COSC174:MACHINE LEARNING PROJECT PROPOSAL

Breast Cancer Predication from Mammographic Findings

Yan Zhao, Rui Tian

October 28, 2014

1 PROBLEM MOTIVATION

Breast cancer is one of the most common cancers with an estimated 230,000 new cases in 2013, and over 1 million diagnosed cases around the world [1]. Mammography is a widely used imaging technique for breast cancer detection, with a reported specificity of 97% [2]. However, its relatively low sensitivity always results in unnecessary biopsies. In order to reduce the false positive rates of mammography, various computer aided detection (CAD) systems have been proposed recently. Machine learning algorithms such as artificial neural networks (ANN) have been used in the process of clinical decision-making. In this project, we propose to make model based prediction from mammographic findings and medical records and evaluate the performance of different models using receiver operating characteristics (ROC) curve analysis.

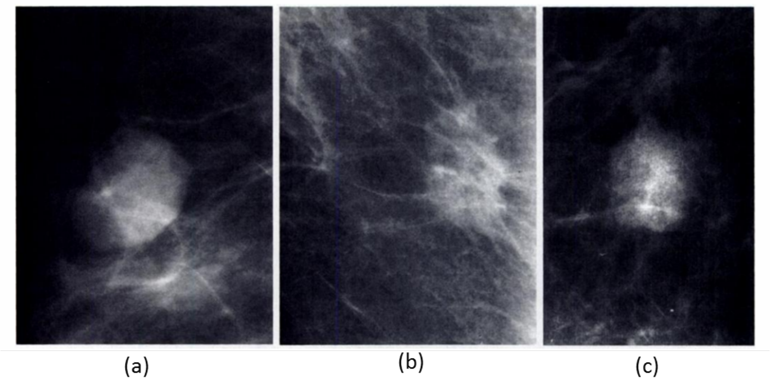

Figure 1.1: Mammographic images of (a) a benign lesion, (b) a malignant lesion, and (c) a mass described as "likely malignant". The lesions are described by mass, shape, position, et al. Figure courtesy of Jay A. Baker [3]

2 DATASET

We will use the well-known UCI machine learning repository [4]. Mammographic findings are standardized by radiologists in the format of BI-RADS score, which is an integer varying from 1 to 5, mass shape (integer), mass margin (integer) and mass density (integer). Furthermore, medical records including patient age and body mass index (BMI) are also available as input attributes. Severity, namely benign or malignant, which is confirmed by biopsy results, will be treated as the output attribute. The dataset including the input and output attributes of a large number of patients will be used for the model training and performance evaluation. Note that some attributes are missing and the effect of their absence will be investigated as well if time allows.

3 METHODS

We intended to implement three different models to make prediction based on the known data. They are Artificial Neural Network, Decision-Tree Learning and Bayesian Networks. These models have been widely applied to medical imaging process and proven to be effective models.

3.1 Random Forest

The concept of Random Forests was raised in Bell Lab [5]. A random forest consists of many decision trees. Each decision tree is generated by part of the initial training set. Both the training samples and the training features are randomly selected to build a decision tree. The advantages of random forest algorithm include

1) Random forests can perform good learning results even with missing data in training set.

2) In most cases, random forests will not have the issue of overfitting. When a training set of a learning problem is not very large, this advantage can lead to good learning results.

3) We do not need to prune the tree.

4) Random forests have accurate prediction results. Our experiment will show that random forests works better than young doctors with professional training and Bayesian Networks algorithm.

3.1.1 Decision-Tree Learning

As its name suggests, a Decision-Tree Learning (ID3) uses a decision tree to present a strategy of classifying and acquiring target values. A node in a decision tree represents a test for a certain attribute. The successor nodes of that node represent the possible values of that attribute. A leaf node in a decision tree represents a subclass. Decision-tree learning has been used to classify medical data and make diagnoses [6].

Figure 3.1 shows an example of a decision tree. The letter a, b, c, d, e denote some features. A sample will first be located at the root node. The conditional expressions along the edges are used to determine which path the sample should follow. At leaf nodes the decision tree will give its prediction for the sample.

