Clustering and Classification of Cancer Data Using Soft Computing Technique

Mr.S.P.shukla and Mrs. Ritu Dwivedi

Abstract: *Clustering and classification of cancer data has been used with success in field of medical side. In this paper the two algorithm K-means and fuzzy C-means proposed for the comparison and find the accuracy of the result. this paper address the problem of learning to classify the cancer data with two different method and information derived from the training and testing .various soft computing based classification and show the comparison of classification technique and classification of this health care data .this paper present the accuracy of the result in cancer data.*
Keywords: *clustering, classification,*

I. Introduction

Cancer data classification and clustering have been the focus of critical research in the area of medical and artificial intelligence. Health care is now days very important for human being. Now everyone is health care unit of system are there to monitor and analyze health status of a human being .As we know health is wealth, if health of a particular is food individual will grow hence society and nation will go a health .This is the reason why soft computing based health care system is to be developed, which has proved it efficiency and performance with conventional system.

Cancer has become one of the major causes of mortality around the world and research into its diagnosis and treatment has become an important issue for the scientific community. The most important issue in classification and clustering of cancer data is deciding what criteria is to be classify against, for example suppose it is desirous to classify cancer disease in describing cancer one will look at its type ,spot, stages and duration and so on .many of these feature are fuzzy and qualitative in nature. For this classification some criterion is to be decided. One can classify cancer on the basis of its type, its human parts of manifestation i.e. mouth, thought, tongue, intestine, image, liver or similar other parts of the body.
The popular method of classification is very well-known as fuzzy C-means (FCM),so named because of its close analog in the crisp word, this method uses concept in n-dimensional Euclidean space to determine the geometric closeness or classes and the determining the distance between the clusters.

In this piece of research work two very important application of research work two very important application, classification and clustering are use on cancer data. It is well known that classification and clustering are the technique to separate same type of data together, classification is a supervisee way to separate same type of data to put similar type of data together. Classification and clustering technique can apply on cancer dataset and find the accuracy of classification and clustering.

The rest of the paper is organised as follows: The 2 section outlines the reason for using ear as biometric for newborn. This section is followed by details of database acquisition in section 3. Covariates of newborn ear is explained in section 4 followed by automated ear masking in section 5. The details of feature extraction and matching are explained in section 6 and this section also explains proposed methodology for ear recognition. Section 7 describes performance evaluation of different algorithms on newborn ear. Finally section 8 and 9 present future direction and key conclusion.

II. Cancer detection algorithm and concept

Our cancer detection system adopts a two FCM and K-means algorithms. In this process we show the class of cancer that mean cancer is benign (2) and malignant(4) .these frames are derived from UCI repository dataset. Which is use to compare the classification and clustering technique and find the accuracy of the result.

2.1 Data Description

In data description number if instances are 699(as of 15 July 1992) that are used for research work. This cancer data contain 10 attribute and their id number value between 1 to 10 .the last attribute of the data are class that has been moved to last column .attribute class are use two value 2 for benign and 4 for malignant.

2.2 MATLAB Software Working

The name MATLAB stands for matrix laboratory. MATLAB was originally written to provide easy access to matrix software developed by the LINPACK and EISPACK projects. Today, MATLAB engines

incorporate the LAPACK and BLAS libraries, embedding the state of the art in software for matrix computation. MATLAB is the tool of choice for high-productivity research, development, and analysis.
MATLAB Toolboxes are comprehensive collections of MATLAB functions (M-files) and our research paper is based on this function. In this research work data set convert in M-file, after the creating M-file MATLAB toolbox perform the comprehensive study of booth.

2.3 Clustering and classification with their   algorithm

Clustering can be considered the most important unsupervised learning problem; so as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

Classification same as classify the cancer data set but it is the type of supervised way and in this process training data has to specify what we are trying to learn so data classification is data reduction technique and data present in class from so classification contain the similar type value in group.

