Palestine Journal of Mathematics

Vol. 10(Special Issue I)(2021) , 135–139

CLASSIFICATION OF BREAST CANCER MAMMOGRAPHIC DATA

Benaki Lairenjam and Yengkhom Satyendra Singh

Communicated by Serkan Araci

MSC 2010 Classifications: 00A06, 06-XX-06-04.

Keywords and phrases: artificial neural network, CMAR (classification based on multiple association Rule), association rule, Bayes theorem.

Abstract Breast cancer is a serious disease that causes death in women if not diagnose early. Early diagnosis of the disease is considered as the best way to save life. Automatic detection of the disease from mammogram image using CAD (computer aided diagnosis) assist radiologist in predicting the disease accurately. In this paper a new model NNC (Neural Network Classifier) based on artificial neural network is developed for automatically detecting Breast cancer from mammography data. The model NNC is a hybrid of CMAR (Classification based on multiple association rule), Bayes theorem with NN (Neural network) for classifying breast cancer mammographic data as benign or malignant. The model consists of three layers: an input layer, a hidden layer and an output layer. CMAR is used in the initial step for creating the structure of the network and Bayes' Theorem is used for calculating initial weights for hidden layer. It is tested on Wisconsin breast cancer data from UCI repository. Experimental results show that NNC model gives better convergence rate as well as classification accuracy in breast data to both benign and malignant.

1 Introduction

Breast cancer is a serious disease that causes death in women if not diagnose earyl [8]. Early diagnosis of the disease is considered as the best way to save life. Automatic detection of the disease from mammogram image using computer aided diagnosis assist radiologist in predicting the disease accurately. Lots of of research work has been conducted to automatically diagnose breast cancer using (CAD). This is created using various data mining techniques such as, artificial neural network (ANN) [1], [10], [14], [17], [18], [19], [20] associative classifier [6], [11], decision trees and statistical classifier [5], genetic programming [10] and Bayesian Networks [3], [7]. Among all the techniques it is found that ANN gives high accuracy in medical dataset and is considered as a popular technique in medical diagnosis [16].

ANN architecture is obtained from the functioning of the human neurons. Its network architecture consists of an interconnected set of input, hidden and output nodes. The network architecture of a simple ANN has three layers: input, hidden and output. Each node in each layer is fully connected with weight that measures the effect of the node in the previous layer to the next layer. These weights are determined by learning the network using training datasets.

Associative classification is an integration of association rule mining and classification. In this process association rules are generated and analyzed for use in classification. Let D be a training dataset with n attributes A_1, A_2, ...A_n and m=|D| instances. The dataset also has a class attribute $C = \{C_1, C_2, ...C_r\}$. An item is characterized by an attribute A_i and its value a_ij denoted as (A_i, a_{ij}), where $j \leq m$. An itemset contains items in the training set $\langle (A_i, a_{i1}), ...(A_i, a_{ik}) \rangle$ where $k \leq m$. In associative classification a rule $r = \langle (A_i, a_{i1}), ...(A_i, a_{ik}), C \rangle$ form where $\langle (A_i, a_{i1}), ...(A_i, a_{ik}) \rangle$ is the rule antecedent and consequent is the corresponding class label.

CMAR is a classification method based on associative classification which comprise of rule creation and classification process. For the rule creation step, CMAR uses FP -growth algorithm satisfying the minimum confidence and support threshold.in the rule generation step it obtain

the total rules and store in the CR-tree. CR-tree is a prefix tree that store and prune rules using confidence, correlation, and database coverage, and can retrieve rules proficiently [2].

In this paper we will develop a model NNC for classifying breast cancer mammographic data. NNC is a hybrid approach of CMAR, Bayes' Theorem and neural network. The architecture of NNC will be created from the rules created by CMAR algorithm. Initial weight will be calculated using Bayes' Theorem for nodes connecting hidden and output layer.For learning the NNC model Backpropagation algorithm will be used. The classifierperformance will be evaluated using classification accuracy, and roc curve.

Our model NNC will be created and tested on Wisconsin Breast Cancer Database [9]. The data set have 9 attributes together with the class attributes i.e., benign and malignant. The dataset have 699 record of which 458 are benign and 241 are malignant.

