Comparative Analysis of Different Classifiers for the Wisconsin Breast Cancer Dataset

Leena Vig

University of Delhi, Delhi, India
Email: leevig@gmail.com

Received 6 July 2014; revised 16 August 2014; accepted 19 September 2014

Abstract

The Wisconsin Breast Cancer Dataset has been heavily cited as a benchmark dataset for classification. Neural Network techniques such as Neural Networks, Probabilistic Neural Networks, and Regression Neural Networks have been shown to perform very well on this dataset. However, despite its obvious practical importance and implications for cancer research, a thorough investigation of all modern classification techniques on this dataset remains to be done. In this paper we examine the efficacy of classifiers such as Random Forests with varying number of trees, Support Vector Machines with different kernels, Naïve Bayes model and neural networks on the accuracy of classifying the masses in the dataset as benign/malignant. Results indicate that Support Vector machines with a Radial Basis function kernel give the best accuracy of all the models attempted. This indicates that there are non-linearities present in the dataset and that the Support vector machine does a good job of mapping the data into a higher dimensional space in which the non-linearities fade away and the data becomes linearly separable by large margin classifier like the support vector machine. These methods show that modern machine learning methods could provide for improved accuracy for early prediction of cancerous tumors.

Keywords

WBCD, Machine Learning, Classification, Naive Bayes, Neural Networks, SVM

Subject Areas: Artificial Intelligence, Bioinformatics

1. Introduction

Breast cancer remains one of the biggest health concerns for women across the world today. Although the risk factors vary with race, location, lifestyle and diet, it remains one of the biggest health concerns for women across the globe. According to an estimate by the National Cancer Insititute of the United States, 13.4 per cent of women born today will be diagnosed with breast cancer at some stage of their lives [1]. Many techniques for

How to cite this paper: Vig, L. (2014) Comparative Analysis of Different Classifiers for the Wisconsin Breast Cancer Dataset. *Open Access Library Journal*, **1**: e660. http://dx.doi.org/10.4236/oalib.1100660

the diagnosis and prognosis of breast cancer have been discussed [2]-[8]. However, even today the method that can confirm malignancy accurately with a high sensitivity is a surgical biopsy, a costly and painful procedure. To this end modern classification techniques attempt to replicate the accuracy of a biopsy, without the negative aspects of a surgical biopsy.

This paper looks at the breast cancer diagnosis problem using the Wisconsin Diagnostic Breast Cancer (WDBC) data set which is available publicly on the web [9]. The data set involves recordings from a Fine Needle Aspirate (FNA) test. The aim of the classification is to provide a distinction between the malignant and the benign masses. The WDBC dataset is the result of the efforts made at the University of Wisconsin Hospital for the diagnosis of breast tumours solely based on FNA test. This test involves fluid extraction from a breast mass using a small-gauge needle and then visual inspection of the fluid under a microscope. **Figure 1** depicts two images, which were taken from fine needle biopsies of breast tumours [10]. **Figure 1(a)** shows a benign instance and **Figure 1(b)** a malignant tumour image. These sample images are provided as part of the dataset. Several modern classifiers were evaluated on this benchmark dataset including Support Vector Machines (SVMs), Neural Networks, Random Forests and Naïve Bayes models.

2. Materials and Methods

2.1. The WDBC Dataset

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset consist of 569 observations with 357 negative (benign) and 212 positive (malignant) observations, where each one represents FNA test measurements for one sample. Each observation has 32 attributes, where the first two attributes correspond to a unique identification number and the diagnosis status (benign/malignant). The remaining 30 features are values of ten real-valued features, along with their mean, standard error and the mean of the three largest values ("worst" value) for each cell nucleus respectively. These ten real values, which are depicted in **Table 1**, are computed from a digitized