This research paper work are address to feed forward neural network are address to feed forward neural network which is work in based on supervised learning .this topic work in this technique ,that have three layer
- Input layer(multiple input data)
- Hidden layer(multiple or one layer)
- Output layer(one layer)

Classification side the three layer are available .first layer input layer have input data in matrix from ,second layer hidden layer ,some process feed forward neural network contain one layer and some contain multiple layer and last layer is output layer which is the resultant layer.
In clustering process the output layer will be hiding so in this condition that process is unsupervised process. Feed forward neural network use some activation function like
- sigmoid activation function

$$f(x) = \frac{1}{1 + e^{-x}}$$

hyperbolic activation function

$$f(x) = \frac{e^{2x} - 1}{e^{2x} + 1}$$

- linear activation function

$$f(x) = x$$

Data classification in clustering side use FCM algorithm .it converts the input matrix in output from.

III.   Experimental Work for Cancer Detection

In the present work the soft computing technique one used for cancer affected object classification purpose .the soft computing technique derived their power due to their clustering and classification an ability to learn in experiment. In experiment work neural network related classification can be used for fairly accurate classification of input data into categories, provided they are previously trained to do so .the accuracy of the classification depend on the efficiency of training. The knowledge gained by the learning experience is stored in the form of connection weights.
The issues need to be settled in designing an ANN for specific application topology of the network training algorithm neuron activation function with weights and bias.
In our topology, the number of neurons in the input layer is 9 by the ANN classifier. The output layer was determined by the number of the class designed .the output are type1 therefore; the output layer of consist of one neurons. The hidden layer is consisting of 12 neurons. Before the training process is started, all the weights and bias are initialized. The training set used LEARNGDM adoption learning function and TRAINLM training function it is work in three layer and after the processing the experimental graph show the 100epochs .in this experiment work ,the training set was formed by choosing 160 data set for the testing process.
Cancer data is classify into one of two type object using the feed forward neural network classification in error back propagation algorithm .after the classification of the cancer object correct and incorrect classification are computer. The next step of classification algorithm is creating the performance matrix.
Same as work perform in clustering side and classify the cancer data find the accuracy of data set but in clustering side only two layer working network because it in type of unsupervised learning so the output layer are hide and find the accuracy of cancer data use fuzzy C-means algorithm and create the performance matrix .

IV.   Training and Testing
The proposed network was trained with 240 data samples. These 240 samples are fed to the network with 9 input neurons, one hidden layer of 12 neurons and one output neuron.MATLAB software version 8 is used to implement the software in current work. When the training process is completed for the training data set the last weights of the network were saved to be ready for the testing process.

The testing process in done for 160 samples, the 160 samples are fed to the proposed network and their output is recorded for calculating of accuracy of data.

In second type clustering use 240 training data set only. It isn't work for testing and the accuracy of data is less with the comparison of classification and finds the performance matrix.

V.   Data Accuracy and Performance
The accuracy of cancer detected data was evaluated by computing the percentages of right classified cancer data .in classification data show the simple number 2 for benign and 4 foe malignant in training and `testing so we remake it data are classify or misclassify when target and actual class are same or differ

The related confusion matrix show the result of EBPA network after training and testing 160 out of 165 samples of benign class are classified correctly while 5 samples are misclassified similarly 73 out of 75 sample of malignant class are classified correctly while 02 samples are misclassified similarly in testing process perform in 160 samples.

In clustering only input the data value and match the target and work the membership function in C1 and C2 (C1 and C2 are benign and malignant class) if output value are high in C1 class for example data membership in C1 is 0.9746 and C2 is 0.0053 so data belong to benign class.

