Machine Learning Algorithm to Predict Survivability In Breast Cancer Patients

Kahksha
Department of Computer Science and Engineering
School of Engineering Sciences and Technology, Jamia Hamdard, New Delhi, India
kahkasha.ahmed@gmail.com

Sameena Naaz
Department of Computer Science and Engineering
School of Engineering Sciences and Technology, Jamia Hamdard, New Delhi, India
snaaz@jamiahamdard.ac.in

Abstract: **Breast cancer is one of the leading cancer and cause of mortality among women in developing countries including India. Data mining is a field of science with the help of which we can recognize the unique patterns and markers from huge heterogeneous database responsible for particular disease. It is helpful to create awareness and to take early decisions based on evidences. In this work we applied different data mining algorithms on UCI Machine Learning Repository dataset which consists of data from case study that was conducted at the University of Chicago's Billings Hospital on the breast cancer patients. In this work models have been developed using decision tree, SVM, and neural network and the accuracy of the developed models were accessed to find the best model for this type of data.**

Keywords: Data Mining, Breast Cancer, SVM, Decision tree, Neural Network.

I. INTRODUCTION

A malignant tumor is a condition when cell division is uncontrolled, cells of an organ divide rapidly to form tumor whereas in normal condition cell genesis and apoptosis are in a controlled manner. In breast cancer patient's breast cell grow abnormally and develop a tumor. Cancer is a complex disorder affecting different organ in our body. Tumours are classified into two categories: malignant and non malignant depending on its affected area. A malignant tumour is the one which affects the surrounding organs as well, whereas a non-malignant tumour is localised in a particular organ and does not spared to other body organs. A malignant tumour is more complicated and difficult to treat. According to WHO reports cancer is a major cause of mortality in the world, accounting for 8.8 million loss of life in 2015. Only 1 in 5 low to and middle-income countries have the requisite data to manage cancer policy [1]. Mortality rate among cancer patients are Lung cancer 1.69 million, Liver cancer 788000, Colorectal cancer 774000, Stomach cancer 754 000 and Breast cancer 571000 deaths in 2015 and this is expected to rise by 70% in next two decades worldwide [2]. The field of data mining has not been particularly much used in medical sciences, interaction of patients, physician, and pharmacy are not much evolved to record large data or to store information. But now in this era of information technology, every professional is equipped with technology, with the help of which it is easy to record and store information which would be helpful in data mining. This trend has started showing up now. Today, healthcare industry delivers a broad measure of complex information with respect to hospitals, patients, electronic patient records, disease prognosis and diagnosis and medical healthcare devices. This huge amount of data needs to be mined and filtered so as to enable us to extract information that can be useful for mankind.

Medical data mining has great application for uncovering the invisible patterns in the data sets of the various disease in medical science. These patterns can be utilized for clinical diagnosis and treatment forecasting. However, the available raw data from hospital information system is heterogeneous in nature as different medical councils have different guidelines for treatment and diagnosis.

II. RELATED WORK

A program by the name of Surveillance, Epidemiology and End Results (SEER) is run by National Cancer Institute (NCI) in USA which records cancer incidences in United States. In [3] a study has been conducted on SEER breast cancer database of 700 records in which authors used C4.5 and Naïve Bayes, algorithm to classify data into benign and malignant cancer. The results of this work show that C4.5 algorithm has more accuracy (98.09%) as compared to Naïve Bayes (95.85%).

The study in [4] condenses technical and fundamental aspects of data mining on breast cancer. The author emphasizes on improvement in the breast cancer detection and prognosis with the help of machine learning. In this work decision tree was created using the Weka J48 and C4.5. With these techniques, the author concluded that it is feasible to find statistically significant associations from a breast cancer data set.

In [5] authors have developed different prediction models for breast cancer survivability. They used three data mining methods: RepTree, Simple Logistic and RBF Network with 10 fold cross-validation. Authors were able to find that simple logistic classification outperformed with an accuracy of 74.4% and 0.62 seconds were taken to build the model.

J. R. Orlando et al [6], used 32 genes to find patients at risk for breast cancer Authours were able to show that it is feasible to find statistically significant associations in breast cancer patients. They utilized 94 cases and 164 controls with 32 SNPs out of which 6 were from TP53, 7 were from BRCA1 and 19 were from BRCA2 gene. They used decision tree to find highly susceptible group for breast cancer.

