Prediction of Breast Cancer Biopsy Outcomes – An Approach using Machine Leaning Perspectives

Sandeep Chaurasia
Sir Padampat Singhania University
India

Prasun Chakrabarti
Sir Padampat Singhania University
India

Neha Chourasia
Mewar University
India

ABSTRACT
Breast cancer is the most frequently diagnosed cancer in USA. Furthermore breast cancer is the second major cause of death for women in USA. Several well established tools are currently used for screening for breast cancer including clinical breast exam, mammograms and ultrasound. Mammography is one of the most effective in terms of accuracy and cost. However the low positive predicted value (PPV) of breast cancer biopsies resulting from mammograms leads to 70% unnecessary biopsies with benign outcomes. In order to reduce the large number of surgical biopsies of breast, several CAD based system has been proposed in the last decades. Using these systems the radiologist gets an aid on their decision to perform breast biopsies. The dataset used is based on BIRADS findings. Prior work achieves good result with decision tree and neural network. The paper use AutoMLP, BP (back propagation) neural network and support vector machine (SVM) approach to predict the outcomes of mammogram with better result. Using SVM the false biopsies should significantly reduced to only 13%.

Keywords
Breast cancer; classifier; BIRADS; decision tree; naïve bayes; neural network; support vector machine.

1. MACHINE LEARNING IN MEDICAL CONTEXT – AN INTRODUCTION
In recent years the mortality rate of breast cancer has significantly increased. Now it is the foremost cause for the casualty in women, with the ratio of one out of every ten women is affected by the breast cancer during their life span. Since 2000 in USA, breast cancer is the second largest cause of cancer deaths among women following the lung cancer. In USA 40,600 deaths from breast cancer in 2009, 400 were men [1] [2]. Mammography is the most effective screening technique available today for breast cancer. Though the less positive predictive value of breast cancer biopsies resulting from mammogram screening leads to an approx 70% of the avoidable biopsies with outcomes as benign [3]. An effective way to reduce the high mortality rate of breast cancer is to detect it at an early stage. Prevention is still a mystery because of censored data and the only way to reduce the mortality rate of patients by early detection. Still the detection of breast cancer at early stages is one of the major challenges in medical science.

Machine learning classifiers are reducing the potential challenges that might be made because of unproven experts provides more comprehensive medical facts for examination in a lesser time [4]. A study showed that if the cancer cells are identified before spreading to any other organs then the survival rate for patients could increase up to 97% [5]. One of the primary goals of machine learning is to device an efficient algorithm for training computers to automatically acquire effective and accurate model from experience. It is providing a technique, method tools that can assist in solving prognosis and diagnosis problems in a variety of medical domains. There are many applications for Machine Learning of which the most significant is computational intelligence and pattern classification.

In this paper, the BIRAD breast cancer dataset has been analyzed over different machine learning principles of classification techniques. This paper is organized as follows, section 2 provides the brief of the related work done, section 3 highlighted a brief introduction of various classification techniques and algorithm, section 4 provide a detailed description of data sets, section 5 shows the comparison statistics of the mention techniques with the acquired results. Finally section 6 concludes the result.

2. RELATED WORK
There has been research with WBCD the breast cancer database on computer aided diagnosis and prognosis of breast cancer. Quinlan J.R has presented an algorithm using C4.5 decision tree method using 10-fold cross validation and reported an accuracy of 94.74% [6]. Hamilton, Shan and Cercone presented a method named Rule induction algorithm based on approximate classifications and reported the accuracy of 94.99% [7]. Nauck D, and Kruse R presented an neuron-fuzzy technique for classification of medical data and reported the accuracy of 95.06% [8]. Abonyi and Szeifet had applied the supervised fuzzy clustering technique and reported an accuracy of 95.57% [9]. Albercht, Lappas, Vinterbo, Wong and Ohno-Machado presented a learning algorithm that combined logarithmic simulated annealing with the perceptron algorithm was used and reported an accuracy of 57% [10]. Guijarro B., B.,Fontenla R. O., Perez S. B, and Fraguela P., presented a learning algorithm by applying linear- least squares method and reported an accuracy of 96% [11]. Karabatak and Cevdet I., presented an automatic diagnosis system for detecting breast cancer based on association Rules (AR) and neural networks (NNs), and reported an accuracy of 97.4% [12].