VI.   Results and Discussion
Figure 1show the training curve with 100 epochs and figure 2 show the bar chart of performance matrix after the comparison between training and testing session the overall performance of classification are decrease in testing session .in training session correct classification of sample are 96.96% and 97.33% but in testing session the classification parameter are reduced and the percent are 94.54% and 94.00%.
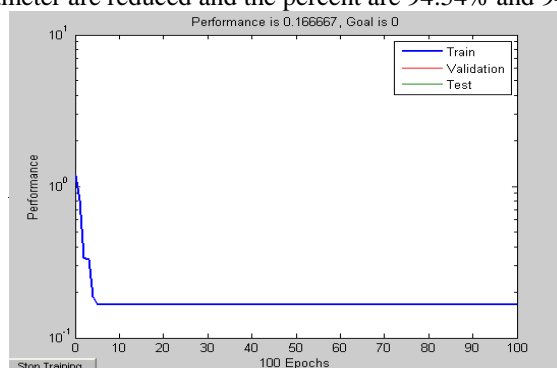

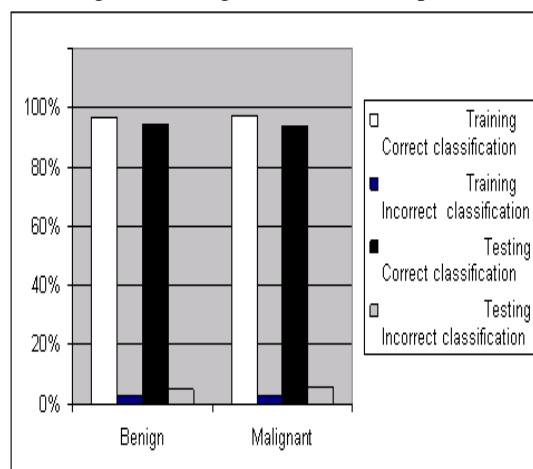Fig 1: Training curve with 100 epochs

Fig 2: bar chart of performance matrix

Fig 3 show the clustering of cancer data after applying FCM algorithm .In this session data perform only training so in which 156 out of 165 sample of benign class are classified correctly which 09 samples are miss-

classified. Similarly 71 out of 75 samples of malignant class are classified correctly which 04 samples are misclassified.
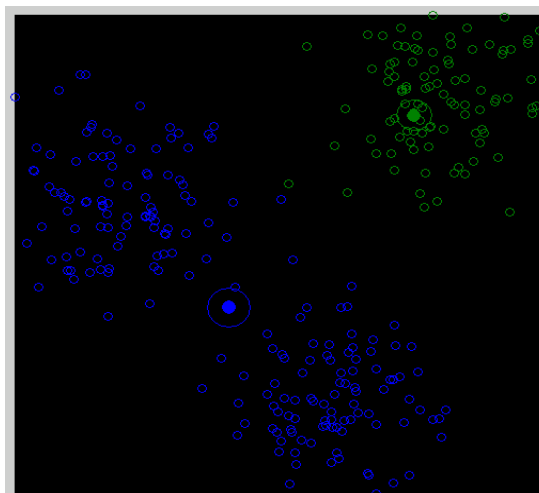

Fig 3: clustering of cancer data after applying FCM algorithm

At last comparison between both EBPA and FCM as simulated above the result is tabulated in table 1.1 from which it is clear that correct classified % in case of EBPA is 97.13% which in case of FCM it is 94.03% which clearly indicate that EBPA algorithm is performing well for classification of cancer related health cancer data.

Table 1.1: Comparison table

Class	EBPA		FCM	
	Correct	Incorrect	Correct	Incorrect
Benign	96.96%	3.04%	94.54%	5.46%
Malignant	97.33%	2.67%	94.66%	5.34%
Average	97.14%	2.85%	94.6%	5.4%

VII.   Conclusions

This paper presented a clustering and classification method for classify the cancer data and find their accuracy .this paper is compare on clustering and classification of soft computing with the area of health care data i.e. cancer data .As a comparison research that author of the current dissertation took bi-direction approach to the problem .In one direction the research studies the supervised manner on classification .the approach lead to constriction of intelligent, less error high performance network due to feed forward and layer architecture of paper.

In second direction, the research studied the unsupervised manner on clustering the approach lead to constriction of perceptional system which is based on fuzzy logic. In this paper exposed the problem of the result and proposed the solution of system by pointing out the attributes of cancer data.