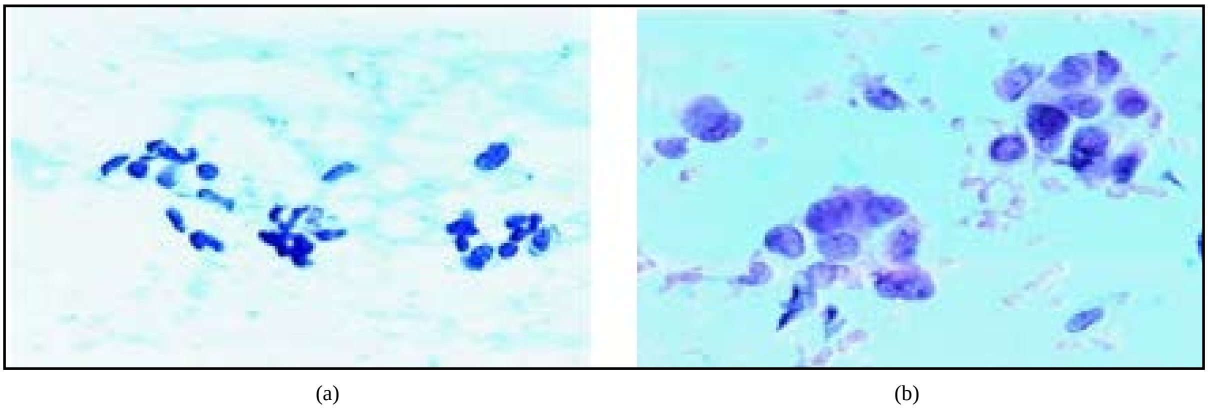

Figure 1. Images taken using the FNA test: (a) Benign; (b) Malignant. Images reproduced from [11].

Table 1. WDBC/WPBC cell nuclei characteristics attributes.

Cell Nuclei Characteristics
1) radius [mean of distances from centre to points on the perimeter],
2) Texture [standard deviation of grey scale values],
3) perimeter,
4) area,
5) smoothness [local variation in radius lengths],
6) compactness[(perimenter)^2/area −1],
7) concavity [severity of concave portions of the contour],
8) concave points [number of concave portions of the contour],
9) symmetry,
10) fractal dimension ["coastline approximation" −1].

image of a fine needle aspirate (FNA) of breast tumor, describing characteristics of the cell nuclei present in the image and are recorded upto four significant digits. For the current problem, the WDBC dataset was used in several publications in the medical literature [10]-[15]. In addition, due to its consistency and robust creation, this dataset has also been used for verification purposes over the classification or prediction performance of information systems in other scientific areas [16] [17].

2.2. Neural Networks

The field of neural networks was initiated due to the efforts by early computer scientists to replicate the neurons in the human brain in an effort to mimic human intelligence. The perceptron [18] by Rosenblatt was the first neural network architecture that was capable of learning a hyperplane for classification. However, the limitation of the perceptron as pointed out by Minsky and Pappart [19] was that it could only learn boundaries for linearly seperable datasets. This limitation was wrongly construed to imply that all neural network architectures suffered from this limitation and this led to a drop in research funding for artificial neural networks. However, in 1985 Rummelhart, Hinton rediscovered the back propagation algorithm [20] and this allowed feedforward neural architectures to learn non-linear decision boundaries. In fact Kolgomorovs theorem [21] shows that any function can be learnt using a three layer feedforward neural network.

Figure 2 shows a typical three layer feedforward network where each layer has a set of independent units that receive weighted inputs from the preceding layer (and direct inputs in the case of the first input layer). The inputs are then combined and an activation function is applied to the weighted sum of the inputs and this is the output of the unit that is sent to the next successive layer (or to the outputs in the case of the output layer). The weights between units in successive layers are the parameters to the network. In order to learn these parameters the backpropagation algorithm is employed.

Backpropagation works by modifying the weights of the feedforward network in proportion to the error sensitivity of the network with respect to a particular weight. The weight update rule for a weight w_{ij}, also called the delta rule is given by:

$$\Delta w_{ij} = \alpha \frac{\mathrm{d}E}{\mathrm{d}w_{ij}} \tag{1}$$

where α is the learning rate.

