Detection of Breast Cancer Progress Using Adaptive Nero Fuzzy Inference System and Data Mining Techniques

Hengameh Mahdavi

Faculty of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran

Abstract

Prediction, diagnosis, recovery and recurrence of the breast cancer among the patients are always one of the most important challenges for explorers and scientists. Nowadays by using of the bioinformatics sciences, these challenges can be eliminated by using of the previous information of patients records. In this paper has been used adaptive nero fuzzy inference system and data mining techniques for processing of input data and the educational combined algorithm for arranging of parameters of input functions. It has used also the downward gradient algorithm for arranging of unlined input parameters and the algorithm of the least of squares for arranging of lined output parameters. It has been used the data the institute of oncology Ljubljana of Yugoslavia that contain the information of 1090 the breast cancer patients. The results show the suggesting system has 89% accuracy in the diagnosis of progressing the breast cancer, which has improved by compared with neural network classification method.

Keywords: Clustering, Classification, ANFIS.

1. Introduction

The breast cancer is considered as one of the most important and dangerous cancers among women and is known the second factor for the women death. The diagnosis of it means separation of malignant glands from benign tumours. This paper has been provided by using of the data techniques for increasing diagnosis and predicting it. The data can be a good guider for physicians in predicting of the breast cancer return. We can increase the number of records with the mixing of several data bases and by using of the more applied variables that are important in diagnosis cause the accuracy of the data mining methods.

2. Literate Review

Dan and colleagues have used the tree of making decision and logistic regression and the artificial neural networks for developing of the predicting models for the breast cancer by the analysis of the great data bases that have been taken from Wisconsin data base. Their research results show that the tree algorithm of making decision has priority on the other methods in extracting the information of the current data[1]. BI and FU YANG have used from the back up machine for finding patterns of diagnosis and found out that this method is a good way for diagnosis of the breast cancer patterns and these results were

* Corresponding author. Email: Mahdavi.hh@gmail.com

compatible with reality [2]. Land in and col have used the logistic regression and the artificial neural networks for predicting of 5, 10, and 15 years for surviving patients. They have considered 951 patients and have regarded the size of tumour, a kind of cell and the age as the input variables[3]. Bend Harker and col have used several methods for considering of the breast cancer patterns. They found out that the data can be used as a precious device in recognizing of similarities with the aim of pre knowing, diagnosis and healing[4]. Tolouii and col have used 3 techniques including the trees of making decision, the supporter vector machines and the techniques of artificial neural networks for predicting of returning breast cancer. Kenar kouhi has considered using of the new software of comparative ANFIS for predicting of causing cancer power of the human papilloma virus.

Harlina and col (2010) have considered the using of ANFIS for diagnosis of surviving in the patients.

Sung and col (2005) in their research have used a hybrid system for diagnosis of the breast cancer.

3. The Suggested Method

As it has been said before, the used data have been taken from the institute of oncology Ljubljana that contain information of 1000 patients that the cancer is being returned in 800 case. The data is about 9 patient s appearance features that have been placed in two groups including the in progress and restored patience. In this paper it has been tried to use ANFIS method for the clustering the patients and has been used for this regard 60% data for education and 40% for evaluating. The table 1 shows the patient s features including nominal and quantative features. These features include: the age period, the postmenopausal conditions, the size of tumour, inv-nodes, deg-malig, the left or right member, the location of glands in member, irradiate. For using of clustering method we need the labelling of nominal variables as quantative. For example in these features we use 1 instead of ‘Yes’ and 0 instead of ‘No’. or for describing the gland place we use 1 for the left-up and 2 for the left-down. It is necessary to appropriate one number instead of the quantative variables and this number can be an average for every data mining. so in this phase we start labelling on the variables for preparing data to clustering and ultimately the data will be provided in a matrix 10*1090.

In the second phase education of network will be taught. The best structure for network will be chose focusing the control data. In the last phase the result will be evaluated by the test data that include 40% of the whole data. The best structure means determining the numbers and the kind of functions membership in every input and the number of returning for achieving the best model.

Table 1

patient’s information about their features.

Name	Type
Age	10-19,20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 0-89, 90-99.
Menopause	It40, ge40, premeno
Tumour-size	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59
Inv-nodes	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35,36-39.
Node-caps	Yes, no
Deg-malig	1,2,3
Breast	Left, right
Breast-quad	Left-up, left-low, right-up, right low. Central
Irradiate	Yes, no
class	No- recurrence- events, recurrence events.