Abdelghani Bellaachia [7] has performed an experiment on SEER database for prediction of survivability rate of breast cancer patients. The database consisted of 151,886 records, from 16 fields, they applied three data mining techniques: Naïve Bayes, back-propagated neural network and C4.5 decision tree. After conducting several data mining experiments, they found that C4.5 algorithm is more superior than other two techniques. 87% accuracy was achieved by authors in this work.

Delen et al [8] has compared three data mining methods for predicting breast cancer survivability. The data mining methods compared in this work are decision tree, artificial neural network and logistic regression. The authors here have used a huge dataset consisting of approximately 2 lakh cases and have used 10 fold cross-validation method to measure the unbiased prediction model. They found that decision tree is best method for prediction with the accuracy of 93.6%, whereas artificial neural network 91.2% and logistic regression method only 89.2% accuracy.

For diagnosis of breast cancer, SEER dataset was classified into benign or precancerous and malignant stage by Rajesh et al using C4.5 algorithm [9].

III. METHODOLOGY

The dataset used in this work consists of 306 multivariate instances with 3 attributes related to age of patients at the time of operation, patients year of operation, and number of positive axillary nodules detected. The target attributes used here is survival status of patients which has been represented by 1 for patients who survived for 5 or more years and 2 for patients who survived for less than five year. Dataset has been taken from the publicly available UCI Machine Learning Repository[10] which was collected from case study conducted on the survival status of breast cancer patients undergone surgery, for a period of 12 years at the University of Chicago's, billings hospital.. R, which is a freely available programing language has been used for analysis of the dataset. In this paper we have examined three different data mining process namely Decision tree, Neural Network and SVM using R. Confusion matrix was created to evaluate the accuracy of the developed model. These models were selected to classify the dataset to find the best model for prediction of survivability among breast cancer patient undergone surgery.

There are millions of neurons in human brain which are connected to each other with synapses and each synapse has its weight. Similarly Neural Network consists of sets of interconnected input/output units and each connected unit has its weight. This Neural Network is used to extract patterns from complicated dataset that are too complex to be noticed by human or other technique[12]. Support Vector Machine is an algorithm used to find linear separator between data point of two class in multidimensional space. It is used to find interaction among features and redundant features. Decision tree is simple to understand in which terminal node represents a decision [13]. We used the above three models to find the most accurate model.

Following steps were involved in the processing of UCI dataset for the development of the above three models.

- Pre-processing
- Preliminary classification
- Selecting Classification technique
- Applying techniques to the test data
- Performance Evaluation

1. Pre-processing

In pre-processing selected UCI breast cancer dataset was converted into a suitable format and dataset was cleaned to utilize in further steps.

2. Preliminary classification

Attributes and class attributes were selected. Class attributes represent target survival status of breast cancer patient undergone surgery.

a. 1 for the patients who survived for 5 or more years.

b. 2 for patients who survived for less than 5 years.

3. Selecting classification technique

Three classification technique Decision tree, Neural Network and SVM were selected in this work for evaluation of model and prediction of survivability.

4. Applying techniques

Experimental analysis has been performed using R i386 3.4.1 software to classify dataset using a decision tree, NN, and SVM techniques. The results of the classification methods are given in the following table:

Table 1: Comparison of Classification Methods

S. No.	*Classification Methods*	*Accuracy*	*Time Taken (Sec.)*
1	Decision Tree	80.80%	0.01
2	SVM	83.64%	0.05
3	NN	77.00%	0.02

From the Table 1, it has been concluded that SVM performs better with the accuracy of 83.60% as compared to Decision tree 80.80% and NN only 77%.

