast cancer. This Chapter describes about process in the current work.

The Architectural diagram of the proposed system is depicted in the figure 4.1.

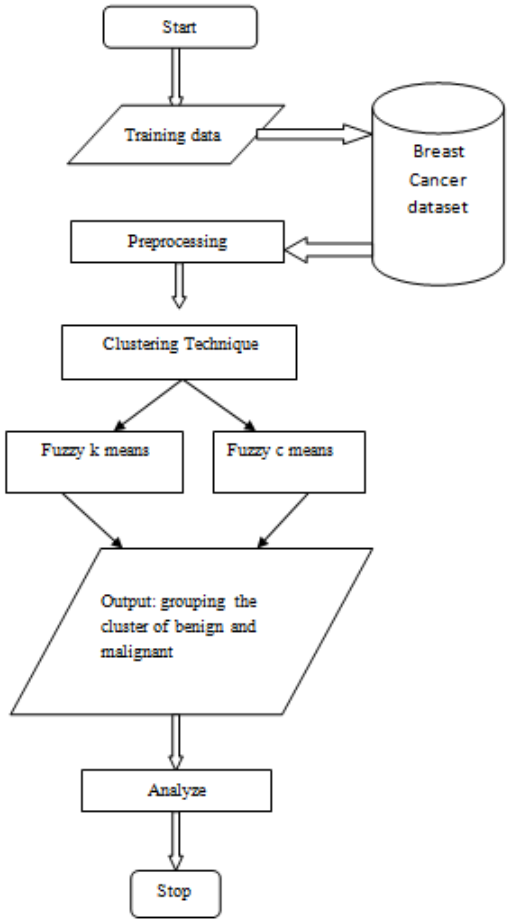

Figure 4.1 Flowchart diagram

The training data of breast cancer dataset is to be preprocessed. In this research work, the dataset is preprocessed to remove the null value in the dataset using median filter. There are many types of data mining techniques in that use clustering method. The proposed work use two algorithm fuzzy c-means and fuzzy k-means using clustering techniques. The results of two algorithms are grouping the cluster of cancer and non cancer. The performance of fuzzy c-means and fuzzy k-means is analyzed and the results can be significant gain in fuzzy k-means with good accuracy then fuzzy c-means.

The Weights in the FKM (Gaussian Weights)

Figure 4.2 shows that the weight of Gaussian is decided not only by the distance between the feature vector and the prototype, but also the distribution of the feature vectors in the cluster. When the feature vectors in a cluster are

centrally and densely distributed around the center, the sigma value will be small. The features that are close to prototype weigh much more than the farther features.

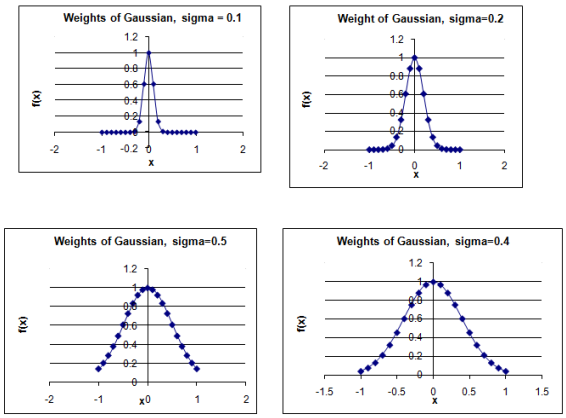

Figure 4.2 Weights of Gaussian for Different Sigma Values of Fuzzy k-means

The Weights in the FCM (Gaussian Weights)

In the FCM algorithm, the weights are defined by above Equation. Figure 4.3 shows that the weight of Gaussian is decided not only by the distance between the feature vector and the prototype, but also the distribution of the feature vectors in the cluster. When the feature vectors in a cluster are centrally and densely distributed around the center, the sigma value will be small. The features that are close to prototype weigh much more than the farther features.

The weight of a feature vector is related to the shape of the Gaussian. If a feature vector has equal distance from two prototypes, it weighs more on the more widely distributed cluster than on the centrally located cluster. Thus, Gaussian fuzzy weights are more immune to outliers and more representative than the other kind of fuzzy weight.