For the given WDBC dataset the inputs are the features and the outputs will have two units corresponding to benign and malignant instance. The correct output unit will have a value 1 and the incorrect output should have a value 0. After training the test set error is evaluated as explained in Section 3.

2.3. Random Forests

Random forest are an ensemble learning method for classification that combine the predictions of several decision trees. Random forests grow a forest of decision trees where each tree is mapped onto a different feature space. Leo Breiman. [22] introduced the idea of random forests [23]. The paper describes a method of building a

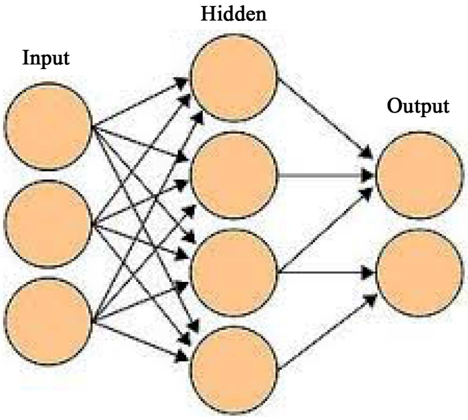

Figure 2. A three layer neural network.

forest of uncorrelated trees using a CART [24] like procedure, combined with randomized node optimization and bagging [25].

The technique of bootstrap aggregating, or bagging, is applied for random forests to generate tree learners. Given a training set $X = x_1, \cdots x_n$ with responses $Y = y_1$ through y_n, bagging repeatedly selects a bootstrap sample of the training set and fits B trees to these samples:

For $b = 1$ through B:

1) Sample, with replacement, n training examples from X, Y; call these X_b, Y_b.

2) Train a decision or regression tree f_b on X_b, Y_b.

After training, predictions for unseen samples x' can be made by taking the majority vote of the predictions from all the individual trees on x': **Figure 3** shows the combination of a set of trees to generate a prediction membership probability.

2.4. Support Vector Machines

Support Vector Machines are an example of a large margin classifier, utilizing the notion of functional and geometric margins. The geometric margin is simply the smallest distance of the decision boundary from all of the data points in either class. The closest points to the decision boundary are called support vectors. The objective is to find the decision boundary that maximizes the geometric margin. The traditional method to achieve this is by reducing the problem to a convex optimization problem, formulating the dual of the problem and solving it using an efficient algorithm called the SMO algorithm [26]. **Figure 4** shows how a large margin classifier operates while separating two classes of data and how a larger margin allows for reduced overfitting.

In order to map the input to a higher dimension often a function called the kernel function is employed. This function allows for an efficient mapping that can be computed efficiently in terms of the inner product for solving the optimization problem to obtain the optimal geometric margin. This is often called the kernel trick [27]. Many kernel functions exist such as the linear, radial basis function and quadratic kernels used in this paper.

2.5. Naïve Bayes Classifier

This is a generative model of classification, which means it does not directly attempt to predict the probability of the class given the data, but attempts to model the data itself and then uses Bayes theorem to extrapolate the results.

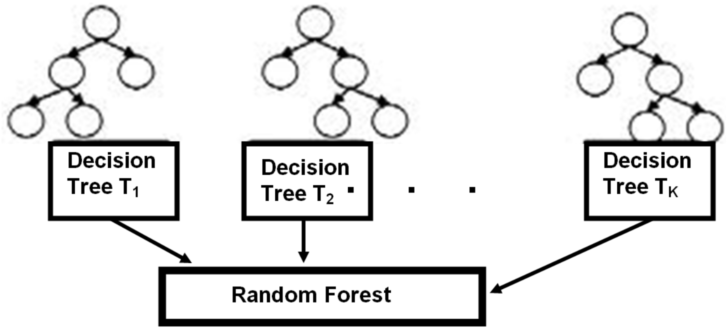

Figure 3. A random forest classifier.

