Int. J. Open Problems Compt Math., Vol. 3, No. 5, December 2010
ISSN 1998-6262; Copyright © ICSRS Publication, 2010
www.i-csrs.org

A note on analysis of Mammography Data

Benaki Lairenjam[1] and Siri Krishan Wasan[2]

[1]Department of Mathematics, Jamia Millia Islamia, New Delhi, India
e-mail: benaki_lai@yahoo.co.in

[2]Department of Mathematics, Jamia Millia Islamia, New Delhi, India
e-mail: skwasan@yahoo.com

Abstract

In this short note we analyze the performance of Backpropagation Neural Network (BPNN), Radial Basis Function Network (RBFN), Classification Based on Multiple Association Rule (CMAR) and Classification Based on Association (CBA) on mammographic mass data from UCI repository. The performance of the classifier is evaluated using sensitivity, specificity and classification accuracy.

Keywords: Breast Cancer, CMAR, Backpropagation Neural Network, Radial Basis Function.

1 Introduction

Breast Cancer is the most common type of cancer among women in the world. It is a cancer that is formed in the tissue of the breast. It is the most frequently diagnosed life-threatening cancer in women in both developing and developed countries and is the principal cause of death from cancer. Detection of breast cancer in the early stage remains the primary defensive measure to prevent the development of breast cancer. Breast cancer if detected early increase the chance of survival. Accurate classification of malignant and benign masses is very important, because it affects the patient management and choice of treatment.

During the past decade, lot of work has been done to help radiologist in detecting suspected structure in breast cancer image. Different data mining techniques are used for creating computer aided diagnosis (CAD), to diagnosis breast cancer accurately and in less time. Some of the data mining techniques widely used for developing computer aided diagnosis are bayesian network [5][14], naives bayes

classifier [3], genetic algorithm [1], artificial neural network [1][2][3][4], associative classification [9][10][16]. Many hybrid approaches are also used to accurately diagnose mammographic findings. In [1] genetic algorithm is hybridized with neural network to classify Wisconsin breast cancer data. In [2] classification based on multiple association rule and neural network are hybridized to create CMARwNN model for classifying mammographic mass data.

Association rule mining discovers various relations that data share between them. An association rule is a relationship of the form $X \to Y$, where X and Y are disjoint itemsets i.e., $X \bigcap Y = \phi$. The support and confidence of the rule are defined as, Support $(X \to Y) = P(X \bigcup Y)$ and Confidence $(X \to Y) = P\left(Y \middle/ X\right)$. The association rule mining task consists of extracting all rules with support and confidence greater than or equal to user-specified thresholds.

Classification is the process of finding a model by analyzing a set of training data, such that the model can be used to predict the class of previously unseen records as accurately as possible. Associative classification is an integration of association rule mining and classification. In this process association rules are generated and analyzed for use in classification.

Classification based on association (CBA) is the earliest and simplest algorithm for associative classification [6]. 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 described by an attribute A_i and a value a_{ij} denoted as (A_i, a_{ij}), where $j \le m$. An itemset is a set consisting of items in the training set $\langle (A_i, a_{i1}), ...(A_i, a_{ik}) \rangle$, where $k \le m$. A rule r in associative classification is of the form $r = \langle (A_i, a_{i1}), ...(A_i, a_{ik}), C \rangle$ where the rule antecedent is a conjunction of items $\langle (A_i, a_{i1}), ...(A_i, a_{ik}) \rangle$ and consequent is the associated class label C. The occurrence of a rule r is the number of instances in D that match the itemset of r. For a given rule r, the support count is the number of instances in D that match the itemsets of r, and belong to the class label C of r. Thus the confidence of a rule r is the percentage of the instances in D satisfying the rule antecedent that also have the class label.