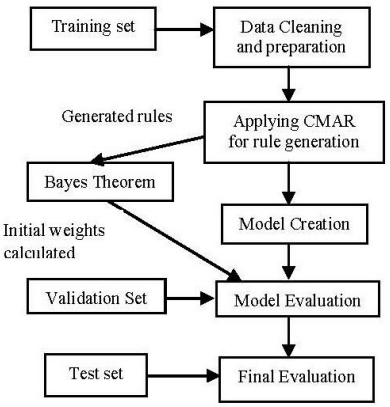

Figure 1. Prototype model for NNC

2 System Approach

Feedforward multilayerd neural network are most frequently used for modeling medical data [4]. They are computational structure consisting of an interconnected node arranged on a multilayerd hierarchical architecture. Multilayerd neural network with backpropagation algorithm has been successfully applied in predicting and classifying breast cancer [10]. The major disadvantages of backpropagation algorithm are slow convergence rate. The performance and convergence of the neural network depends on learning rate, initial weights, connection between the nodes and the number of layers.

In this paper a new model is created by combining CMAR, Bayes' Theorem and NN as a hybrid approach for classifying breast cancer mammographic data. CMAR algorithm is applied on the training datasets. The CMAR algorithm is applied on the training dataset stating an initial support of 10% and confidence 50%. From the rules generated on the dataset the network architecture is created with the number of attribute values present in the rule as input nodes and the number of rules as the nodes in the hidden layer. The number of classes will represent the number of nodes in the output layer.Weight for input to hidden layer are given random initial weights and initial weights for hidden and output layer is given the value calculated using Bayes' Theorem.The model is learn using backpropagation algorithm.The network were trained until the error is low with a learning rate of 0.1. The NNC model can improve in the number of average iteration by putting initial weight in the hidden layer to the output layer. The overall process flow chart in NNC prototype model are shown in Fig. 1.

3 Model Creation and Learning

We create a classifier NNC using CMAR algorithm, Bayes theorem and backpropagation neural network. The model creation process can be divided into three steps: (see Fig. 2.)

- Generation of CMAR rule from the training dataset.
- Creation of the neural network model.

- Calculation of initial weight using Bayes theorem.

In NNC model each node in the input layer representing the characteristic are connected to the hidden nodes representing the rule. Nodes in the hidden layer and output are fully connected.

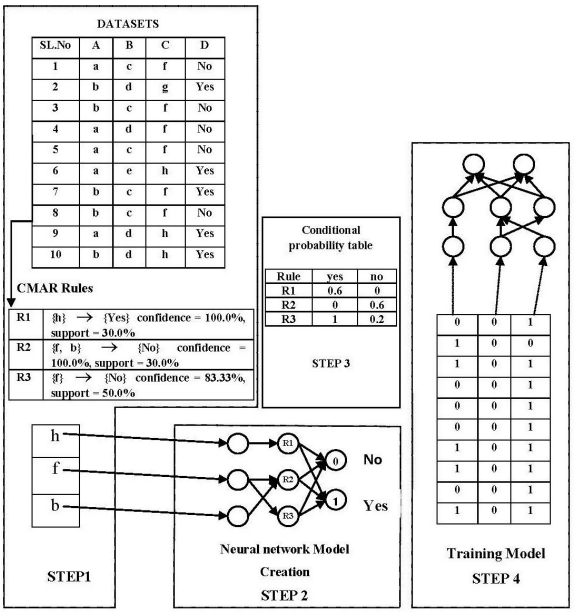

Figure 2. NNC model creation.

In the third step for each rule generated using Bayes' Theorem conditional probability of each rule condition on the class label is calculated. The probability of rule (R), $P(R)$ is expressed as,

$P(R) = \Sigma_{i=1}^{i=m} P(C_i)P(R|C_i)$

with $P(C_i)P(R|C_i)$ denoting the probability of classifying patterns from class i. The conditional probability of rule R condition on class C_i is,

$P(C_i|R) = \frac{P(R|C_i)P(C_i)}{P(R)}$

After the creation of NNC model, it is learn usingbackpropagation neural network algorithm. Initial random weights are given for the input and input weight for hidden and output layer is the value calculated using Bayes theorem. Sigmoid activation is used for both the hidden and output layer. 10-fold cross validation is used for learning NNC. Each model is learned repeatedly on the test set with 10-fold cross-validation until the error is low. It is tested by using the test set.