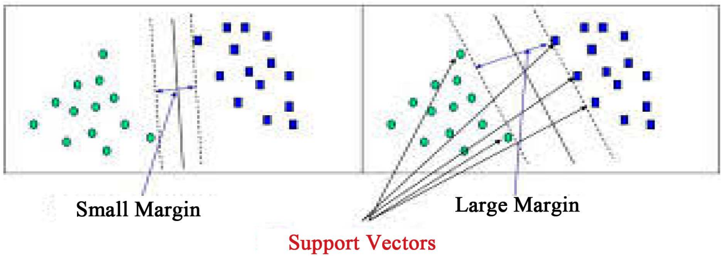

Figure 4. Margins and support vectors for a two class problem.

$$P(y \mid x) = \frac{P(x \mid y) P(y)}{P(x)} \tag{2}$$

Here a prior probability over the class to be predicted is evaluated *i.e.* *P*(*y*) is estimated from the training data as the probability of any instance in the training data being positive or negative, subsequently the probability of the feature vector *x* is evaluated given the value of *y*. This is where the Naïve Bayes assumption plays a role, the Naïve Bayes assumption implies that all of the features of x are independent given the class of the feature vector *y*. The probabilities of each of the features of *x* can be easily evaluated by only considering the instances with the correct value of *y* and the proportion of each feature value in the set. The denominator in Equation (2) need not be evaluated as it is simply a normalizing constant.

3. Experiments

The data was split into a training set (70%), testing set (15%) and validation sets (15%) several times and a Double Cross Validation (DCV) Approach was utilized to evaluate the accuracy, sensitivity and specificity of each of the classifiers. The proportion of positive and negative examples was kept the same in each of the splits. Simple Cross Validation (CV) has been shown to have a high bias and is therefore not considered an adequate measure of classification accuracy. In DCV we divide the dataset several times into a test set and a training set, and perform a K(10) fold cross validation within the training set and use the model generated to predict on the test set.

4. Results

Experimental Results for Accuracy, Sensitivity, Specificity are provided in **Table 2**. From the table it is evident that a random forest classifier with 100 decision trees provides the best results with 95.64% accuracy. The true positive rate and true negative rate are also very accurate for this classifier. Support vector machines seem to also work well for this kind of classification and ANN's tend to give a high accuracy as well. Naïve Bayes does quite poorly in comparison to the other classifiers and reason could be that the naïve bayes assumption is too strong for this dataset.

Figure 5 shows the Receiver Operating Characteristics (ROC) curve of the different classifiers and the curves of the Random Forest Classifier have much greater area than the other classifiers.

5. Conclusion

The WBCD dataset has been at the focus of several research efforts aimed at improving accuracy for Breast Cancer detection. However, to the best of our knowledge a comprehensive analysis using the latest techniques had not been performed. This paper presents an analysis using Random Forest classifiers, Artificial Neural Networks, Naïve Bayes and Support Vector Machines. Results show that ANN's, Random Forests and SVMs are able to yield models with high accuracy, sensitivity and specificity whereas Naïve Bayes performs poorly.

Table 2. Accuracy, sensitivity, specificity for the different classifiers.

Classifier	Accuracy (%)	TPR	TNR
SVM (linear)	78.45	0.72	0.81
SVM (rbf)	93.63	0.94	0.92
SVM (quadratic)	72.93	0.68	0.75
Random Forest (100)	95.64	0.97	0.94
Random Forest (10)	90.13	0.92	0.89
ANN	92.44	0.93	0.92
Naïve Bayes	65.27	0.57	0.72