The strength of the class association rule CAR is measured using support and confidence threshold. CBA generates the complete set of CAR's for classifying new instances. In the classification process, if more than one rule fit a certain cases then CBA will classify the class from the rule with the highest confidence. If the confidences are same then the rule having the highest support will be used to classify the case. CBA saves a default class to deal with the case when no CARs can classify.

CMAR is an associative classification method that performs classification based on multiple association rules. It consists of two steps rule generation and classification. In the rule generation step, CMAR finds the complete set of rules using FP[1]-growth algorithm [7] in the form $R: X \rightarrow C$, where X is a pattern in the training dataset and C is the class label satisfying the minimum confidence and support threshold. Once a rule is generated, it is stored in a CR-tree. CR-tree is a prefix tree data structure which store and retrieve rules efficiently and prune rules based on confidence, correlation and database coverage. Whenever a rule is inserted into the CR-tree, it prunes all rules and only selects subsets of high quality rule for classification [9].

Artificial Neural Network or Neural Network (NN) are computing models for information processing that mimic the little understanding of biological neurons of human brain. It is a non-linear model that learns through training. Important classes of neural network models include feedforward multilayer networks, Hopfield networks and Kohonen's self organizing maps. Multilayer feedforward neural network also called Multilayer Perceptron (MLP) are the most widely studied and used neural network model [15]. MLP is a directed graph with a set of input/output nodes, which are interconnected to each other, and each connection is associated with a weight representing the strength of the connection between the nodes. These weights are determined by means of a learning process on the training data where the training algorithm iteratively adjusts the connection weights. The most popular algorithm is backpropagation algorithm. Backpropagation algorithm uses sigmoid function as activation function and gradient technique to modify the weights. Gradient descent is a technique to modify the weight in the graph. Its basic idea is to find the set of weights that minimizes the MSE (Mean Square Error).

The structure of a three layer multilayer feed forward neural network is shown in Fig. 1, where $X_1,...X_n$ are input to the input node, O_j is an output of the j^{th} node in the hidden layer and O_i is an output of the i^{th} node in the output layer. The nodes in the hidden layer and output layer processes information from several inputs and then converts it into outputs. Each nodes process information by combining the inputs $X_1,...X_n$, with the weights w_{ij} (weight of the connection of node j and node i in the previous layer) to form a weighted sum of inputs and weights of connecting links and using transfer function converts into output [6]. Training the networks is an optimization problem of finding the set of network parameters (weights) that provide the best classification performance. Backpropagation trains the multilayer feed-forward neural network by modifying the weights and bias so as to minimize the MSE (Mean Square Error) and the output node. Backpropagation starts after calculating MSE at the output nodes

[1] Frequent pattern is the frequent item sets that satisfy specified threshold i.e., minimum support and minimum confidence.

using learning rules to calculate derivatives of the squared error w.r.t weights and bias at the output and hidden layers in the network.

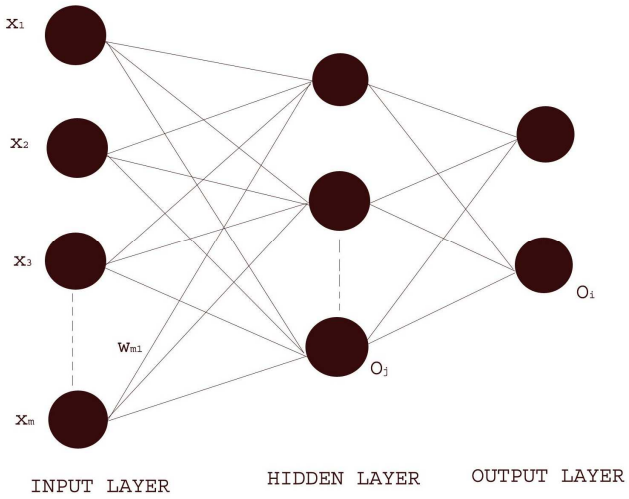

Fig. 1 Structure of three layer neural network

Radial Basis Function (RBF) is a popular type of feedforward neural network [8]. It has three layers: an input layer, a hidden layer and an output layer. The hidden layer in RBF network is nonlinear and output layer is linear. Gaussian activation function is commonly used activation function in the hidden nodes and sigmoid function is used in the output nodes similar to multilayer perceptron. It is a fully connected network. The network input represents features and output corresponds to a class. Learning the network is to optimize the network parameters in multidimensional space to fit the network outputs to the given input. The fit is evaluated by mean squared error. The parameters that RBF learns are (i) centers and width of the RBF and (ii) weight connecting the hidden nodes and the output nodes.