3.1.2 Information Gain and Gini impurity

A essential issue for building a decision tree is to decide which feature should be used in the conditional expression at a given node. In other words, we need to figure out a feature to

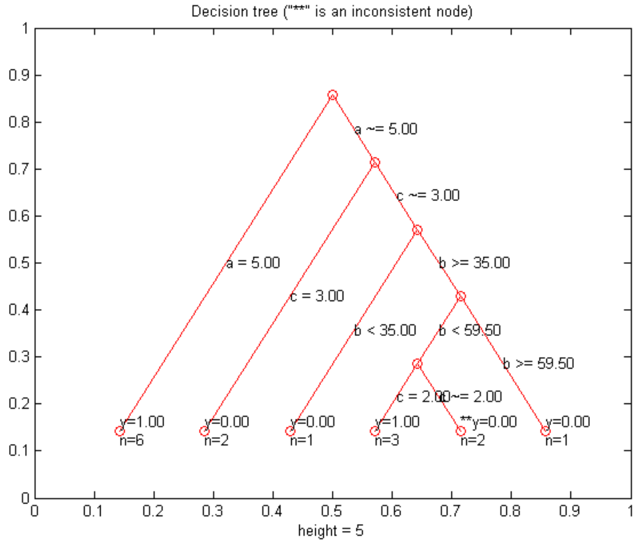

Figure 3.1: A demonstration of decision tree in MATLAB

split a node into two. The concept of information gain will address this issue. In general, we hope that we always pick a feature that is the most predictive to separate a group of samples. Information flow defines how much information a feature contains about the prediction results.

To quantify the information gain, there two popular ways to compute it. They are entropy and Gini impurity. Information entropy was raised by the founder of information theory, Cluade E. Shannon [7]. However, in our project, we used another way to quantify the information gain in building decision trees as Gini impurity works well for classification and regression tree. Gini impurity can also make it faster to get the proper feature to split on [].

To calculate the Gini impurity, suppose we have a set of labels $1, 2, ..., k$ and let f_i denote the fraction of the samples with labeled value i, then the Gini impurity are given by

$$I_G(f) = \sum_{i=1}^{k} f_i(1-f_i) = \sum_{i=1}^{k} f_i - \sum_{i=1}^{k} f_i^2 = 1 - \sum_{i=1}^{k} f_i^2 \tag{3.1}$$

Since decision trees are binary trees, we need to figure out a point to divide a particular feature. For discrete features, it is easy to compute all Gini impurities for each point. For continuous feature, the strategy is to use the values appeared in samples to divide the samples into two group and calculate Gini impurity respectively.

3.1.3 Build a random forest from decision trees

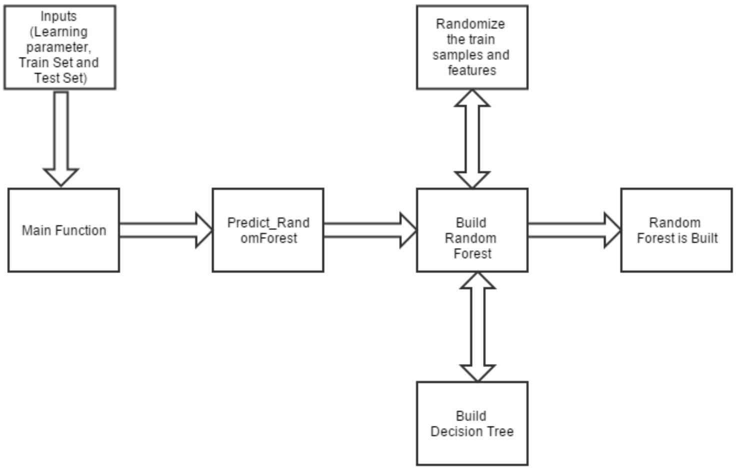

Figure 3.2: A demonstration of decision tree in MATLAB

Figure 3.2 illustrates the structure of our implementation. At first, we have a $main$ function to receive inputs, include learning parameter like the number of tress and the weighting factors, train set and test set. Then the $main$ function will call $Predict_RandomForest$ which is used save output results and generate figures. $BuildRandomForest$ is called to build a random forest consists of a number of decision trees. $BuildDecisionTree$ is used to build a single tree. A helper function is coded to randomized the features and samples of the given train set for a particular decision trees. Finally a random forest is built.

3.2 Bayesian Networks

Bayesian network utilizes the probability theory from Bayes' Rule, and it has various applications in the clinical decision-making. The ability to deal with uncertainty is one of the key advantages of Bayesian network. In this project, we will construct a Bayesian network for mammographic decision support.

4 Result Analysis

We first investigate the dependence of prediction accuracy on the number of trees selected in the Random Forest model. As shown in 4.1, the prediction accuracy is plotted versus number of trees from 10 to 100. We are able to get overall prediction accuracy above 80% using the random forest model, which is competitive to the prediction made by young radiology resident. Although it has been suggested that one could use as many trees as possible in the random forest model, there does not exist a monotonically increase of prediction accuracy with an increasing number of trees. The optimal number of trees (N=90) with highest prediction accuracy does not appear with highest number of trees. By further increasing N from 90 to 100, the prediction accuracy is weakened. It will be further investigated in our future work.

