ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

# Breast Cancer Survivability Predictor Using Adaboost and CART Algorithm

R.K.Kavitha[1], Dr.D.DoraiRangasamy[2]

Ph.D Research Scholar, Vinayagar Mission University, Tamilnadu, India[1]

Professor , Dept. of CSE, Vinayagar Mission University, Tamilnadu, India [2]

**ABSTRACT-Breast cancer is the second leading cancer for women in developed countries including India. Many new cancer detection and treatment approaches were developed, the cancer incidences and death of breast cancer decreased constantly. The patients are concerned about survival time after diagnosis in order to plan regarding their treatments. It is difficult for a physician to have accurate answers about prognosis. Data mining techniques are used to obtain useful information from the large amounts of data which can help the physician for decision making regarding the prognosis. This paper studies the performance comparison of Adaboost algorithm which classifies data as linear combination and CART (Classification and regression trees) which classifies data by constructing decision tree in predicting the survivability of breast cancer patients.**

**Keywords— prognosis, Adaboost, survival**

## I INTRODUCTION

Breast cancer is the second most common cause of cancer death in women in developed countries. The most effective way to reduce breast cancer deaths is detect it earlier. Many treatments have developed to reduce the number of mortalities and increase the survival time for patients. In order to predict the survivability of cancer patients, data mining algorithms obtain useful information from the large amounts of data which helps the physician for decision making regarding the prognosis. This paper compares the performance of Adaboost algorithm and CART (Classification and regression trees) algorithm.

## II ADABOOST

As a successor of the boosting algorithm, it is used to combine a set of weak classifiers to form a model with higher prediction outcomes .AdaBoost is the most popular ensemble method and has been shown to significantly enhance the prediction accuracy of the base learner. With this method, medical practitioners are able to focus on finding weak learning algorithms that only should be better than the original algorithm (weak learner). It is a learning algorithm used to generate multiple classifiers and to utilize them to build the best classifier. AdaBoost technique has become an attractive ensemble method in machine learning since it is low in error rate, performing well in the low noise data set. The advantage of this algorithm is that it requires less input parameters and needs little prior knowledge about the weak learner. As a result, several research studies have successfully applied the AdaBoost algorithm to solve classification problems in object detection, including face recognition, video sequences and signal processing systems. AdaBoost algorithm is not only used for predicting in Classification tasks, but also for presenting self-rated confidence scores which estimate the reliability of their predictions. This algorithm requires user less knowledge of computing in order to improve accuracy of models over data sets.

## III CLASIFICATION AND REGRESSION TREE

CART stands for Classification And Regression Trees, a decision-tree procedure representing a classification system or predictive model introduced in 1984 by statisticians, Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone. . CART builds classification and regression trees for predicting continuous dependent variables and categorical or predictor variables, and by predicting the most likely value of the dependent variable. The decision tree produced by CART is strictly binary, contain two branches of each decision node. CART recursively partitions the record into subsets of record with similar values of target attributes. The CART algorithm grows the tree by conducting for each decision node, an exhaustive search of all variables and all possible

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

splitting values, selecting the optimal split. It gives an estimate of the misclassification rate.

## IV COMPARISON OF ADABOOST AND CART

Gentle AdaBoost, which is originated from the setting of weights over the training set. The training set (x1,y1),…(xn,yn) where each xi belongs to instance space X,and each label yi is in the label set Y, which is equal to the set of {-1,+1}. It assigns the weight on the training example i on round k as Dk(i). The same weight will be set at the starting point (Dk(i)=1/N, i=1,…,N). Then the weight of the misclassified example from base learning algorithm (called weak hypothesis) is increased to concentrate the hard examples in the training set in each round.

The AdaBoost algorithm is presented in seven steps below:

1) Assign N example
(x1,y1),..,(xn,yn); xi $\in$ X , yi $\in$ {-1,+1}

2) Initialise the weights of D1(i)=1/N, i=1,…,N

3) for k=1,…,K

4)Train weak learner using distribution Dk

5)Get weak hypothesis hk:X R with its error :
$\varepsilon k = \sum Dk(i)$
6) choose $\varepsilon k = R$

7) Output the final hypothesis:

$$H(x) = \text{sign}\left(\sum_{k=1}^{K} \alpha_k h_k(x)\right)$$

Classification and regression tree (CART) is a rule-based method that generates a binary tree. through a binary recursive partitioning process that splits a node based on the yes and no answer of the predictors. Although some variables may be used many times, others may not be used at all. A single variable is used to split the tree by using purity criterion. The rule generated at each step is to maximize the class purity within the two resulting subsets. Each subset is split further based on the independent rules to find the threshold among the descriptive variables at the node of all dimensions and they separate the training sample with least error. The steps of constructing the tree included:
1.Create root node;
2.Select leaf with Largent error;
3.Create node, using only those training samples, that are associated with the chosen leaf;
4.Replace selected leaf with created node;
5. Repeat 2-4 until leaves become zero

## V SRATIFIED 10 FOLD CLASS VALIDATION

Stratified 10-fold cross-validation is a common validation method used to minimise bias and variance associated with the random sampling of the training and test sets. Moreover, it is a popular method for data selection in data mining related to medical research. In this study the process of stratified 10-fold cross-validation consists of four steps:
1) divide the data set into a set of subclasses;
2) assign a new sequence number to each set of subclasses;
3) randomly partition the subclass into 10 subsets or folds and;
4) combine each fold of each subclass into a single fold. Therefore, the size of each single fold is approximately equal to that of the original data set.