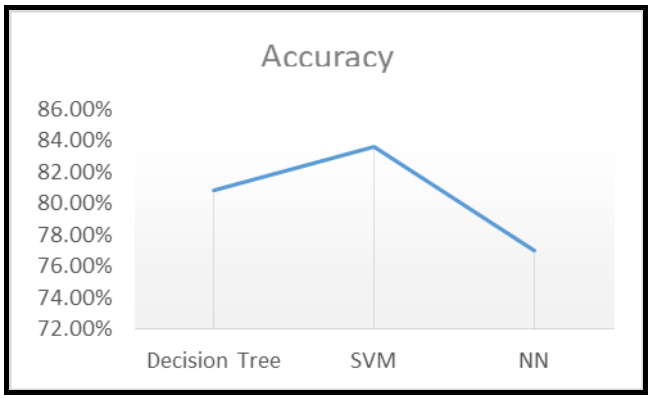

Figure 1. Percentage accuracy of Decision tree, SVM, and NN

IV. RESULTS AND DISCUSSION

Performance has been evaluated using Confusion Matrix. In this work, performance Matrix of SVM has been calculated. This Matrix can be obtained from confusion Matrix and can be converted into true Positive (TP) and false positive (FP), confusion Matrix is also called error matrix.

Table 2: Confusion Metrix

	Predict		
Actual	1	2	Total
1	163 (TP)	2 (FN)	165
2	33 (FP)	16 (TN)	49
Total	196	18	214

In this research work Accuracy, sensitivity, specificity, and precision have been used as performance measures

Accuracy is proportions of correct classification (true positive and negative) from overall numbers of case

Accuracy = (TP+TN) / (TP+TN+FP+FN) (1)

Accuracy = (163+16) / (163+16+33+2)

= 0.8364

Sensitivity is proportions of correct positive classifications (true positive) from case that are actually positive

Sensitivity = TP/ (FN+TP) (2)

Sensitivity = 163/ (2+163) = 0.987

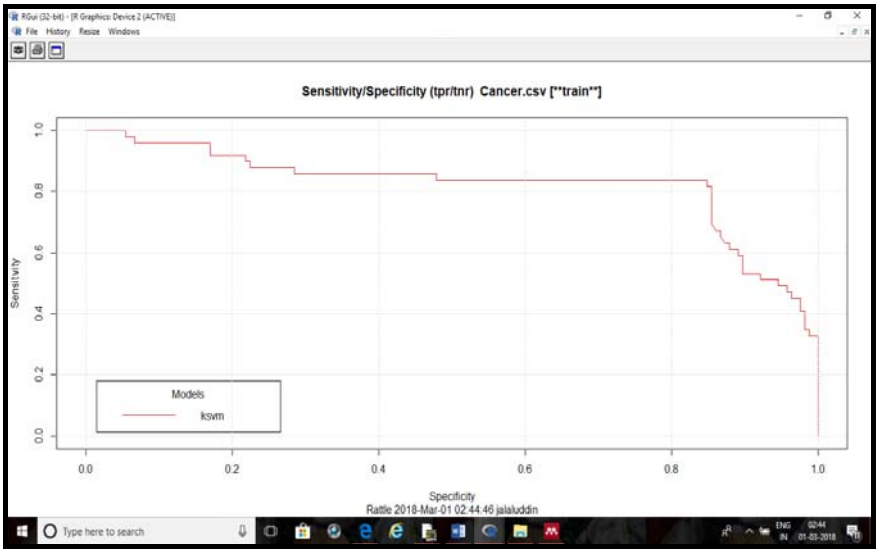

Figure 1. Sensitivity Vs specificity

Specificity is the proportions of positive records classified correctly out of all positive records

Specificity = TN/ (FP+TN) (3)

Specificity = 16/ 33+16 = 0.326

Precision is the proportions of the correct positive classification (true positive) from cases that predicate as positive

Precision = TP/ (TP+FP) (4)