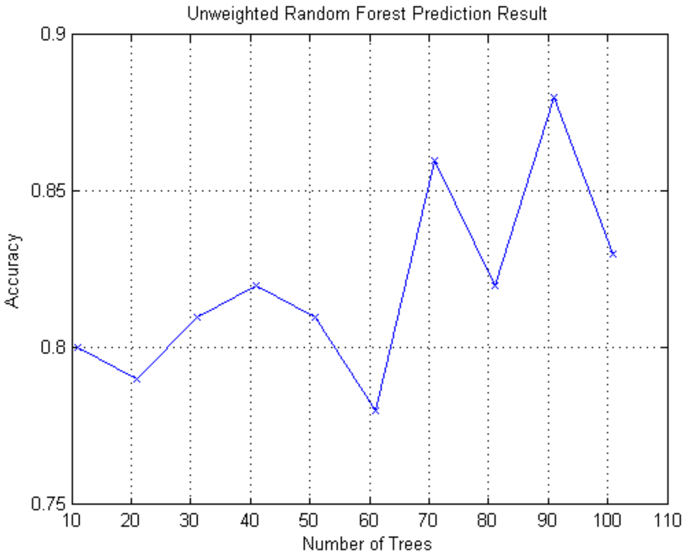

Figure 4.1: Accuracy versus number of trees, without weighting factors

Receiver operating characteristic (ROC) curve is a power tool to evaluate the performance of a prediction model in various biological and medicine applications. True positive rate (sensitivity) is plotted versus false positive rate (specificity), where a curve is fitted. The area under curve (AUC) is utilized as a metric to quantitatively depict the performance of a prediction/correlation model. We could get an AUC of 0.87 for N=70. By now we

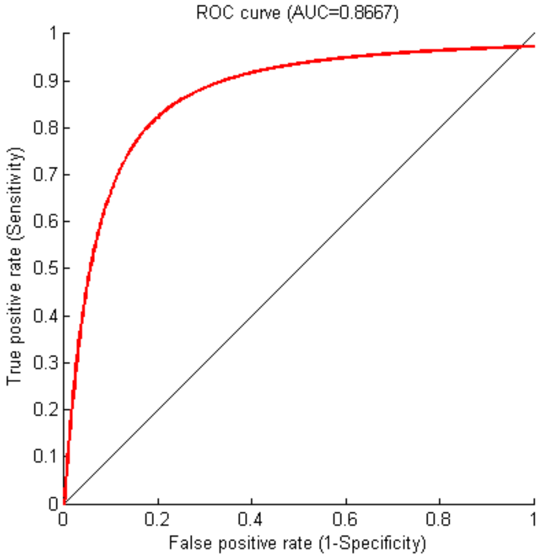

Figure 4.2: ROC curve for unweighted prediction accuracy with N = 70

have successfully implemented the random model in the breast cancer prediction, with competitive prediction accuracy. Next we tried to further improve the performance by adding a weighting factor to the input features. The input features in our problem include BIRADS score, age, shape, margin and density. By conducting statistical analysis of the input dataset, we find that the lesion is 90.2% malignant for the group with a BIRADS score of 5. While the rate of malignancy is only 16.7% for the group with a BIRADS score of 3. The left input features do not show as strong effect on the malignancy as BIRADS score. As a result, we expect the BIRADS score as the dominant decision factor for splitting when a tree is built. However, the patient age, which is an integer input varying from 18 to 96, turns out to be the most dominant factor for an unpruned tree in the random forest model. To reveal the relationship between input features and output label properly, we used a weighted input for the the random forest model training and prediction. As we can see in Figure 4.3, random forest is able to provide better overall prediction accuracy, with the addition of weighting factor into the input features, compared with that shown in Figure 4.1. AUC increases from 0.87 (Figure 4.2 to 0.95 (Figure 4.4), with 70 trees used in the random forest model.

