An Improved Algorithm To Predict Recurrence Of Breast Cancer

Umang Agrawal[1], Ass. Prof. Ishan K Rajani[2]

[1] *M.E Computer Engineer, Silver Oak College of Engineering & Technology, Gujarat, India.*

[2] *Assistant Professor, Computer/IT Department, Silver Oak College of Engineering & Technology, Gujarat, India.*

ABSTRACT

In today's fast growing world people are becoming more and more prone to diseases, whether they live in developed or third world countries. The tremendous advancement in technology has led medical information systems in hospitals and medical institutions become larger and larger and in turn the process of extracting useful information is becoming more and more difficult and time consuming. Breast cancer is the one of the most common cancer in women and thus the early stage detection in breast cancer can provide potential advantage in the treatment of this disease. Early treatment not only helps to cure cancer but also help in its prevention of its recurrence. Datamining algorithm can provide great assistance in prediction of early stage breast cancer that always has been a challenging research problem. The main objective of this research is to find how precisely can these datamining algorithms predict the probability of recurrence of the disease among the patients on the basis of important stated parameters. An approach is proposed for disease prediction that combines Support vector machine and bagging using Ensemble learning. The research highlights the performance of Support vector machine with using bagging method in ensemble learning. Experiments show that Support vector machine is the best predictor in all classification algorithms while combining with bagging .The result indicates that accuracy of Support Vector Machine is 81% without using ensemble learning but by combining with bagging in ensemble learning, the accuracy of Support Vector machine is 84.61% which is the highest accuracy that have predicted.

Keyword: - *Datamining, SVM, Bagging, Breast cancer, Classification, Ensemble and machine learning.*

1. INTRODUCTION

Breast cancer is the most common cancer in the world among women according the World health organization's Globocan2012 report [1]. As per the report, Indian women are the most affected by this disease and therefore, it is the most common cause of death too. Early detection of this cancer increased the survivability chances of patients suffering from this disease. Many biological techniques can be used for early detection of breast cancer so that preventive measures can be taken. In this paper, we use different data mining algorithms to predict all those cases of breast cancer that are recurrent using Wisconsin Prognostic Breast Cancer (WPBC) dataset from the UCI machine learning repository [2].

Data mining and knowledge discovery from data (KDD) is the process of extracting knowledge from large amounts of data and have been successfully applied to different classification tasks including, but not limited to, decision making, fault detection, pattern recognition, weather forecasting and image processing. Extracting knowledge from data aims at building a model from the data to predict the future behavior.