Briefly, the first layer in the ANFIS structure is the fuzzy system and the second layer is fuzzy AND that performs the fuzzy rules. in the third layer making common of membership functions will be done. The fourth layer is the final section of the fuzzy rules and eventually the last output layer of network will be estimated. There are the whole parameters of the comparative and following hypothesis in the ANFIS structure. For achieving to these parameters two passes should be stepped: in the first step that named forward pass, the whole comparative parameters are considered stable and the whole following parameters will be estimated by the least square error. In the second step that named the backward pass, the whole following parameters are considered stable and the whole following parameters will be estimated by the gradient descent. These operations that are being done in every phase of education, is named the epoch. With calculating of the model parameters, the output amount of model is achieved on based on the arranged couples that with name education have been gave to the model. This predicted amount will be compared with the real amount so the function of education error of ANFIS will be estimated. In the follow the structure and planning of installed model has been showed for the whole patient's data. This system is educated by 9 input (features) and 1 output (the aim) that all of them have been showed in Table 1.

In this phase a model will be needed that can extract the best pattern from the recorded data. TSK for operating in function approximation and the high accuracy in reaching to final answer is chose. The educational algorithm is combination of error back propagation for arranging the input data and least square estimator algorithm for arranging the output parameters. The figure 1 shows the ANFIS model the kind of TSK.

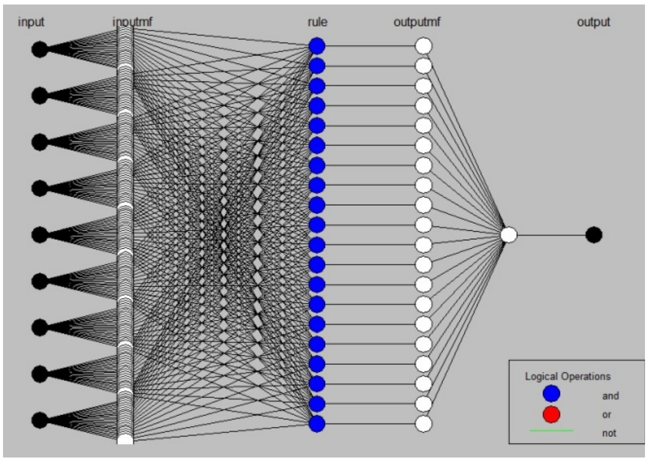

Fig. 1. TSK type model ANFIS

For every recorded data, the whole data will be normalized in the compass of -1 to +1. Then the more percent of data (60%) will be appropriated for training data and the rest for the test data. In this setting phase is used from the subtract clustering methods and FCM algorithm that these methods at first will regard and determine one rule for each cluster by the data clustering in several clear data. The number of gauss functions for every dimension of inputs in these methods is determined by the number of clusters spontaneously. The distance criteria in these algorithms are the Euclidean distance. After determining of the number of training and test data, the model will be educated by the combined algorithm that is a criterion for the average of mean square error. For evaluating the model operation, the method of fold errors validations have been used that for every recorded data in 10 times performance is divided to 60% training data and 40% test data. Also the criterion of MSE and R are used for validation of the model operation rate that is defined as the Equation1:

$$MSE = \frac{\sum_{m=1}^{n}(y_{pre,m} - t_{mea,m})^2}{n} \quad (1)$$

In the Equation 1 $y_{pre,m}$, m is the determined amount for the number 'm' and $t_{mea,m}$ is expected amount for the number m data. In fact this Equation is the differences average between the expected amounts and obtained amounts from the model. The Equation of the R accuracy rate is defined as the Equation 2:

$$R = \frac{\sum_{m=1}^{n} (y_{pre,m} - t_{mea,m})^2}{\sum_{m=1}^{n} (t_{mea,m})^2} \qquad (2)$$

In the Equation 2 $y_{pre,m}$ is the obtained amount for watching of number 'm' and $t_{mea,m}$ is favoured amount for the number m watching and n is the total of data, the obtained amount of the above Equation is referring to the accuracy rate of the method in determination process that is determined between 0 to 1. If this difference is more, it means the less being different of model in estimating of output amounts so we succeed to do the diagnosis according to the favoured amounts. In this model the gauss membership function is used for the system of fuzzy conclusion that its Equation is so:

$$f(x,\sigma,c) = e^{\frac{-(x-c)^2}{2\sigma^2}} \qquad (3)$$

In this Equation x is input, a and c are parameters that should be arranged that in this model they are arranged by using of the after spreading error algorithm for every input and their changes.