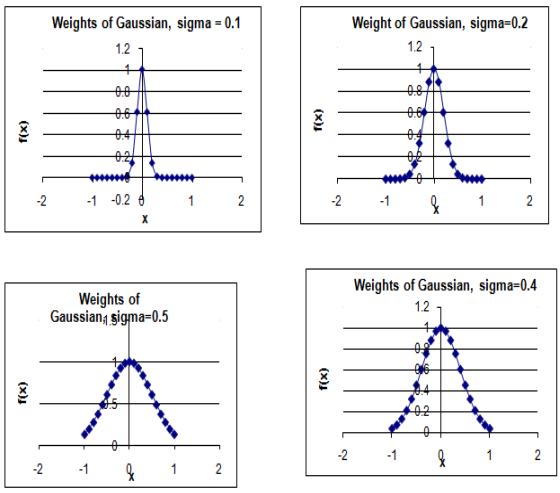

Figure 4.3 Weights of Gaussian for Different Sigma ValuesOf Fuzzy C-means

ACCURACY

Accuracy is the important parameter for analysis the any process it help to find the quality of our output here we give the formula to find the accuracy of the process.

$$Accuracy \frac{Noofcorrectclassifieddata}{Totaldata} \times 100$$

Where, no of correct classified data is correctly classify in given data 699 dataset

(eg: $\frac{540}{699}$ X 100)

Total data is total number of data in dataset 100 is constant value.

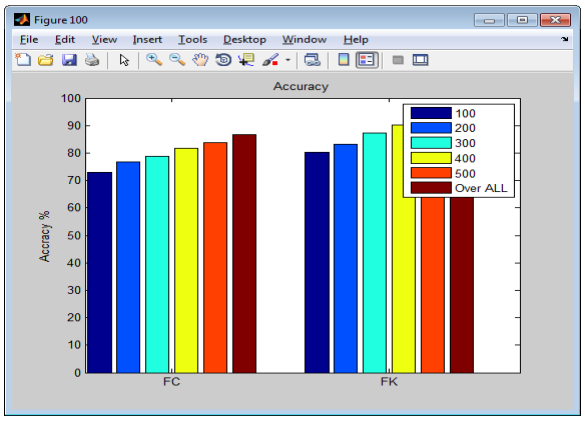

Fig 4.4 Accuracy

F-MEASURE

F-measure is a measure of a test's accuracy. It consists both the precision and the recall of the compute the score.

