On Comparative Study of Breast Cancer Classification Using Ensembles in Statistical Modelling

[1]Milan Joshi, [2]Anurag Joshi
[1]Dept. of Applied Mathematics, MPSTME, NMIMS University, Shirpur, Maharashtra, India
[2]Dept. of Computer Engineering, MPSTME, NMIMS University, Shirpur, Maharashtra, India

Abstract
Distinguishing different stages in cancer for medical professionals will require a reliable prediction methodology to diagnose cancer. To name a few, breast cancer is one of the most ordinary disease among women that leads to death. Hence, diagnosing it in earlier stage has become the certainty in cancer research,as to provide succeeding clinical approach to the patients. The classification of breast cancer is based on large number of parameters that characterize the tumour's appearance. This helps the physician for diagnosis of breast cancer more easily. Hence automation of diagnostic system is needed for diagnosing tumours. Using automated computer tools and in particular machine learning to facilitate and boost medical investigation and detection is a promising and important area. A variety of these techniques, including Artificial Neural Networks (ANNs), Decision Trees (DTs) and Support Vector Machines (SVMs) and have been broadly applied in cancer research for the development of predictive models, resulting in effective and accurate decision making. Our Papers aims to use a modern and effective technique called Random Rotation Ensembles as well as Deep Learning .In this domain we show that the performance of this method is better than that of previous methods, therefore encouraging a more comprehensive and generic approach for cancer diagnosis.

Keywords
SVM, RF, DT, RRE, R Language

I. Introduction

Cancers figure among the leading causes of morbidity and mortality worldwide, with approximately 14 million new cases and 8.2 million cancer related deaths in 2012 (1). The number of new cases is expected to rise by about 70% over the next 2 decades. Among men, the 5 most common sites of cancer diagnosed in 2012 were lung, prostate, colorectum, stomach, and liver cancer.
Among women the 5 most common sites diagnosed were breast, colorectal, lung, cervix, and stomach cancer. Around one third of cancer deaths are due to the 5 leading behavioural and dietary risks: high body mass index, low fruit and vegetable intake, lack of physical activity, tobacco use, alcohol use. Tobacco use is the most important risk factor for cancer causing around 20% of global cancer deaths and around 70% of global lung cancer deaths. Cancer causing viral infections such as HBV/HCV and HPV are responsible for up to 20% of cancer deaths in low- and middle-income countries (2).More than 60% of world's total new annual cases occur in Africa, Asia and Central and South America. These regions account for 70% of the world's cancer deaths (1). It is expected that annual cancer cases will rise from 14 million in 2012 to 22 within the next two decades (1).

A. Stages & TNM

The stage of a cancer describes how far it has grown and spread at the time it is diagnosed. Stages are used to describe the spread of solid tumours, like breast, bowel or lung cancers. Blood cancers behave differently and are staged in different ways.

B. Cancer Staging

Staging is the process of measuring how far a cancer has spread when it is first diagnosed. It often involves having scans and other tests. Knowing the stage of a cancer is important as it helps doctors to work out the best treatment options. It also means the person with cancer can fully understand their situation and discuss any concerns they have. There are different staging systems for different cancers, but they generally use either the: Number Cancer Stage System, TNM System

C. Numbered System

The numbered system uses stage numbers to identify how far cancer has spread:

- **Stage 0 Cancer:** Often referred to as 'in-situ' cancer means the cancer cells are still in the place where they started and have not spread at all.
- **Stage 1 Cancer:** Is small and has only spread a little into nearby tissues. It has not spread to any lymph nodes or other body areas.
- **Stage 2 and 3 Cancer:** Means it is larger or has spread into nearby tissues or lymph nodes.
- **Stage 4 Cancer:** Has spread to other areas of the body. Stage 4 cancer is also called metastatic cancer or advanced cancer.

TNM System: In the TNM system, there are three categories: T = tumour N = lymph nodes M = metastases. Each of these categories is given a score, and together these scores show how far the cancer has spread.

II. Introduction

A. Random Forest

Among current classification algorithms Random forest is unexcelled in accuracy. However, it runs efficiently on large databases as it can handle thousands of input variables without variable deletion. It provides estimates of variables which are important in classification. As the forest building progresses it generates an internal unbiased estimate of the generalization error. By design, bagging lends itself nicely to parallelization. Hence, these methods can be easily applied on a very large dataset in a cluster environment. One of the major complaints against tree-based methods is the difficulty with pruning the trees to avoid over fitting. Big trees tend to also fit the noise present in the underlying data and hence lead to a low bias and high variance. However, when we grow a lot of trees and the final prediction is an average of the output of all the trees in the ensemble, we avoid these problems. In this paper, we will see a powerful tree-based

ensemble method called rotation forest. A typical random forest requires a large number of trees to be a part of its ensemble in order to achieve good performance. Rotation forest can achieve similar or better performance with less number of trees.