For considering and showing of the obtained results of model, a various kinds of test will be done on the model that by depicting of the output vector of model and wanted output that we can watch the model difference with the real amounts. In the figure 2, the ANFIS model diagram is showed before the education of MSE error diagram:

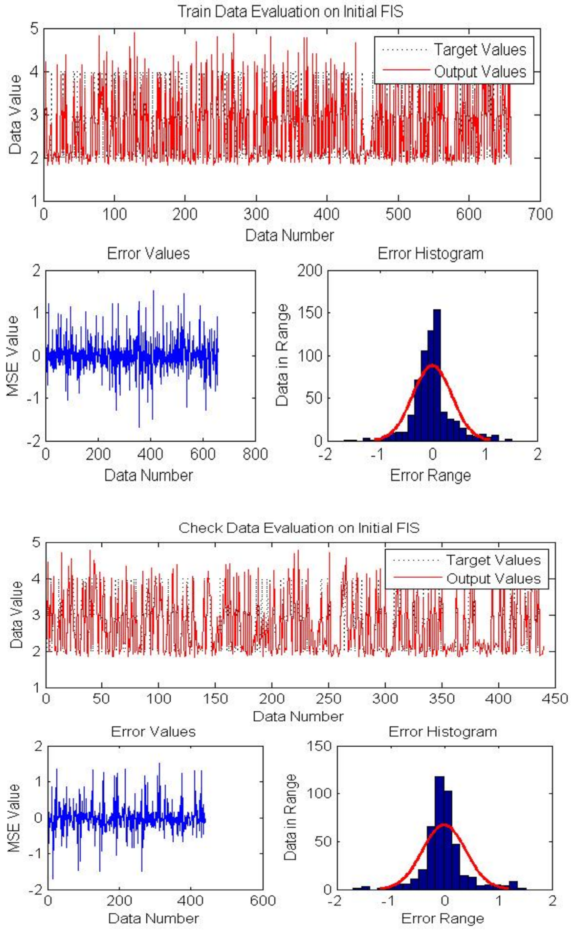

Fig. 2. Diagram ANFIS model for training data to training and ANFIS model diagram

As it is obvious, the MSE error has been so high before the model education an account of not being arranged of parameters of membership functions for inputs and outputs and has changed between 0 to 2/5.this error has been decreased by the model education and it's effect on parameters. In the figure 3 the ANFIS diagrams has been showed for the educational and testing data:

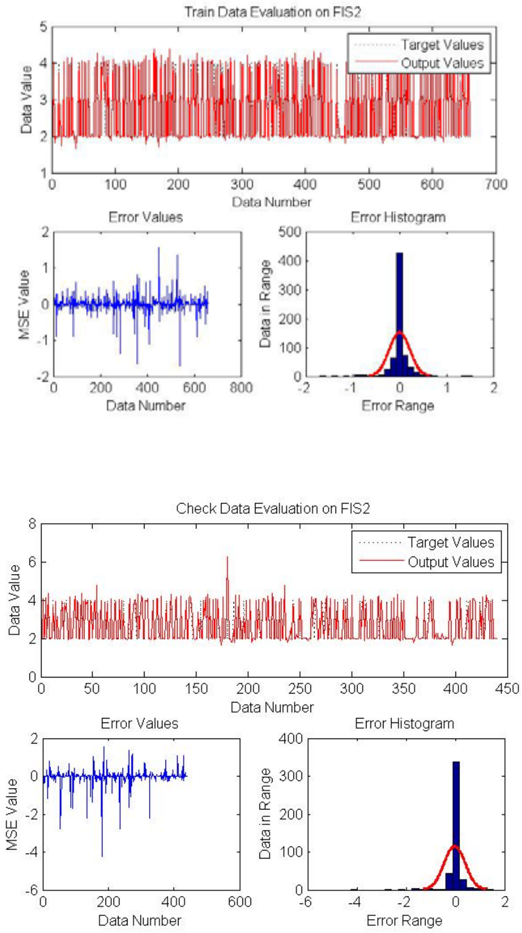

Fig. 3. Diagram ANFIS model for training data after the training, after the training charts ANFIS model to experimental data

As we can see in the figure 3, determining error for the educational and testing data has been less after the education. Also we can find by the histogram error diagram that the most error amounts have been occurred on the zero error that proves this method being suitable in the diagnosis process.

Table 2

The error values in different modes of education model

ANFIS Model	MSE Train	MSE Test
Initial FIS	0.1680	112.2
BP-Train	0.0280	5.236
BP-LSE-Train	0.0112	0.1828

In the table 2, error for educational and testing data in the case that model has not been educated, has been offered.

As it is obvious in the table 2 when the model has not been educated, the MSE error had been so high for both educational and testing data. But after education this error has been decreased that the difference between the favoured output and obtained output for educational data is 0/0128 and for testing data is 0/1828.