In this paper we analyze mammographic mass data from UCI repository [19]. We have used the same dataset in our earlier papers [2] [3] for classifying the data into benign and malignant, using neural network, associative classifier and Naïve Bayes classifier. CBA, CMAR and three layered BPNN and RBF networks are used to differentiate malignant from benign findings from mammographic mass data. The database was randomly divided into training and validation samples using ten-cross validation, to construct CBA, CMAR, BPNN and RBF networks and validate their performance. Sensitivity, specificity and classification accuracy are used to evaluate the performance of the classifier.

2 Analysis of Mammography Data

2.1 Data Cleaning

A Mammography data of 961 instances is taken from UCI repository [19], and name of the dataset is mammographic mass data. The dataset consists of 516 instances classified as benign and 445 instances as malignant. This database contains BI-RADS assessment, patient's age and three BI-RADS attributes together with the ground truth i.e. benign and malignant those have been identified on full field digital mammograms collected at the Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006.

BI-RADS (Breast Imaging Reporting and Data system), developed by the American College of Radiology provides a standardized classification for mammographic studies. The system demonstrates good correlation with the likelihood of breast malignancy. The BI-RADS system can inform domain expert about key findings, identify appropriate follow-up and management [13]. The BI-RADS attributes are mass shape, mass margin and mass density [19]. Masses can be circumscribed, microlobulated, round, oval, lobular, irregular etc and radiologist will need to take a good look to find the suspicious lesions. Therefore in order to classify a mass, the BI-RADS attributes recorded were used as an input feature. The attributes as per the relevant datasets are the following:

- BI-RADS assessment: 1 to 5 (ordinal)
- Age: patient's age in years (integer)
- Shape: mass shape: round=1 oval=2 lobular=3 irregular=4 (nominal)
- Margin: mass margin: circumscribed = 1, microlobulated = 2, obscured = 3, ill-defined = 4, speculated = 5 (nominal)
- Density: mass density high = 1, iso = 2 low = 3, fat-containing = 4 (ordinal)
- Severity: benign = 0 or malignant = 1 (binominal)

There is some missing information in the dataset. The number of missing value in BI-RADS assessment: 2, Age: 5, Shape: 31, Margin: 48, Density: 76 and Severity: 0. Therefore two different cleaning processes has been carried out on the dataset. In the first process rows containing missing information are removed from the dataset. We called this dataset A. The dataset A undergo normal statistical cleaning process where all the attributes are distributed normally to check outliers and extreme. Histogram and normal curve graph were projected to ensure that all data are normally distributed.

The second process was to fill in the missing values and retain all the instances. This dataset is called B. Dataset B undergo normal statistical cleaning process where all the attributes are distributed normally to check outliers, extreme, noisy

or missing values. These values are replaced with the attribute's mean or average depending on its suitability. Histogram and normal curve graph were projected to ensure that all data are normally distributed. Analyses were done on both the datasets using CBA, CMAR, BPNN and RBF networks, and the results are compared.

2.2 Classification and Testing

In this section different classification model are considered and tested. For each classifier classification is done on the dataset A and B. The five features of mammographic mass data were used in CBA, CMAR, BPNN and RBF network classifier to predict breast mass into benign and malignant.