B. Rotation Forest

Rotation Forest[2] is a current proposed methodology for building ensemble classifiers which uses trained independent trees and is found to be more accurate than Ada Boost and Random forest on standard data sets. A random forest ensemble consists of decision trees trained on bootstrap samples from the data set. Additional diversity is introduced by randomising the feature choice at each node. Rotation forest draws upon the Random Forest idea. The base classifiers are also independently built decision trees, but in Rotation Forest each tree is trained on the whole data set in a rotated feature space.

C. Random Rotation Ensemble

In machine learning, ensemble methods combine the predictions of multiple base learners to construct more accurate aggregate predictions. Established supervised learning algorithms inject randomness into the construction of the individual base learners in an effort to promote diversity within the resulting ensembles. An undesirable side effect of this approach is that it generally also reduces the accuracy of the base learners. The idea of ensemble learning is to build a prediction model by combining the strengths of a collection of simpler base models. We have already seen a number of examples that fall into this category. Bagging and random forests[4] are ensemble methods for classification, where a committee of trees each cast a vote for the predicted class. Boosting in was initially proposed as a committee method as well, although unlike random forests, the committee of weak learners evolves over time, and the members cast a weighted vote. Stacking is a novel approach to combining the strengths of a number of fitted models. In fact one could characterize any dictionary method, such as regression splines, as an ensemble method, with the basis functions serving the role of weak learners. Bayesian methods for nonparametric regression can also be viewed as ensemble methods: a large number of candidate models are averaged with respect to the posterior distribution of their parameter settings (e.g. (Neal and Zhang, 2006)). Ensembles[9] are well established as a method for obtaining highly accurate classifiers by combining less accurate ones. This paper has provided a brief survey of methods for constructing ensembles and reviewed the three fundamental reasons why ensemble methods are able to out perform any single classifier within the ensemble. The paper has also provided some experimental results to elucidate one of the reasons why performs so well.

Table 1: Data Mining Techniques

Decision Trees		Support Vector Machine	
PROS	CONS	PROS	CONS
Intuitive Decision Rules	Highly biased to training set	Can handle large feature space	Not very efficient with large number of observations
Can handle non-linear features	No ranking score as direct result	Handle non-linear feature interactions	It can be tricky to find appropriate kernel sometimes
Variable interactions taken into consideration		Do not rely on entire data	

III. Data Mining Techniques

The data mining consists of various methods. Different methods serve different purposes, each method offering its own pros and cons as discussed in Table 1. However, most data mining methods commonly used for this paper are of classification category as the applied prediction techniques assign patients to either a ”benign” or a ”malignant” group and generate rules for the same.
The commonly used methods for data mining classification tasks can be classified into the following groups [4].

IV. Methodologies Used

We used following methodologies:

A. Data Collection

Wisconsin Dataset used which is referred from UCI repository as this dataset includes various attributes for which if it is analysed will give important results.

B. Data Storage

Data is first converted from .xls file to .csv as the tool used R will give desired results. Various pacakages are inbuilt in R studio viz. dplyr which is used for predictive modelling.

C. Data Processing

Various processing techniques include Data cleaning, Data Integration, Data transformation exists in machine learning but we had done all this in one package so as to optimize the results.

D. Software Version used

R 3.1.2 platform is used to carry out the results.

V. System Design

1. Formulate Problem

- Load libraries
- Load dataset
- Split-out validation dataset

2. Summarize Data

- Descriptive statistics
- Data visualizations

3. Prepare Data

- Data Cleaning
- Feature Selection
- Data Transforms

4. Evaluate Algorithms

- Test options and evaluation metric
- Spot Check Algorithms
- Compare Algorithms

5. Improve Accuracy

- Algorithm Tuning
- Ensembles

6. Finalize Model

- Predictions on validation dataset
- Create standalone model on entire training dataset
- Save model for later use

VI. Results

Data has been read thoroughly and converted to factor variable. We have carried out pre-processing through Caret package. We categorise the type of cancer as can be seen from Fig. 1, that proportion of Benign is more i.e. 62.7% as compared to Malign which is 37.3%.