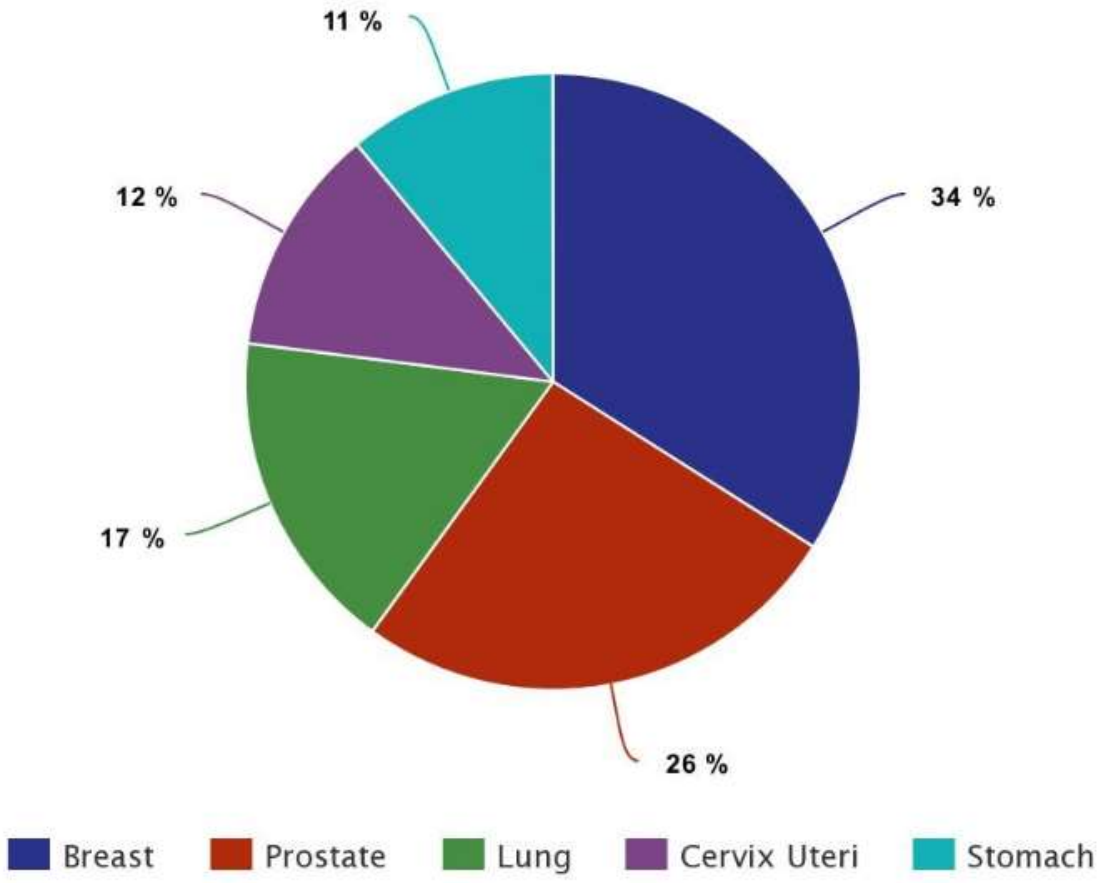

Fig -1: Most Common Cancer Types Prevalent in Humans

2. BACKGROUND

2.1 Overview of related work

The past and current research reports on medical data using data mining techniques have been studied. All these reports are taken as a base for this paper. Ojha U, Goel S. [3] have compared various classification and clustering algorithms on Wisconsin Breast Cancer Prognosis dataset. Their result demonstrate that Support vector machine and C5.0 produce 81% accuracy. Jacob et al. [4] have compared various classifier algorithms on Wisconsin Breast Cancer diagnosis dataset. Their results demonstrate that Random Tree andC4.5 classification algorithm produce 100% accuracy. However they have used, 'Time' attribute (Time to recur/ Disease-free Survival) along with other parameters to predict the outcome of recurrence or non-recurrence of breast cancer among patients. In this paper, 'Time' attribute has not been relied upon for prediction of recurrence of the disease. Delen et al. [5] used the SEER data (period of 1973-2000 with 202,932 records) of breast cancer to predict the survivability of a patient using 10fold cross validation method. The result indicated that the decision tree (C5) is the best predictor with 93.6% accuracy on the dataset, artificial neural network (ANN) also showed good performance with 91.2% accuracy The logistic regression model was less successful with 89.2% accuracy as compared to other two. Chih-Lin Chi et al. [6] used the ANN model for Breast Cancer Prognosis on two dataset. They predicted recurrence probability of breast cancer and grouped patients with good (>5years) and bad(<5 years) prognoses. Falk et al. [7] has explored the results of Gaussian Mixture Regression (GMR) on WPBC dataset and has concluded that the GMR performance is better than the performance of Classification and Regression Trees (CART) in predicting breast cancer recurrence in patients. Pendharkar et al.[8] used several data mining algorithms for discovering patterns in breast cancer. They

showed that data mining could be used in discovering similar patterns in breast cancer cases, which could be a great help in early detection and prevention of this disease.

2.2 Support Vector Machine (SVM)

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well (look at following below snapshot).