Furthermore, we compared the performance of the prediction accuracy using random forest model and that using another widely used machine learning algorithm, Bayesian

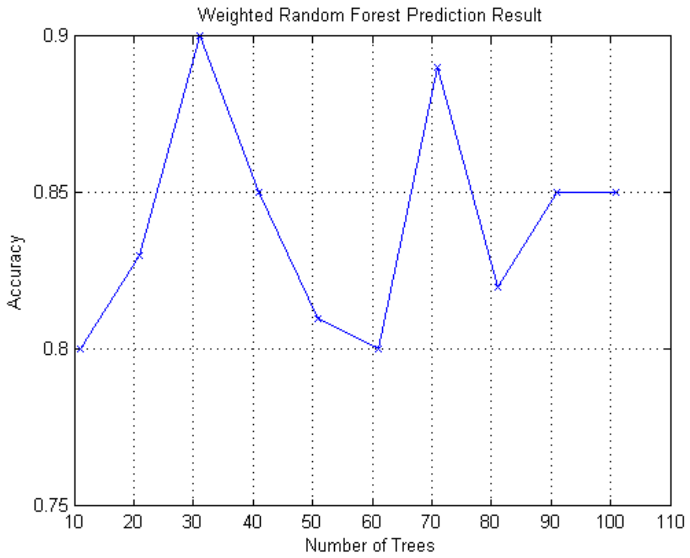

Figure 4.3: Accuracy versus number of trees, with weighting factors

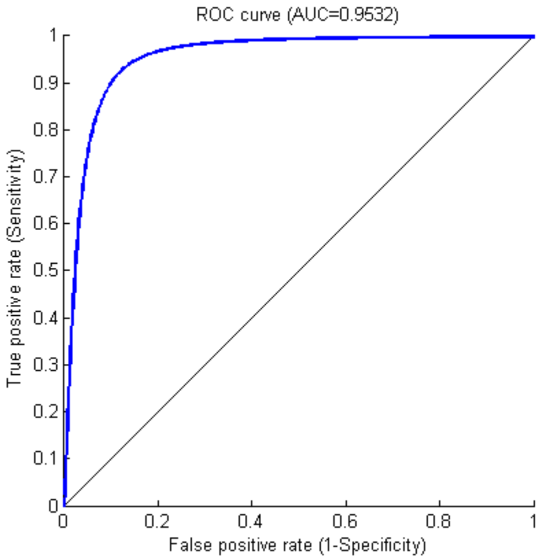

Figure 4.4: ROC curve for weighted prediction accuracy with N = 70

Network. First we randomly choose a fixed number of training samples from the whole dataset, and used the left samples as test samples. And then, we trained the random forest model and Bayesian Network model to make prediction on the test set. This process was repeated on 50 randomly selected training sets with corresponding test sets. As shown in Figure 4.5, we can find that the weighted random forest model is more likely (70%) to have higher prediction accuracy than Bayesian Network model. It demonstrates the benefit of using random forest model in a dataset with relatively small sample size.

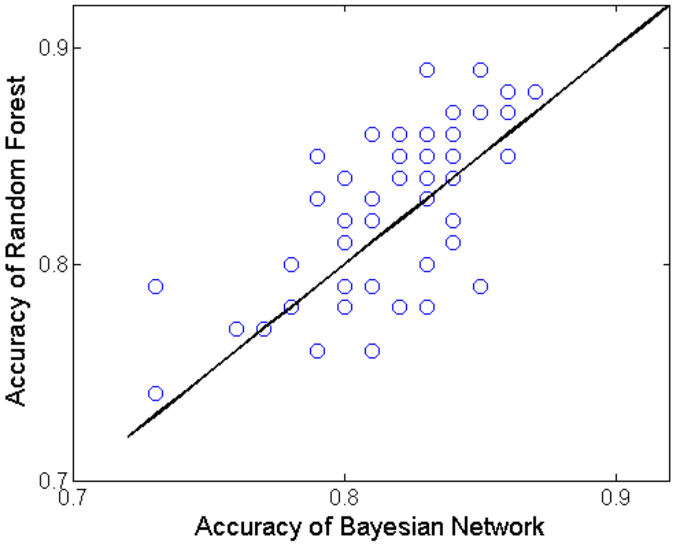

Figure 4.5: Accuracy of random forest is plotted versus that of Bayesian network. Blue circle represents a given training set with corresponding test set. random forest and Bayesian Network model has the same accuracy on the black line.

We have shown that a cursory chosen weighting factor can improve the prediction accuracy efficiently. We intend to pay more attention for choosing a proper weighting factor as this will combine prior knowledge and background information for our project.

The motivation of this project is to make model based prediction from mammographic findings and medical records. As we already know, it is most difficult for radiologists to make correct decision on whether the tumor is malignant or benign for the group with a BIRADS score of 3. From Table 4.1 we can find that even for the group with a BIRADS score of 3,

Table 4.1: Detailed information for the feature of BIRADS score

BIRADS score	2	3	4	5	5
Test frequency	11	34	589	356	8
Correct prection frequency	10	28	475	315	6
False positive	1	1	52	31	2
False negative	0	5	62	10	0

we are able to get a relatively low false positive rate (2.9%) and false negative rate (14.7%). The prediction accuracy is 82.4% for this group. The random forest model also has a higher accuracy for the other groups.