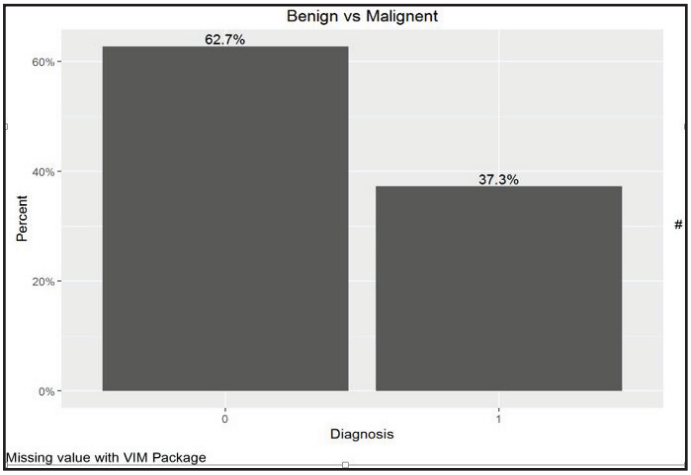

Fig. 1: Proportion of Benign vs Malignant

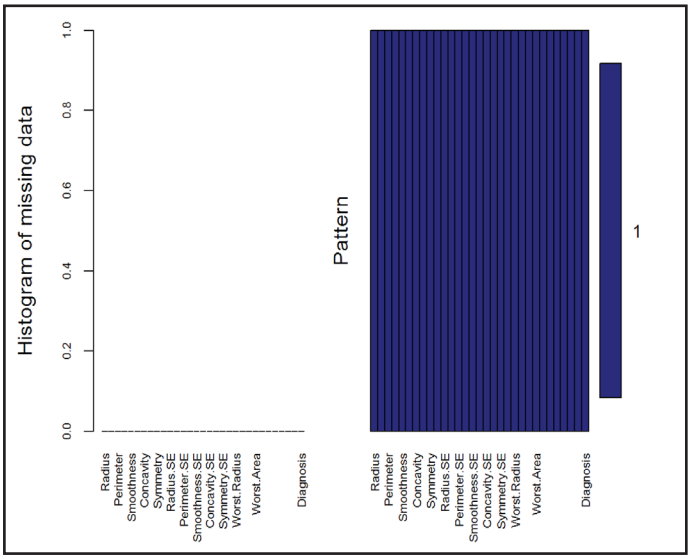

Fig. 2: Histogram Missing Value Representation

Histogram in fig. 2 showing the exact pattern through various pre-processing technique.

To make it more concrete we had applied predictive modelling techniques whose results are shown as below in Fig. 3, Fig. 4, Fig. 5.

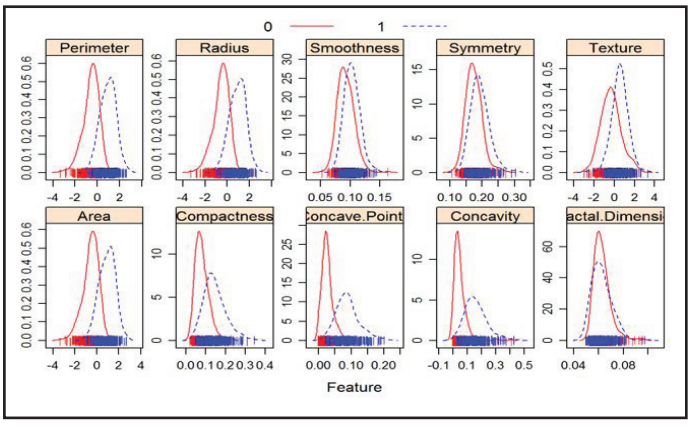

Fig. 3: Feature Plot

Curves above for individual attribute contributing to its categorisation i.e. Malign and Benign. Same has been carried out for other attributes as well.

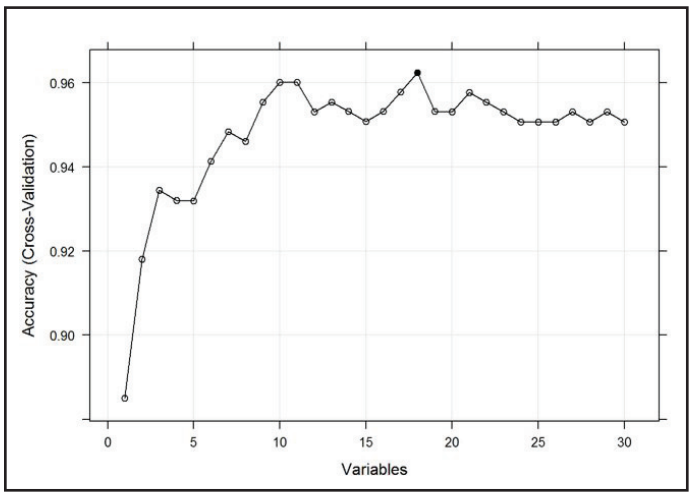

Fig. 4: Accuracy Curve

By summarizing the co-relation matrix accuracy curve is generated which is shown in Fig. 4, accuracy is 0.96 till 18th variable and further it has no effect.