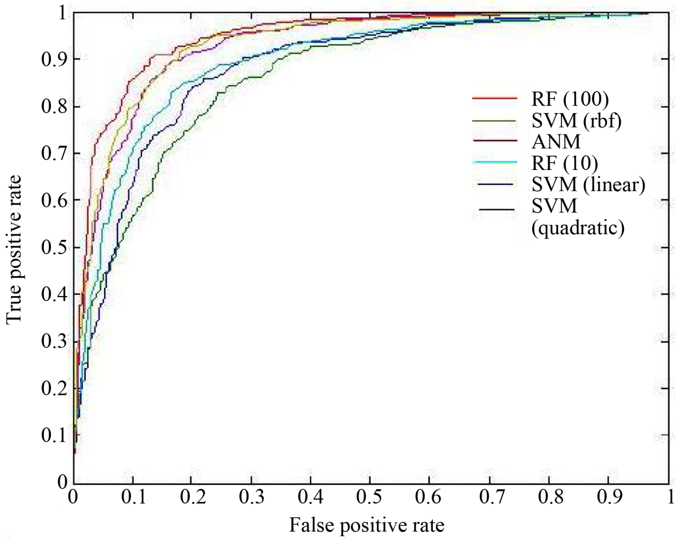

Figure 5. ROC curves for the different classifiers.

These could be incorporated in the medical research field to potentially benefit patients suspected of having breast cancer.

References

[1] Estimated New Cancer Cases and Deaths for 2004. http://seer.cancer.gov/cgi-bin/csr/1975_2001/search.pl#results

[2] Wang, T.C. and Karayiannis, N.B. (1998) Detection of Microcalcifications in Digital Mammograms Using Wavelets. *IEEE Transactions on Medical Imaging*, **17**, 49-509. http://dx.doi.org/10.1109/42.730395

[3] Huo, Z., Giger, M., Vyborny, C., Wolverton, D., Schmidt, R. and Doi, K. (1998) Automated Computerized Classification of Malignant and Benign Mass Lesions on Digital Mammograms. *Academic Radiology*, **5**, 155-168. http://dx.doi.org/10.1016/S1076-6332(98)80278-X

[4] Cheng, H.-D., Lui Y.M. and Freimanis, R.I. (1998) A Novel Approach to Microcalcification Detection Using Fuzzy Logic Technique. *IEEE Transactions on Medical Imaging*, **17**, 442-450. http://dx.doi.org/10.1109/42.712133

[5] Pendharkar, P.C., Rodger, J.A., Yaverbaum, G.J., Herman, N. and Benner, M. (1999) Association, Statistical, Mathematical and Neural Approaches for Mining Breast Cancer Patterns, Expert Systems with Applications, 17, 223-232. DRAFT VERSION of paper to Appear at the Oncology Reports, Special Issue Computational Analysis and Decision Support Systems in Oncology, Last Quarter 2005.

[6] Setiono R. (2000) Generating Concise and Accurate Classification Rules for Breast Cancer Diagnosis. *Artificial Intelligence in Medicine*, **18**, 205-219. http://dx.doi.org/10.1016/S0933-3657(99)00041-X

[7] Chen, D., Chang, R.F. and Huang, Y.L. (2000) Breast Cancer Diagnosis Using Self-Organizing Map for Sonography. *Ultrasound in Medical Biology*, **26**, 405-411. http://dx.doi.org/10.1016/S0301-5629(99)00156-8

[8] Giger, M., Huo, Z., Kupinski, M. and Vyborny, C. (2000) Computer-Aided Diagnosis in Mammography. In: Sonka, M. adn Fitzpatrick, J., Eds., *Handbook of Medical Imaging*, *Medical Image Processing and Analysis*, *Vol.* 2, SPIE Press, 386-408.