5 Future Work

5.1 Optimize The Computation Efficiency

Both the computation efficiency and the prediction accuracy are dependent on the number of decision trees but they are inconsistent. The tradeoff between the computation efficiency and prediction accuracy is worth more exploring. An intuitive principle is that we need to try our best to minimized the false negative results since that means we would treat a malignant tumor as benign.

More work are needed to optimize the computation efficiency includes improving the code and data storage. By now we investigated the running time versus the number of trees which seems to be the most critical parameter on the running efficiency. Besides, the number of trees is also a parameter that is chosen by users. Figure 5.1 presents these results.

5.2 Improve the prediction by optimizing the weighting factor

We will further correlate the performance of the weighted random forest model with input features such as BIRADS score. Model parameters such as number of trees and weighting factor will be optimized to get the optimal performance of a specific group with given input

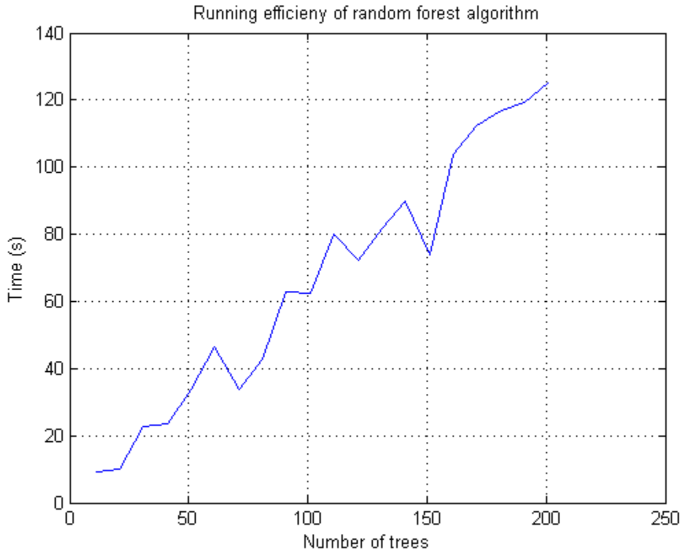

Figure 5.1: Running time versus the number of decision trees for random forest algorithm

features.

5.3 Compare Random Forest with Other Machine Learning Algorithms

Artificial Neural Network (ANN) is a computational model first introduced in 1943 but recently it attracted more and more attention. The typical structure of an ANN consists of a series of interconnected elements which can be called as nodes or neurons. These neurons can be capable of executing the necessary calculation based on activation function and input data in a parallel way [8]. In medical imaging, ANN is widely used in medical image processing and recognition for disease prediction and diagnose [9].

We also intend to investigate more information regarding this mature and widely used machine learning algorithm and compare its performance with random forest algorithm.

Reference

[1] American Cancer Society. Cancer facts and figures. 2013. `http://www.cancer.org/research/cancerfactsfigures/cancerfactsfigures/cancer-facts-figures-2013`.

[2] Per Skaane, Solveig Hofvind, and Arnulf Skjennald. Randomized trial of screen-film versus full-field digital mammography with soft-copy reading in population-based screening program: Follow-up and final results of oslo ii study 1. *Radiology*, 244(3):708–717, 2007.

[3] Jay A Baker, Phyllis J Kornguth, Joseph Y Lo, Margaret E Williford, and Carey E Floyd Jr. Breast cancer: prediction with artificial neural network based on bi-rads standardized lexicon. *Radiology*, 196(3):817–822, 1995.

[4] Catherine Blake and Christopher J Merz. {UCI} repository of machine learning databases. *University of California, Irvine, Dept. of Information and Computer Sciences*, 1998. `http://archive.ics.uci.edu/ml/`.

[5] Tin Kam Ho. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282. IEEE, 1995.

[6] Chin-Yuan Fan, Pei-Chann Chang, Jyun-Jie Lin, and JC Hsieh. A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. *Applied Soft Computing*, 11(1):632–644, 2011.

[7] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[8] IA Basheer and M Hajmeer. Artificial neural networks: fundamentals, computing, design, and application. *Journal of microbiological methods*, 43(1):3–31, 2000.

[9] Jianmin Jiang, P Trundle, and Jinchang Ren. Medical image analysis with artificial neural networks. *Computerized Medical Imaging and Graphics*, 34(8):617–631, 2010.