3. CLASSIFICATION TECHNIQUES
The classifier's accuracy is most often based on prediction evolutions. There are various methods to evaluate the accuracy by dividing the data set, two-third for training and one-third for testing. Another method as cross validation, the training set is split into mutually exclusive and equal sized subset and for each subset the classifier is trained on the union of all others subset. And the last is leave-one-out validation is a modified case of cross validation [13]. A large number of methods have been developed based on logic or symbolic

based techniques, perceptron based techniques, bayesian network and support vector machine.

3.1 Logic and Symbolic Based

Decision tree are trees that classifies the instances by sorting them based on the feature value of an instance to be classified, each branch represents an attribute value that the node has initialized. Instances are classified starting at the root node having the best information gain and sorted based on their feature value [14]. C4.5 is a well known algorithm and is an expansion of Quinlan's ID3 algorithm with having a very good combination of error rate and speed [15].

3.2 Statistical Learning

Naïve bayes are simple Bayesian network composed of dag (direct acyclic graph) with one parent and several children. The naïve assumption of independency among child nodes is usually almost wrong and for this reason the quick learner naïve bayes classifier are usually less accurate than other sophisticated learning algorithm. Conditional probabilities: Pi(xi|C = c), the probability that the feature value in the i-th position is equal to xi given class c, were estimated using kernel density estimation (KDE) from a set of labeled training data (X, C). KDE is a non-parametric way of estimating the probability density function population [16]. The probability Pi(xi|C = c) was estimated using Equations.

$$P_i(x_i \mid C = c) = \frac{1}{N_c h} \sum_{j=1}^{N_c} K(x_i, x_{j|i|c})$$

$$K(a,b) = \frac{1}{\sqrt{2\pi}} e^{\frac{(a-b)^2}{2h^2}}$$

where K is a Gaussian function kernel with mean zero and variance 1, Nc is the number of the input data X belonging to class c, xj|i|c is the feature value in the i-th position of the j-th input X = (x1 x2 … xi … xn) in class c, and h is a bandwidth, or a smoothing parameter. To optimally estimate the conditional probabilities, h was optimized on the training data set.

3.3 Perceptron Based Techniques

Artificial neural network (ANN) depends on input data, its activation function and weight of each input connection. There are several algorithms by which a network can be trained [17], but the most popular algorithm is back propagation (BP) algorithm. The back propagation algorithm will perform a number of weight modification before it concludes with a good weight configuration for n training instances and w weights each epoch in learning takes O(nw) time. The auto multilayer perceptron the technique combines the concept of stochastic optimization and genetic algorithm. The process creates small ensembles of multilayer perceptron networks with different numbers of hidden units and with different learning rates. The different parallel networks are trained for small number of training cycle then the error rate is evaluated on testing set. After few cycles the worst performers are substituted with copies of best networks, modified to have different numbers of hidden neurons and learning rates.

3.4 Support Vector Machine

The support vector machine is originally a binary classification method developed by Vapnik et.al at Bell laboratories [18]. Typically for binary classification, the training data point $\{x_i, y_i\}, i = 1 \ldots l$, $y_i \in \{-1, 1\}, x_i \in R^d$. Suppose for some hyperplane that separates or classify the positive label from the negative labels with a separating hyperplane. The points x which is lie on the hyperplane satisfy w·x + b = 0, where w is normal to the hyperplane |b|/||w||, is the perpendicular distance from the hyperplane to the origin, and ||w|| is the Euclidean norm of w [4]. For non linear separable case, selection of appropriate kernel function is important because kernel function is responsible to transformed feature space in which training set instance will be classified. Training the SVM is performed by solving nth dimensional QP problem, where N is the number of samples in training dataset which involves large matrix operations.

4. BREAST CANCER DATASET OVERVIEW

Data set can be used to predict the outcome as benign or malignant for a mammographic mass lesion from BI-RADS attribute. BIRADS is a quality assurance tool for mammographic screening which record the value from 1 to 5.

Table 1. Attributes of the BIRADS Database

Attribute	Attribute description	Value of attribute
1	BI-RADS assessment	1 to 5
2	Age: patient's age in years	Integer
3	Shape: mass shape	round=1 oval=2 lobular=3 irregular=4
4	Margin: mass margin	circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 speculated=5
5	Density : mass	high=1 iso=2 low=3 fat-containing=4
6	Severity:	benign=0 or malignant=1

It contains BI-RADS assessment, the patient's age and three BI-RADS attributes together with the ground truth (the severity field) for 516 benign and 445 malignant masses that have been identified on full field digital mammograms collected at the Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006. Number of Instances are 961 with number of attributes are 6. Attribute 3 to 5 are recorded during BIRAD assessment.