[9] Wisconsin Diagnostic Breast Cancer (WDBC) Dataset and Wisconsin Prognostic Breast Cancer (WPBC) Dataset. http://ftp.ics.uci.edu/pub/machine-learning-databases/breast-cancer-wisconsin/

[10] Tourassi, G.D., Markey, M.K., Lo, J.Y. and Floyd Jr., C.E. (2001) A Neural Network Approach to Breast Cancer Diagnosis as a Constraint Satisfaction Problem. *Medical Physics*, **28**, 804-811. http://dx.doi.org/10.1118/1.1367861

[11] Wolberg, W.H., Street, W.N., Heisey, D.M. and Mangasarian, O.L. (1995) Computer-Derived Nuclear Features Distinguish Malignant from Benign Breast Cytology. *Human Pathology*, **26**, 792-796. http://dx.doi.org/10.1016/0046-8177(95)90229-5

[12] Wolberg, W.H., Street, W.N. and Mangasarian, O.L. (1994) Machine Learning Techniques to Diagnose Breast Cancer from Image-Processed Nuclear Features of Fine-Needle Aspirates. *Cancer Letters*, **77**, 163-171. http://dx.doi.org/10.1016/0304-3835(94)90099-X

[13] Wolberg, W.H., Street, W.N. and Mangasarian, O.L. (1995) Image Analysis and Machine Learning Applied to Breast

Cancer Diagnosis and Prognosis. *Analytical and Quantitative Cytology and Histology*, **17**, 77-87.

[14] Jiang, Y., Nishikawa, R., Wolverton, D., Metz, C., Giger, M.L., Schmidt, R. and Doi, K. (1996) Malignant and Benign Clustered Microcalcifications: Automated Feature Analysis and Classification. *Radiology*, **198**, 671-678. http://dx.doi.org/10.1148/radiology.198.3.8628853

[15] Taylor, P., Fox, J. and Todd-Pokropek, A. (1998) Evaluation of a Decision Aid for the Classification of Microcalcifications. In: *Digital Mammography*, Kluwer Academic Publishers, Nijmegen, 237-244.

[16] Hoya, T. and Chambers, J.A. (2001) Heuristic Pattern Correction Scheme Using Adaptively Trained Generalized Regression Neural Networks. *IEEE Transactions on Neural Networks*, **12**, 91-100. http://dx.doi.org/10.1109/72.896798

[17] Kaban, A. and Girolami, M. (2000) Initialized and Guided EM-Clustering of Sparse Binary Data with Application to Text Based Documents. 15*th International Conference on Pattern Recognition*, **2**, 744-747.

[18] Rosenblatt, F. (1958) The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. Cornell Aeronautical Laboratory, *Psychological Review*, **65**, 386-408.

[19] Minsky, M.L. and Papert, S.A. (1969) Perceptrons. MIT Press, Cambridge.

[20] Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) Learning Representations by Back-Propagating Errors. *Nature*, **323**, 533-536. http://dx.doi.org/10.1038/323533a0

[21] Kolmogorov, A.N. (1957) On the Representation of Continuous Functions of Many Variables by Superposition of Continuous Functions of One Variable and Addition. *Doklady Akademii Nauk SSSR*, **144**, 679-681. *American Mathematical Society Translation*, **28**, 55-59 [1963].

[22] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. http://dx.doi.org/10.1023/A:1010933404324

[23] Ho, T.K. (1995) Random Decision Forest. *Proceedings of the* 3*rd International Conference on Document Analysis and Recognition*, Montreal, 14-16 August 1995, 278-282.

[24] Chipman, H.A., George, E.I. and McCulloch, R.E. (1998) Bayesian CART Model Search. *Journal of the American Statistical Association*, **93**, 935-948. http://dx.doi.org/10.1080/01621459.1998.10473750

[25] Breiman, L. (1996) Bagging Predictors. *Machine Learning*, **24**, 123-140. http://dx.doi.org/10.1007/BF00058655

[26] Platt, J. (1998) Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, Advances in Kernel Methods. Support Vector Learning, MIT Press, Boston.

[27] Altman, N.S. (1992) An Introduction to Kernel and Nearest Neighbor Nonparametric Regression. *The American Statistician*, **46**, 175-185.