For depicting of the educational parameters effect on the model, different amounts of parameters of clusters numbers in the FCM method and parameter of the clusters center in additional algorithm of clustering have been showed in the table 3.

When the RAD parameter is less, it means that distance between clusters centers is more. So the number of clusters is more and this will causes increasing of rules number of if-when in the total of the fuzzy rules. But we cannot always expect that the less amount of this parameter is, the higher accuracy of model is because it may cause the problem of dimensions higher processing by extreme increasing of model rules.

Table 3

The effect of education parameters on the error and accuracy rate in the model

Algorithm	FCM N Cluster	Learning Rate	MSE	R
BP	-	0.9	.04475	0.8754
BP-LSE	-	0.8	0.0252	0.8825
BP	10	0.9	0.0208	0.8878
BP-LSE	20	0.8	0.0156	0.8998

RAD Parameter shows the effect rate of classifications centers in determining of output and input dimensions. Also N-Cluster in FCM algorithm determines the number of clusters that data should be clustered in it.

4. Comparison with the Similar Methods

To evaluate the suggested method, we have compared the results obtained with those of similar methods such as neural network and naïve Bayes methods. The result of these comparisons were summarized and put into Table4.

Table 4

The results of the comparisons done

Total Accuracy	*Criterion MSE*	*The evaluation method*
97.80%	*0.0447*	*BP*
98.07%	*0.0156*	*BP-LSE*
96.25%	*0.0234*	*ANN-Classifier*
94.03%	*0.0513*	*NB-Classifier*

As it is obvious in the table4, the suggested method has offered better results than the similar methods so the total accuracy has improved 3% in comparison with the other mentioned methods and the MSE criterion has also been more accurate as 4%.

5. Results

In this paper ANFIS has been used for diagnosis of the cancer patients. This method focuses on the ANFIS and the effect of train and test data on the results. This method also uses from the different parameters obtained from patients. In this modelling the fuzzy neural system of TSK model of the first grade with gauss membership functions for input data has been used. The downward gradient algorithm has been used for arranging of unlined output parameters and the least square algorithm has been used for arranging of lined output parameters. Results from modelling prove the high quality method in diagnosis that the accuracy in this method has been reached to 89% whereas it has been reached 46% by using of the neural network.

References

[1] D. Delen, G. Walker, A. Kadam. Predicting breast cancer survivability: a comparison of three data mining methods. J. Artificial Intelligence in Medicine 2010.

[2] W . Yi, W . Fuyong. Breast cancer diagnosis via support vector machines. 2007

[3] M. Lundin, J. Lundin, HB. Burke, S. Toikkanen, l. Pylkkanen, H. Joensuu. Artificial neural networks applied to survival prediction in breast cancer.Oncology 1999.

[4] PC. Pendharkar, JA. Rodger, GJ. Yaverbaum, N. Herman, Benner M. Associations statistical, mathematical and neural approaches for mining breast cancer patterns. Expert Systems with Applications 1999.

[5] A. Toluee Ashlaqy, A. Poor Ebrahimi, M. Ibrahimi. Prediction of recurrence of breast cancer via three data mining techniques. Journal of Chest Diseases of Iran, Issue IV, 2013.

[6] H. Soleimanjahi, S. Fallahi, H. Riahi, Z. Ashkat. New Intelligent adaptive neural fuzzy inference system (ANFIS) to predict the strength of carcinogenic Human Papilloma Virus.2011.

[7] S.M. Hosseini, R. Hassan Nezhad, Khade ol Ghorani Sh.Identification and modeling of patterns of metastatic breast cancer in women referred to the center of Sayyid Alshhda.Journal of Researchs of the health system. Seventh year. No. VI. Health special name in 2012.

[8] R. Kavita, A Soft Computing Genetic-Neuro fuzzy Approach for Data Mining and Its Application to Medical Diagnosis. International Journal of Engineering and Advanced Technology 2013.

[9] H. Hazlina, Adaptive Neuro-Fuzzy Inference System (ANFIS) in Modeling Breast Cancer Survival. IEEE World Congress on Computational Intelligence2010.

[10] Kh. Shweta, A. Shika, S. Sunita, Naive Bayes Classifiers: A Probabilistic Detection Model for Breast Cancer. International Journal of Computer Applications Volume 92 – No.10, April 2014.

[11] Howida Ali Abd Elgader , Mohammed Hassen Hamza , Breast Cancer Diagnosis Using Artificial Intelligence Neural Networks. J.Sc. Tech, 2011.