$$f\ measure = 2 * \frac{Precision * recall}{Precision + Recall}$$

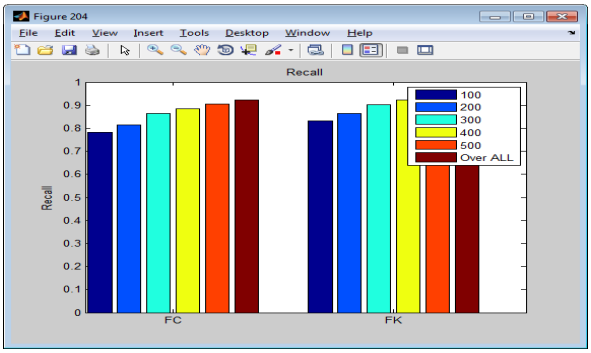

Fig 4.5 Sensitivity

ROC Curve

ROC curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

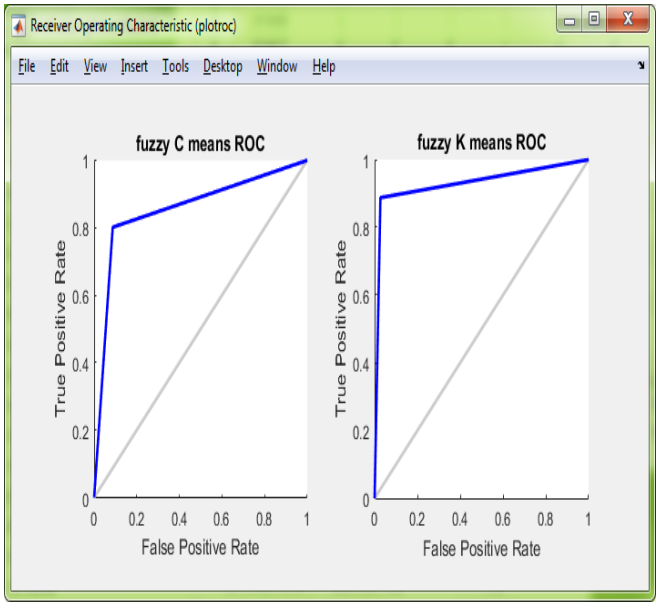

Fig 4.6 ROC curve

Comparison Table

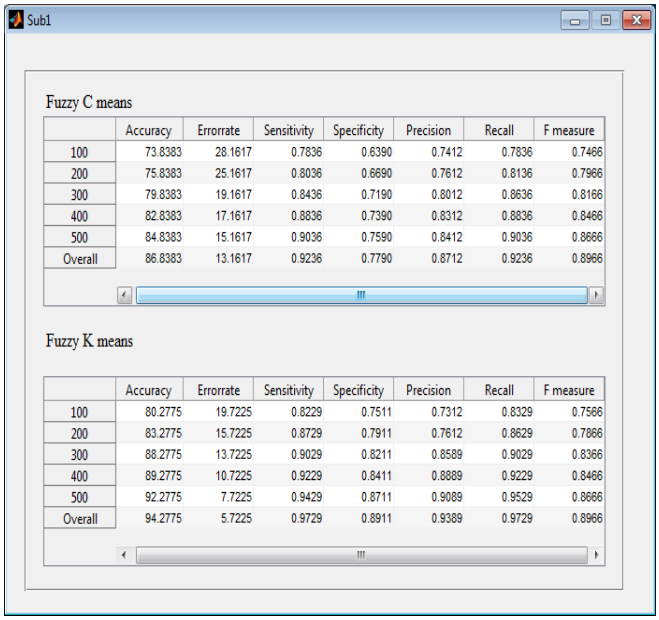

Fig 4.7 Comparison Table

V. CONCLUSION

In this paper, two novel algorithms are developed to speed up fuzzy k-means clustering through using the information of center displacement between two successive partition processes. A cluster center estimation algorithm is also presented to determine the initial cluster centers for the proposed algorithm fuzzy c-means and fuzzy k-means. Experimental results show that the proposed approaches and fuzzy k-means can obtain the same clustering result. The proposed methods are used to reduce the computational complexity of conventional fuzzy c-means clustering. Therefore the Euclidean distance is used as the distortion measure. The result shows that the fuzzyk-means algorithm increase accuracy of process during the cluster the breast cancer dataset.

REFERENCES

[1] AbdelghaniBellaachia, ErhanGuven, "Predicting Breast Cancer Survivability Using Data Mining Techniques", 2009.

[2] D. Auber, and M. Delest, "A clustering algorithm for huge trees," Advances in Applied Mathematics, vol. 31, no. 1 pp. 46-60 2003.

[3] VikasChaurasia, Saurabh Pal, "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivablity" International of Computer Science and Mobile Computing, Vol. 3, Issue 1, January 2014, Pg. 10 -22

[4] VikasChaurasia, Saurabh Pal, " A Novel Approach for Breast Cancer Detection using Data Mining Techniques" International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2 Issue 1, January 2014

[5] D. Lavanya, Dr. K Usha Rani, "Analysis of Features Selection with Classification:Breast Cancer Datasets" Indian Journal of Computer Science and Engineering, Vol.2 No. 5, Nov 2011.

[6] A. Y. Al-Omary, and M. S. Jamil, "A new approach of clustering based machine-learning algorithm," Knowledge-Based Systems, vol.19, no.4, pp. 248-258, 2006.

[7] RonakSumbaly, N. Vishnusri, S. Jeyalatha, " Diagnosis of Breast Cancer using Decision Tree Data Mining Technique", International Journal of Computer Application, Vol 98 No 10,July 2014

[8] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: analysis and implementation," IEEE

Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 881-892, 2002.

[9] R. Radha, P. Rajendiran, " Using K Means Clustering Technique To Study of Breast Cancer", IEEE Transactions on World Congress on Computing and Communication Technologies, March 2014.

[10] M. Halkidi, and M. Vazirgiannis, "Clustering validity assessment: finding the optimal partitioning of a data set," Proceedings IEEE International Conference on Data Mining. 2001.

[11] Z.-H. Zhou, "Three perspectives of data mining," Artificial Intelligence, vol. 143, no. 1, pp. 139-146, 2003.

[12] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: an efficient data clustering method for very large databases," in Proceedings of the 1996 ACM SIGMOD international conference on Management of data, 1996, Montreal, Quebec, Canada.