5. RESULTS

The original data is present in the form of analogue values with different range of data. The data are converted to their equivalent integer or real number form. Then the mean and the standard deviation are calculated to normalize the data. Then the label field is identified for dataset 2 it is 0 for benign and 1 for malignant. If any missing data is encountered then the missing value is replaced by the mean value of the column. To measure the performance of the breast cancer diagnosis of the classifiers used in this investigation, the process divides the evaluation into two parts first is to determine performance result accuracies by means of classification accuracy [19], analysis of specificity and sensitivity, and confusion matrix and the second is by the performance results in term of ROC, related to ROC curve analysis and area under the curve (AUC). The performance measure methods are explain in the following sections:

5.1 Performance Result Accuracy

Classification accuracy: In this study the classification accuracy for each data sets are calculated using the following equation:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- True positive (TP): An input instance is actual malignant that were correctly classified as malignant.
- True negative (TN): An input instance is malignant that were correctly classified as benign
- False positive (FP): An input instance is benign those were incorrectly classified as malignant.
- False negative (FN): An input instance is malignant that were incorrectly classified as benign.

5.2 Sensitivity and Specificity

For measuring performance by means of sensitivity and specificity analysis, the following expressions is used.

$$Sensitivity = \frac{TP}{TP + FN}(\%)$$

$$Specificit\,y = \frac{TN}{FP + TN}(\%)$$

5.3 Confusion matrix

A confusion matrix contains information about actual and predicted classifications in the matrix form as performed by a classifier. Table II shows the confusion matrix for a two class classifier.

Table 2. Representation of Confusion matrix

Actual	Predicted	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Table 3. Classification accuracies of classifiers

Type classifier	Classification accuracies (%)		
	Specificity	Sensitivity	Accuracy
Decision tree	74.84	86.44	78.79
Naïve bayes	82.53	80	81.27
Naïve bayes kernel	85.36	79.48	82.41
Neural net	82.23	79.68	81.56
Auto MLP	83.07	79.24	81.27
SVM	84.54	81.88	83.25

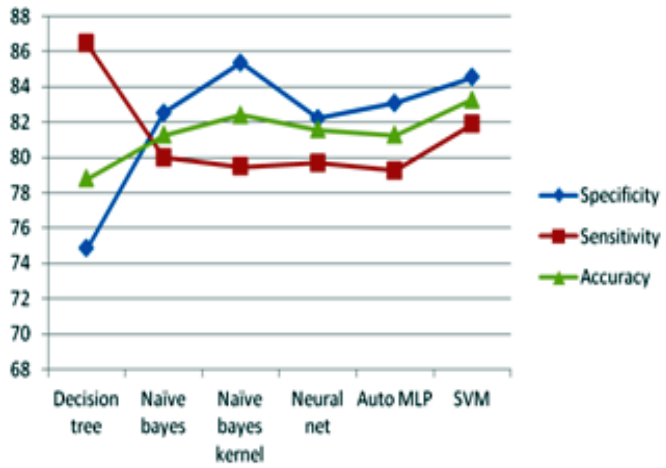

Figure 1. Classifiers with their accuracy, sensitivity and specificity

Figure 1 shows that the accuracy of naïve Bayes using kernel method, auto multilayer perceptron and support vector machine are comparable, though the accuracy is a tradeoff of sensitivity and specificity and it is depending on factors such as missing data, nature of data and distribution of labels across the feature vectors. In the dataset number of observation were 961 with 87 attributes were with missing, the support vector machine is best classifier with an accuracy of 83.25% and the area under the curve (AUC) as a performance measure is 0.8419 followed by naïve Bayes kernel when there was no missing data. If the missing data was present then naïve Bayes classifier predict the best accuracy.