Supervised and unsupervised are the two apposite techniques for classification of data but with the help of MATLAB software 8 used to implement the software in the current work. Supervised need both only input pattern. The EBPA and FCM are compared in terms of performance .the EBPA performance accuracy is 97.14% which in case of FCM accuracy is 94.6% which in case of FCM is having low performance due to unsupervised manner of classification.

References

[1]   MATLAB software in URL address:\\www.mathworks.com\\The Math Works,
[2]   Using Cancer data set from UCI repository data set ,the URL address: \\WWW.UCI.com\\
[3]   George j.klir / boyuan "fuzzy set and fuzzy logic" theory and application, year 2003, pages 50-61,
[4]   MATLAB software in URL address: \\www.mathworks.com\\The Math Works, MATLAB 7.5.0(R2007b) help file.
[5]   Zadeh, Lotfi A., "Fuzzy Logic, Neural Networks, and Soft Computing," Communications of the ACM, March 1994, Vol. 37 No. 3, pages 77-84.
[6]   Takagi, H.: "Fusion Technology of Fuzzy Theory and Neural Networks: Survey and Future Directions" IIZUKA90: International Conference on Fuzzy Logic and Neural Networks. pp. 13-26, Iizuka, Japan 1990.
[7]   Tanaka, Makoto: "Application of The Neural Network and Fuzzy Logic to The Rotating Machine Diagnosis" Fusion of Neural Networks, Fuzzy Sets, and Genetic Algorithms: Industrial Applications. CRC Press LLC, CRC Press LLC, Boca Raton, FL, USA 1999.
[8]   Lee, S. and E. Lee: "Fuzzy Sets and Neural Networks" Journal of Cybernetics. Volume 4, No. 2, pp. 83-013, 1974.
[9]   Zadeh, Lotfi: "The Role of Soft Computing and Fuzzy Logic in the Conception, Design, Development of Intelligent Systems" Plenary Speaker, Proceedings of the International Workshop on soft Computing Industry. Muroran, Japan, 1996.
[10]   Zadeh, Lotfi: "What is Soft Computing" Soft Computing. Springer-Verlag Germany/USA 1997.
[11]   Kacpzyk, Janusz (Editor): Advances in Soft Computing. Springer-Verlag, Heidelberg, Germany, 2001.
[12]   Learning internal representations by error propagation by Rumelhart, Hinton and Williams (1986).

[13] Vikram Chandramohan and Tuan D. Pham James Cook University School of Math, Physics and IT. Of published " Cancer Classification using Kernelized Fuzzy C-means" research paper

[14] Xiao Ying Wang, Jon Garibaldi, Turhan Ozen Department of Computer Science and Information Technology ,The University of Nottingham, United Kingdom of published " Application of the Fuzzy C-Means Clustering Method on the Analysis of non Preprocessed FTIR Data for Cancer Diagnosis " Research paper.

[15] Dave Anderson and George McNeill Kaman "ARTIFICIAL NEURAL NETWORKS TECHNOLOGY" developed by Sciences Corporation address of 258 Geneses Street Utica, New York

[16] Aleksander, Igor and H. Morton: An Introduction to Neural Computing Chapman and Hall, London, UK 1990

[17] Bonissone, Piero: "Soft Computing: The Convergence of Emerging Reasoning Technologies" Soft Computing. Springer-Verlag, Germany/USA 1997.

[18] Lee, S. and E. Lee: "Fuzzy Sets and Neural Networks" Journal of Cybernetics. Volume 4, No. 2, pp. 83-013, 1974.

[19] Gurney, Kevin: An Introduction to Neural Networks. UCL Press, London, UK 1999.

[20] Fausett, Laurene: Fundamentals of Neural Networks: Architectures, Algorithms, and Applications. Prentice Hall, NJ, USA 1994.

[21] Hertz, J.A., Krogh, A. & Palmer, R. Introduction to the Theory of Neural Computation (Addison-Wesley, Redwood City, 1991)