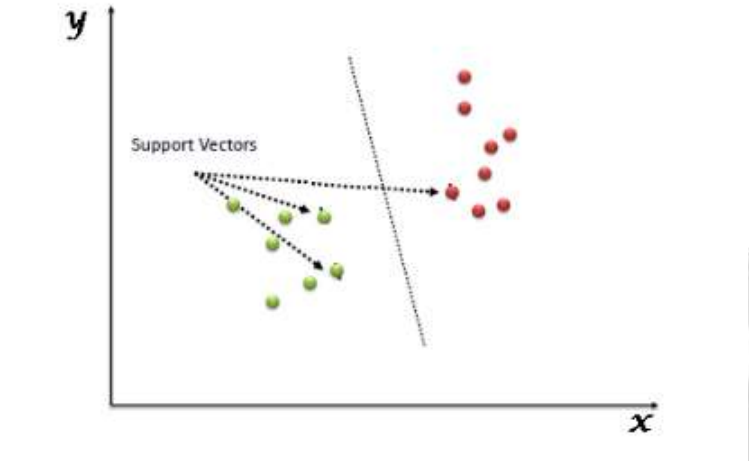

Fig -2: Support Vector Machine

SVM works really well with clear margin of separation and it is effective in high dimensional spaces. Also it uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

2.3 Bagging

Bootstrap aggregating also known as bagging, is one of the most widely used Machine Learning algorithm which is basically used for improving accuracy and stability of algorithms and is deployed in statistical classification and regression. It works by reducing variance for avoiding Overfitting. Mostly deployed with decision tree methods but can be applied with any machine learning techniques. It is one of the special cases of model averaging approach. Suppose there is Dataset (training dataset) 'D' of size 'n' , bagging approach creates 'm' new training datasets each of size 'n', suppose newly created datasets are designated as 'D' and generated after uniform sampling from D with replacement. When we generate new training examples by sampling with replacement some observations may be repeated in Di. In case n' = n, for large value of 'n' D is equivalent to have fraction (1-1/e) (≈ 64%) that is only 64% contains unique examples rest being duplicates. This type of sampling is known as Bootstrap Sampling. Total 'm' bootstrap samples are used to fit 'm' models and combined by averaging the output (for regression) or voting (for classification) [9].

3. EXPERIMENTS

3.1 Data Source

In order to find the best predictor model that can predict recurrent cases of breast cancer, the authentic dataset has been used. In WPBC dataset, Out of 35 attributes, the 'Outcome' is the target attribute (class label); and, all other 32attributes (except ID) are decisive attributes whose value helps in predicting the recurrence of the disease. This data set consists of 198 records of patients out of which, the value of the attribute 'Lymph node' status was missing in 4 records. Since lymph node value is an important factor in determining the breast cancer status. Thus the records rather than removing this attribute itself. Thus the final dataset contains 194 records in which 148 were non recurrent and 46 were recurrent cases.

3.2 Evaluation Methodology

In classifying an unknown case, depending on the class predicted by the classifier and the true class of the patient (Control or HCC), four possible types of results can be observed for the prediction as follows:

- True positive—the result of the patient has been predicted as positive (disease type) and the patient has particular disease.
- False positive—the result of the patient has been predicted as positive (disease type) but the patient does not have disease.
- True negative—the result of the patient has been predicted as negative (Control), and indeed, the patient does not have particular disease.
- False negative—the result of the patient has been predicted as negative (Control) but the patient has disease.

Let TP, FP, TN, and FN, respectively, denote the number of true positives, false positives, true negatives, and false negatives. For each learning and evaluation experiment Accuracy defined as below is used as the fitness or performance indicators of the classification [10],[11].

Accuracy = (TP + TN) / (TP + TN + FP + FN).

In this research work, Applied SVM and Bagging on given data set where dataset split ratio is 80:20.From this study it has been found that our proposed ensemble approach generates better result in terms of accuracy. It creates individuals for its ensemble by training each SVM classifier on a random subset of the training set. For a given data set, k random bootstrap samples are drawn with replacement. SVM classifiers are trained independently on each randomly selected subsets and aggregated via an aggregation technique. A test set is predicted on each of the SVM classifiers and the predicted class labels are determined using aggregated results likely in training sets. Normalized and transformed datasets are used as input to this SVM-Bagged classifier.