Table 4. Comparative statistics of number of false negative cases and false positive cases

BIRADS Levels	SVM		AUTO MLP		Neural Net	
	FN	FP	FN	FP	FN	FP
Level ≤ 2	1	0	2	3	2	1
Level ≤ 3	6	0	6	1	6	6
Level ≤ 4	112	6	66	41	9	42
Level ≤ 5	1	43	3	44	7	44
Wrong Biopsies	13.06%		38.10%		40.58%	

Table 4 shows that the false positive cases (saying you have a disease when you don't) is exactly 0 in level 2 and 3, where in Auto MLP and neural network have some values. More importantly for level 4, SVM gives only 6 as false positive in comparison to other classifier. And at last level 5 SVM predictions is equal with Auto MLP and NN classifier. Because at level 5 chances of being malignant tissue is around 100%. So at level 5 the surgical biopsies could not be avoided, but at level 4 the performance of SVM is best to reduce or avoid the biopsies.

6 CONCLUSION

Using support vector machine classifier the number of unnecessary surgical biopsies could be reduced. The prediction of breast biopsies resulting from mammogram interpretation leads to approximately reduce the unnecessary biopsies by 13.06% in support vector machine as compared to neural network. An automated method has been proposed to prevent the unnecessary surgical biopsies using support vector machine has significantly reduced the number of the wrong

biopsies as compared to other classifiers. Therefore an aid has been provided to the physician in the prognosis of mammographic interpretation using SVM.

7 REFERENCES

[1] Nasseer M. B. and Mustafa H. M., "Classification of Breast Masses in Digital Mammograms Using Support Vector Machines", International Journal of Advanced Research in Computer Science and Software Engineering, 3(10), 57-63 (2013)

[2] Chaurasia S., Chakrabarti P. and Chourasia N., "An Application of Classification Techniques on Breast Cancer Prognosis",International Journal of Computer Applications,59(3),6-10 (2012)

[3] Siegel R., Naishadham D. and Jemal A., "Cancer Statistics, 2013", CA: A Cancer Journal for Clinicians, 63(1), 1-30 (2013)

[4] Chaurasia S. and Chakrabarti P., "An Approach with SVM using Variable Feature Selection on Breast Cancer Prognosis" ,International Journal of Advanced Research in Artificial Intelligence,2(9),38-42 (2013)

[5] American Cancer Society Hompage, http://www.cancer.org/ 2008

[6] Quinlan, J. R., "Improved use of continuous attributes in C4.5," Journal of Artificial Intelligence Research, 4(1), 77-90 (1996)

[7] Hamiton, H. J., Shan, N., & Cercone, N. (1996). RIAC: A rule induction algorithm based on approximate classification. In International conference on engineering applications of neural networks, University of Regina.

[8] Nauck D. and Kruse R. , "Obtaining interpretable fuzzy classification rules from medical data", Artificial Intelligence in Medicine, 16(1), 149–169 (1999)

[9] Abonyi J. and Szeifert F. "Supervised fuzzy clustering for the identification of fuzzy classifiers". Pattern Recognition Letters, 24(1), 2195-2207 (2001)

[10] Albrecht A. A., Lappas G., Vinterbo S. A., ., Wong C. K. and Ohno-Machado L., Two applications of the LSA machine, In Proc. ICONIP '02, November 18-22 (2002) p. 184.

[11] Guijarro-Berdias, B., Fontenla-Romero, O., Perez-Sanchez, B., & Fraguela, P. (2007). A linear learning method for multilayer perceptrons using leastsquares. Lecture Notes in Computer Science, 365–374. 10.1007/978-3-540-77226-238.

[12] Karabatak M. and Cevdet-Ince M., "An expert system for detection of breast cancer based on association rules and neural network," Expert Systems with Applications, 36(1), 3465–3469 (2009)

[13] Kotsiantis S. B., "Supervised machine learning: A review of classification techniques", Informatica,31,249-268(2007)

[14] Murthy, (1998), Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey, Data Mining and Knowledge Discovery 2: 345–389.

[15] Quinlan, J.R. (1993). C4.5: Programs for machine learning. Morgan Kaufmann, San Francisco

[16] Parzen E. On estimation of a probability density function and mode. Ann. Math. Stat.1962; 33:1065-1076.

[17] Neocleous, C. & Schizas, C., (2002), Artificial Neural Network Learning: A Comparative Review, LNAI 2308, pp. 300–313, Springer-Verlag Berlin Heidelberg.

[18] Vapnik V. The nature of statistical learning Theory, 2nd Ed. Springer, NewYork, 1999.

[19] A. Marcano-Cedeño, J. Quintanilla-Domínguez, D. Andina, WBCD breast cancer database classification applying artificial metaplasticity neural network, Expert Systems with Applications, Volume 38, Issue 8, August 2011, Pages 9573-9579, ISSN 0957-4174