On the training data set CMAR algorithm is applied with a support of 10% and confidence of 50%. The NNC is created from the rules generated.The model of NNC has an input-a hidden-an output nodes. Each rule comprises of attribute values and they are called as Characteristic. The number of input nodes are determined by the number of characteristics present in the rule and each characteristic represent each input nodes. The number of rules in CMAR algorithm gives the number of hidden nodes that must be put in the hidden layer and the number of classes present in the dataset determines the number of output nodes in NNC.

4 Evaluation of the classifier on the mammographic data

We experiment on Wisconsin Breast Cancer Database taken from UCI repository [9]. The dataset has 699 records, of which 458 records are benign and 241 records are malignant. The datasets have 9 attributes : (1) Lump Thickness; (2) Uniformity of Cell Size; (3) Uniformity of Cell Shape; (4) Marginal Adhesion; (5) Single Epithelial Cell Size; (6) Bare Nuclei; (7) Bland Chromatin; (8) Normal Nucleoli; (9) Mitoses. The class attributes have two values benign (non-cancerous) and malignant (cancerous).

The dataset is pre-process to check outliers, noisy, extreme or missing data using normal statistical cleaning process in SPSS. These values are replaced with the attribute's mean or average depending on its suitability.

Out of the 699 records two third of the datasets are taken for training and one third of the datasets are kept for testing i.e., 466 datasets for training and 233 datasets for testing. The network model is train with the training dataset using ten-fold cross validation. Ten-fold cross validation is used to demonstrate the error rate of networks.

Additionally, we created and trained the model NNwCMAR using the training dataset, following the method given in Benaki and Wasan [17]. We evaluate the average iteration of convergence of both NNC AND NNwCMAR. Fig. 3 shows the training error of both the networks. We see that model NNC with initial weights converge in 910 and NNwCMAR in 1990 iteration. The results have been shown in Fig. 4.

Table 1. Error Convergence for NNC and NNwCMAR

Ten split Test	Error in NNC 910 ITERATION	Error in NNwCMAR 1990 ITERATION
1	0.010245	0.014233
2	0.022052	0.026601
3	0.026872	0.029122
4	0.019845	0.021725
5	0.020521	0.022827
6	0.035945	0.03504
7	0.020011	0.022843
8	0.022325	0.022193
9	0.038856	0.044095
10	0.018195	0.020964
Avg. error	0.023487	0.025964

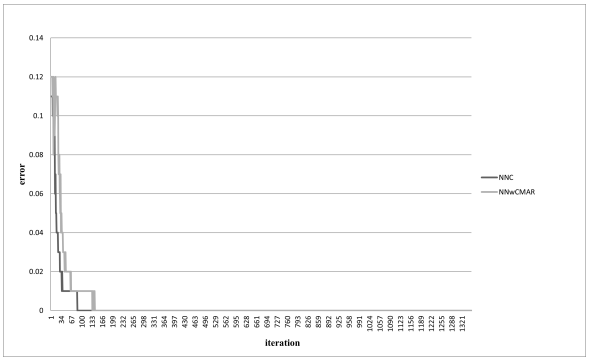

Figure 3. Error convergence for NNC and NNwCMAR Model.

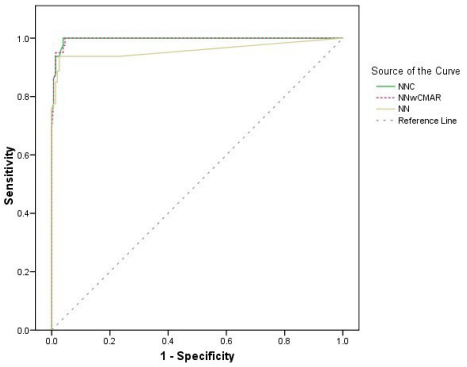

Figure 4. ROC curves of NNC, NNwCMAR and NN.

References

[1] D.S. Sudhir, A.G. Ashok, P.P. Amol, Neural network aided breast cancer detection and diagnosis, *Proc. 7th WSEAS International Conference on Neural Networks 2006*, 158—163 (2006).