Several model architecture of CBA, CMAR, BPNN and RBF network are considered in the classification phase. Architecture of CBA and CMAR are model setting a starting support of 5% and confidence of 50%. The neural network architecture of RBF use Gaussian activation function in the internal node and sigmoid activation function in the output node and BPNN use sigmoid activation function in the internal and output node. Both the network architecture consists of three layers i.e., input layer, hidden layer and output layer.

Classifications were conducted on dataset A and B. Both the datasets are randomly divided into training and validation sample using ten-cross validation to construct CBA, CMAR, BPNN and RBF network model and validate its performance.

2.3 Performance of Classifier

The analysis of different classification algorithms are done on Dataset A and B. The classification performance of each classifier is evaluated using three statistical measures: sensitivity, specificity and classification accuracy. Sensitivity refers to the portion of people with disease who have a positive test result that is tp [8]. Specificity refers to the portion of people without disease who have a negative test result which is (1-fp) [8].The formula for sensitivity and specificity are:

$$\text{Sensitivity (tp)} = TP/(TP + FN), \text{Specificity (1-fp)} = TN/(TN + FP)$$

Here TP stands for true positives, TN for true negatives, FP for false positives, FN for false negatives. Classification accuracy is defined as the ratio of correctly classified instances and is equal to the sum of TP and TN divided by the total number of instances N.

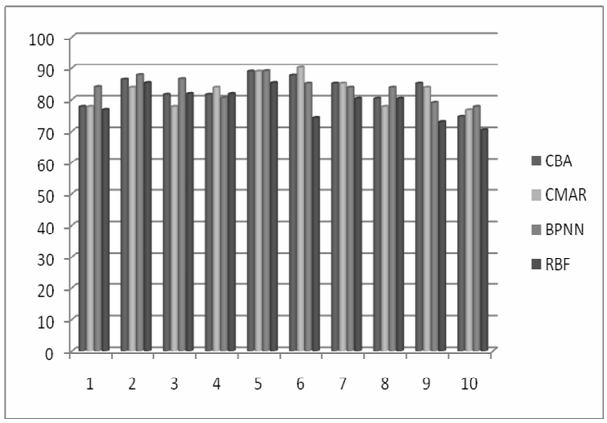

Fig. 2: Classification performance over the ten cross split on dataset A

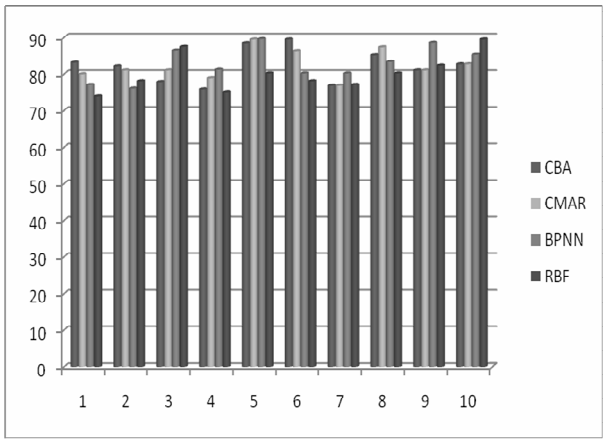

Fig. 3: Classification performance over the ten cross split on dataset B

$$\text{Accuracy} = (TP + TN)/N$$

Figure 2 and 3 shows the classification performance of the ten split on the validation set of both the datasets.

The average classification accuracy of the four classifiers is shown in Table 1 and Table 2. From the result we see that BPNN produce high classification accuracy on dataset A. Classification accuracy of CBA and CMAR are found to be similar in both the dataset. We also see that the classification accuracy of CBA and CMAR are consistent on both the dataset. The classification accuracy of RBF slightly improved in dataset B but comparatively lower then CBA, CMAR and BPNN.