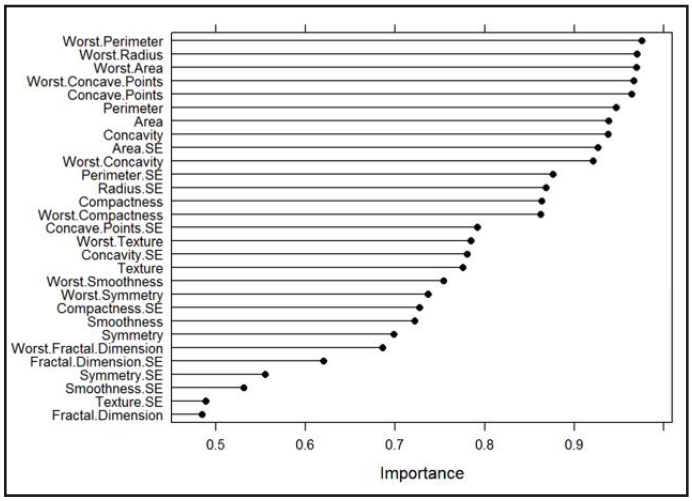

Fig. 5: Attribute Importance Chart

Fig. 5 shows that last four factor i.e. Fractional Dimension, Texture SE, Smoothness SE, Symmetry SE is not contributing to the model building and hence we ignored it in further processing.

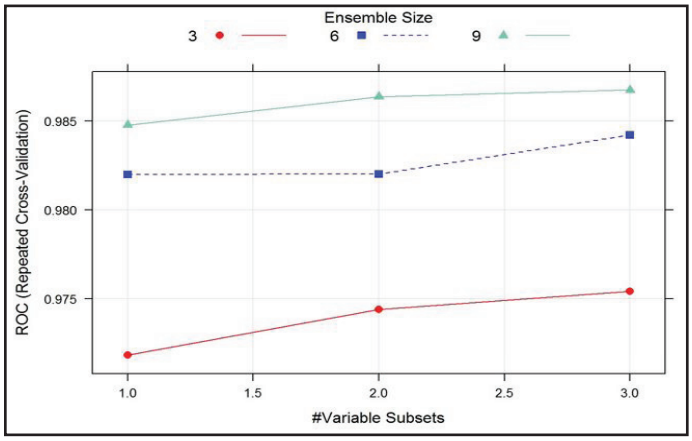

Fig. 6: Accuracy of Randomly Selected Predictors

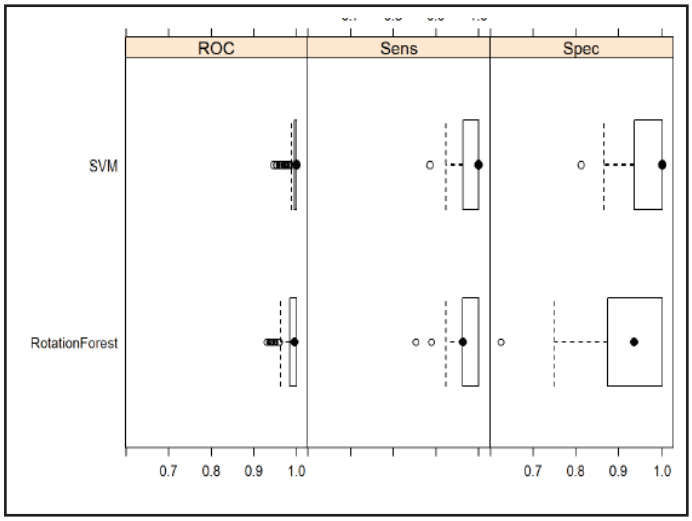

Fig. 7: ROC for SVM vs Rotation Forest

Fig. 7 shows statistics of attributes whose processing has been done through SVM[3] and Rotation forest[4] so as to check or identify which model is better. To make it more generic we will show confidence level which is 0.95 carried out by each model as shown in fig.1.8 named Confidence level modelling.