[2] W. Li, J. Han and J. Pei, Accurate and efficient Classification Based on Multiple Class-Association Rules, *IEEE ICDM, 2001* 283-299 (2001).

[3] N. Ferreira, M. Velikova and P. Luca, Bayesian Modelling of Multi-View Mammography,*Proc. ICML/UAI/COLT 2008 Workshop on Machine Learning for Health-Care Applications, Helsinki, Finland, 2008*.

[4] O. Maimon and Rokach, Data Mining and Knowledge Discovery Handbook,*Springe*r, 193-230 (2005).

[5] I.H. Witten and E. Frank, Data Mining Practical Machine Learning Tools and Techniques, *2nd ed., Elsevier: Morgan Kaufmann Publishers*, 83-112 (2005).

[6] M.L. Antonie, O.R. Zaïane, and A. Coman, Associative Classifiers for Medical Images,*LNCS, Mining multimedia and Complex Data*,**2797**, 68-83 (2003).

[7] E.S. Burnside, D.L. Rubin, J.P. Fine, R.D. Shachter, G.A. Sisney and W.K. Leung, Bayesian Network to Predict Breast Cancer Risk of Mammographic Microcalcifications and Reduce Number of Benign Biopsy, *Radiology*, **240**, 666-673 (2006).

[8] National cancer institute http://www.cancer.gov/cancertopics/pdq/treatment/breast-cancer-and-pregnancy/healthprofessional.

[9] UCI Repository, http://www.archive.ics.uci.edu/ml/datasets/ Mammographic + Mass

[10] A. Adam, K. Omar, Computerized Breast Cancer Diagnosis with Genetic Algorithms and Neural Network, *Proc. of the 3rd International Conference on Artificial Intelligence and Engineering Technology (ICAIET)*, 533-538 (2006).

[11] O.R. Zaiane, M.L. Antonie, A. Coman, Mammography Classification by an Association rule based Classifier, *Proc. International Workshop on Multimedia Data Mining (MDM/KDD '2002) in conjunction with ACM SIGKDD*, 62–69 (2002).

[12] http:://as.webmd.com/event.ng/

[13] M.H. Dunham and S. Sridhar, Data Mining Introductory and Advanced topics, *1st Impression*, pp. 48-52.

[14] A. Maria-Luiza, R. Osmar Zaiane, C. Alexandru, Application of data mining techniques for medical image classification, *Proc. Of Second Intl. Workshop on Multimedia Data Mining (MDM/KDD '2001) in conjunction with Seventh ACM SIGKDD, USA*, 94-101 (2001).

[15] The LUKAS- KDD Implementation of the CMAR Algorithm, http: // www.csc.liv.ac.uk / ~frans / KDD / Software / CMAR / cmar.html

[16] Neural Network Applications in Medical Research, http://www.nd.com/apps/medical.html

[17] B. Lairenjam, and S.K. Wasan, Neural Network with Classification Based on Multiple Association Rule for Classifying Mammographic Data, *Proc. Intelligent Data Engineering and Automated Learning-IDEAL 2009, LNCS*, 465-476 (2009).

[18] Y. A. Hamad, K. Simonov and M. B. Naeem, Breast Cancer Detection and classification Using Artificial Neural Networks, *In Proceedings of 1st International Conference on Information and Sciences*, (2018).

[19] N. Tariq. Breast Cancer Detection using Artificial Neural Networks. *Journal of Molecular Biomarkers & Diagnosis*, **9**, 1-6 (2018).

[20] J.M.O. Rodriguez, C.G. Mendez, M.R.M. Blanco, S.C. Tapia, M.M. Lucio, R.J. Martinez, L.O.S. Sánchez, M.L. Fierro, I.G. Veloz, J.C.M. Galvan and J.A.B. Garcia. Breast Cancer Detection by Means of Artificial Neural Networks,*Intechopen* (2018).

Author information

Benaki Lairenjam and Yengkhom Satyendra Singh, Department of Mathematics, REVA University, Bangalore-560064, India.
E-mail: benaki_lai@yahoo.co.in and yengkhom123@gmail.com

Received: Dec 20, 2020.
Accepted: Feb 26, 2021.