Table 1: Average classification performance on dataset A

Type	*Accuracy*
CBA	82.87%
CMAR	82.61%
BPNN	84%
RBF	78.89%

Table 2: Average classification performance on dataset B

Type	*Accuracy*
CBA	82.27%
CMAR	82.48%
BPNN	82.9%
RBF	80.19%

We also present sensitivity and specificity analysis in Table 3 and Table 4. From the sensitivity and specificity analysis we see that sensitivity of CMAR is consistent in both the dataset. The respective results of sensitivity, specificity and accuracy of 84%, 85.45% and 84.12% obtained from BPNN on dataset A were comparable to the result obtained for CBA and RBF.

Table 3: Dataset A

Type	*Sensitivity*	*Specificity*
CBA	84.23%	81.28%
CMAR	87.02%	80.14%
BPNN	85.45%	84.12%
RBF	81.93%	76.06%

Table 4: Dataset B

Type	*Sensitivity*	*Specificity*
CBA	82.36%	79.94%
CMAR	87.27%	78.8%
BPNN	79.86%	85.44%
RBF	79.4%	80.84%

The experiment is made on a computer with a single 1.73 GHZ Core (TM)2-CPU and 1280 MB memory and BPNN and RBF networks are run on weka. The source program of CMAR is downloaded from [17] and CBA is downloaded from [18].

3 Conclusion

We perform an analysis of mammographic mass data from UCI repository using associative classifier and artificial neural network. Cleaning process has been carried on the dataset to prepare for mining process. We have found that BPNN perform better in the dataset where missing record has been removed. But CMAR perform consistently on both the dataset. It will be interesting to examine a hybrid approach combining the features of BPNN and CMAR.

4 Open Problem

Many works has been done to develop a classifier model that classifies breast cancer mammographic data into benign and malignant. Accurate classification of malignant and benign masses is very important, because it would lead to tremendous benefits both in terms of lives saved each year, and in terms of reduced workload of radiologist.

References

[1] A. Adam, K. Omar, Computerized Breast Cancer Diagnosis with Genetic Algorithms and Neural Network.

[2] 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*, LNCS, Vol. 5788/2009, pp. 465-476

[3] B. Lairenjam and S.K. Wasan, Naïve Bayes Associative Classification of Mammographic Data, *Proc. International conference on education and network technology*, IEEE, 2010, pp. 276-281.

[4] Carey E. Floyd and Daniel C.Sullivan, Predicting of Breast Cancer Malignancy Using an Artificial Neural Network, *CANCER*, Vol. 74, No.11 (1994).

[5] 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*, Vol. 240, No. 3 (2006), pp.666-673.

[6] J. Han and M. Kamber, *Data Mining, Concepts and Techniques*, Morgan Kaufmann, 2001.

[7] J. Han, J. Pei, and Y. Yin, Mining frequent patterns without candidate generation, *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, Vol. 29, No. 2 (2000), pp.163-170.

[8] Ian H. Witten and F. Eibe, *Data Mining Practical Machine learning Tools and Techniques*, Elsvier, 2005.

[9] W. Li, J. Han and J. Pei, CMAR, Accurate and Efficient Classification Based on Multiple Class-Association Rules, *Proceeding of First IEEE International Conference on Data Mining (ICDM'01)*, icdm, 2001, pp.369.

[10] Maria-Luiza, A., Osmar, R., Zaiane, Alexandru, C., 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*, San Francisco, USA (2001) , pp. 94-101.

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

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

[13] M. Margaret, Eberl, C.H. Fox, MD, S.B. Edge, C.A. Carter, and M.C. Mahoney, BI-RADS Classification for Management of Abnormal Mammograms, *The Journal of the American Board of Family Medicine19*, 2006, pp.161-164.

[14] 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.

[15] O. Maimon and Lior Rokach, *Data Mining and Knowledge Discovery Handbook*, Springer, 2005.

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

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

[18] The LUCS-KDD Implementations of the CBA Algorithm, http://www.csc.liv.ac.uk/~frans/KDD/Software/CBA/cba.html.

[19] UCI Repository, http://archive.ics.uci.edu/ml/d.