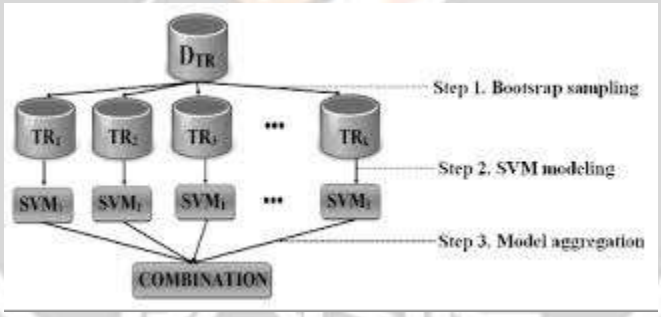

Fig -3: Proposed system

3.2 Experimental Result

```
In [62]: bg = BaggingClassifier(SVC(C=1000),max_samples= 0.2, max_features = 1.0)
         bg.fit(x_train,y_train)
         Accuracy=bg.score(x_test,y_test)
         print('The Accuracy of bagged svm is',Accuracy)

         The Accuracy of bagged svm is 0.8461538461538461
```

Fig -4: Result of Proposed System

4. CONCLUSIONS

Using prediction model to classify recurrent or non-recurrent cases of breast cancer is a research that is statistical in nature. Still this work can be linked to bio medical evidences. In this paper, WPBC dataset is used for finding an efficient predictor algorithm to predict the recurring or non-recurring nature of disease. This might help Oncologists to differentiate a good prognosis (non-recurrent) from a bad one (recurrent) and can treat the patients more effectively. Hence in this research work we focus to improve accuracy by using ensemble learning in which an improved svm algorithm has shown 84.61% accuracy in classifying the recurrence of the disease.

5. REFERENCES

[1]J. Ferlay, Globocan 2012 v1.0 Cancer Incidence and Mortality Worldwide: Iarc Cancer base no. 11, 2014, [online] Available: http://globocan.iarc.fr.

[2]A.Frank and A.Asuncion, "UCI machine learning repository," 2010.[online] Available: http://archive.ics.uci.edu/ml.

[3]Ojha U, Goel S. A study on prediction of breast cancer recurrence using data mining techniques. In Cloud Computing, Data Science & Engineering Confluence, 2017 7th International Conference on 2017 Jan 12 (pp. 527-530). IEEE.

[4]Shomona G. Jacob and R. Geetha Ramani, "Efficient Classifier for Classification of Prognosis Breast Cancer Data Through Data Mining Techniques," Proceedings of the World Congress on Engineering and Computer Science 2012, Vol. I, October 2012.

[5]D. Delen, G. Walker, A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods", Artificial Intelligence in Medicine, vol. 34, no. 2, pp. 113127, 2004.

[6]C.L. Chi, W. N. Street, W. H. Wolberg, "Application of artificial neural network-based survival analysis on two breast cancer datasets", American Medical Informatics Association Annual Symposium, pp. 130-134, Nov. 2007.

[7]T. H. Falk, H. Shatkay, and W.-Y. Chan, "Breast cancer prognosis via gaussian mixture regression," in Canadian conference on Electrical and Computer Engineering, CCECE'06, 2006.

[8]Breiman, “Bagging predictors,” Machine Learning, Vol. 24, 1996, . pp. 123-140.

[9]Efron B., and Tibshirani R., “An Introduction to the Bootstrap,” Chapman & Hall, 1993.

[10]Vapnik V., Statistical Learning theory. New York: Wiley, 1998.

[11]Yugal K. Sahoo G.,”Study of Parametric Performance Evaluation of Machine Learning and Statistical Classifiers” International Journal of Information Technology & Computer Science , Vol. 5 Issue 6, 2013, pp.57-64.