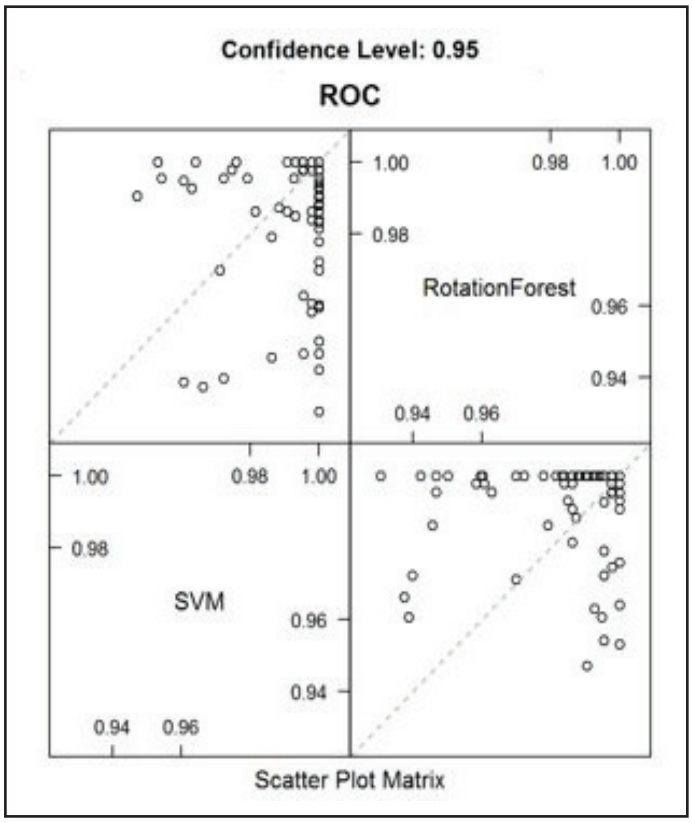

Fig. 8: Confidence Level Modelling

VII. Conclusions

Ensemble techniques are widely used in Machine learning now a days. We have used techniques viz. SVM and Rotation forest as to identify which model performs best on particular attributes. The obtained results shows that accuracy of Rotation Forest is better as compared to SVM .Hence proper usage of Machine Learning techniques will help in finding the proofs in order to find out attribute importance.

Techniques	Accuracy	Sensitivity	Specificity	K-Value
Rotation Forest	0.978	0.966	1.000	0.9554
Random Forest	0.954	0.976	0.976	0.901
Random Rotation Ensemble	0.9648	0.988	0.924	0.9239

References

[1] El-Sebakhy A. Emad, Faisal Abed Kanaan, Helmy T., Azzedin F., Al-Suhaim F.,"Evaluation of breast cancer tumor classification with unconstrained functional networks classifier," Computer Systems and Applications, IEEE International Conference, pp. 281 – 287, 2006.

[2] Rodriguez, Juan José, Ludmila I. Kuncheva, Carlos J. Alonso,"Rotation forest: A new classifier ensemble method", IEEE transactions on pattern analysis and machine intelligence 28.10, pp. 1619-1630, 2006.

[3] Liu, Kun-Hong, De-Shuang Huang,"Cancer classification using rotation forest", Computers in Biology and Medicine 38.5, pp. 601-610, 2008.

[4] Leslie, Christina S., Eleazar Eskin, William Stafford Noble, "The spectrum kernel: A string kernel for SVM protein classification", Pacific symposium on biocomputing. Vol. 7. No. 7. 2002.

[5] Abdelaal Ahmed Mohamed Medhat, Farouq Wael Muhamed, "Using data mining for assessing diagnosis of breast cnacer," In Proc. International multiconfrence on computer science and information Technology, pp. 11-17, 2010.

[6] Chang Pin Wei, Liou Ming Der,"Comparision of three Data Mining techniques with Ginetic Algorithm in analysis of Breast Cancerdata". [Online] Available: http://www.ym.edu.tw/~dmliou/Paper/compar_threedata.pdf

[7] Sudhir D., Ghatol Ashok A., Pande Amol P., "Neural Network aided Breast Cancer Detection and Diagnosis",7th WSEAS International Conference on Neural Networks, 2006.

[8] Gandhi Rajiv K., Karnan Marcus, Kannan S.,"Classification rule construction using particle swarm optimization algorithm for breast cancer datasets," Signal Acquisition and Processing. ICSAP, International Conference, pp. 233 – 237, 2010.

[9] "Random Rotation Ensembles" Journal of Machine Learning Research 17 (2016) 1-26Rico Blaser Piotr Fryzlewicz Department of Statistics London School of Economics Houghton Street London, WC2A 2AE, UK