## VI EXPERIMENTAL ANANLYSIS

In this study, the models were evaluated based on performance measures including accuracy, sensitivity and specificity. The results were achieved by using stratified
10-fold cross-validation for each model, and were averaged from the test set (the remaining fold), for each fold. Our experiments were done in MATLAB 7 release 14 with GML AdaBoost MATLAB. The experiment results show that the accuracy of the Real and Gentle are decreasing rapidly. Comparing of accuracy, sensitivity and specificity for each classifier was measured by using our breast cancer data set. The same training and test. sets were utilized in all experiments with stratified10-fold cross-validation and selecting 10 iterations, in order to compare the performance of classification tasks. The experiment results show that Modest AdaBoost outperforms Bagging,

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

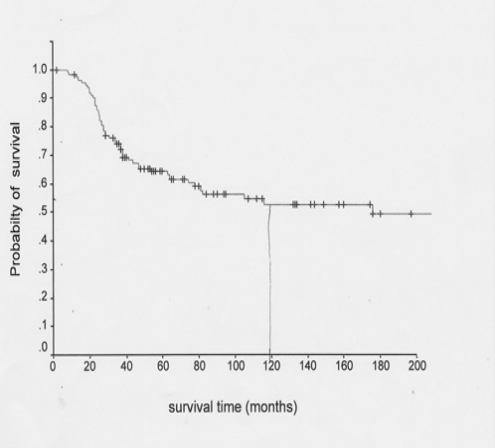
International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

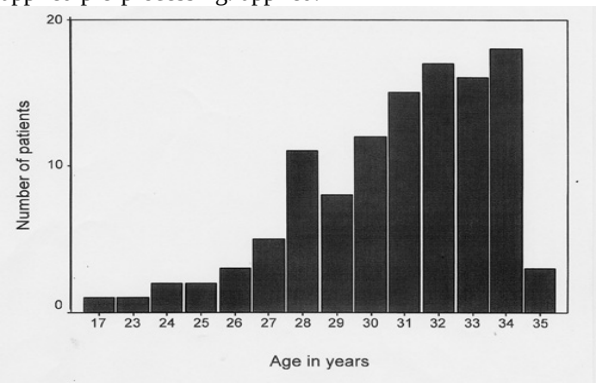
On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

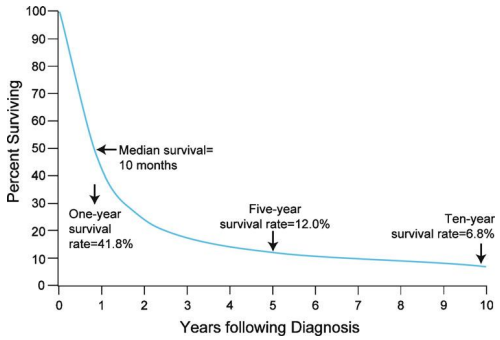



Survival of patients treated for breast cancer (Kaplan-Meier analysis).

In using AdaBoost algorithms to extract breast cancer survivability patterns in breast cancer databases at Hospital, we have successfully utilized stratified 10-fold cross-validation to divide the data set into 10 groups, with the same number in each class. Then presented the accuracy, sensitivity and specificity of classifiers in breast cancer survivability. We found that the accuracy and sensitivity of the models generated from Modest algorithm slightly improved (about 4%) after applied pre-processing. applied.




Presentation of patients with breast cancer by age.




## VII CONCLUSION

Adaboost was introduced to achieve better accuracy. Experimental results conducted on the collected breast cancer data set demonstrated the effectiveness of the proposed techniques. Particularly, the proposed method mainly aims to predict the single-point in unknown data rather than estimate multiple-point survival rate in known data, which is usually done by Kaplan-Meier curve. This paper is expected to be of benefit for medical decision making systems to give an alternative choice for medical practitioners to construct more accurate predictive models and stronger classifiers.

## REFERENCES

[1].American Cancer Society, Cancer facts and figures 2006. http://www.cancer.org/downloads/STT/CAFF2006PWSecured.pdf. Accessed 24 Jul 2007. [3].Borovkova, S., Analysis of survival data. http://www.math.leidenuniv.nl/~naw/serie5/03/dec2002/pdf/borovkova.pdf. Accessed. [4].Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984): Classification and regression trees. Wadsworth, Belmont.

[5]Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. and Zanasi, A. (1998): Discovering data mining from concept to implementation. Upper Saddle River, N.J., Prentice Hall...