Precision = 163/ (163+33) = 0.83

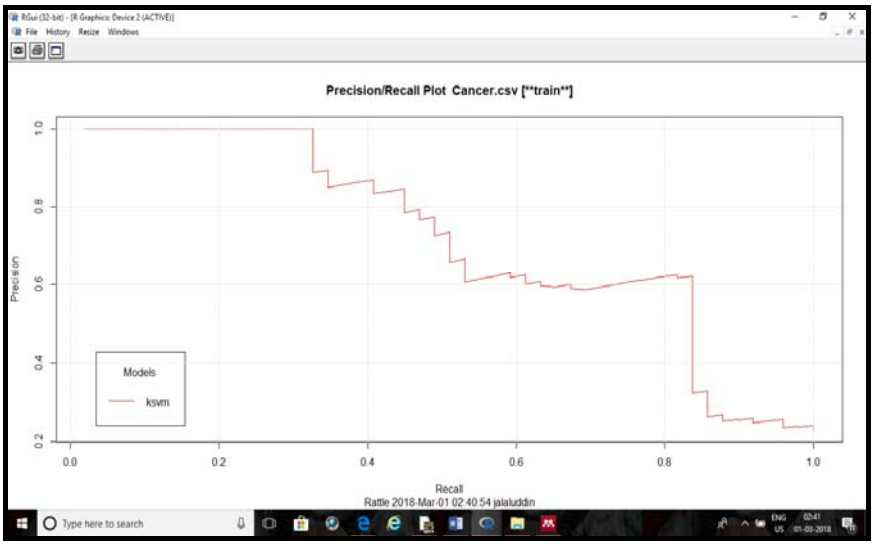

Figure 2: Precession Vs sensitivity (Recall)

Survival status of patients

In this study, 73.52 % patients survived more than five years (Survival Status 1) whereas 26.47% of the patients were survived less than five years (survival status 2) as shown in “Fig. 3”

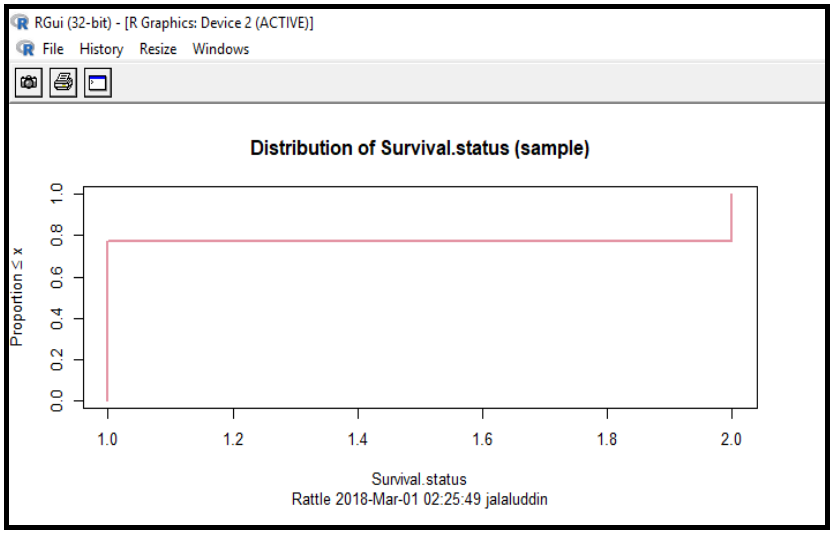

Figure 3. Distribution of survival status

V. CONCLUSION AND FUTURE WORK

In this research work we tried to classify the UCI data to find survival status of the patients undergone surgery due to breast cancer into survival status 1(patients survived more than 5 years) and 2 (patients survived less than 5 years), we have been developed Decision tree, NN, and SVM model over the dataset, finding of the developed model was that SVM has the highest accuracy 83.64% as compared to other two algorithm. Further enhancement of this work would be to apply different data mining methods on large dataset with more number of attributes along with missing values.

VI. REFERENCES

[1] WHO: International agency for research on cancer, “International cancer community welcomes Global Initiative for Cancer Registry Development in low- and middle-income countries,” IARC News, no. 10, pp. 1–2, 2011.
[2] M. H. Forouzanfar et al., “Global, regional, and national comparative risk assessment of 79 behavioral, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the global burden of disease Study 2015,” Lancet, vol. 388, no. 10053, pp. 1659–1724, 2016.
[3] K. Yeulkar, "R Analysis of SEER Breast Cancer Dataset Using Naive Bayes and C4. 5 Algorithm," IJST vol. 8491, pp. 43–45, 2017.
[4] S. Kharya, “Using data mining techniques for diagnosis and prognosis of cancer disease,” Int. J. Comput. Sci. Inf. Technol., vol. 2, no. 2, pp. 55–66, 2012.
[5] V. Chaurasia and S. Pal, “Data mining techniques: To predict and resolve breast cancer survivability.” 2017. [Online] Available:https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2994925
[6] J. R. Orlando Anunciação, Bruno C. Gomes, Susana Vinga, Jorge Gaspar, Arlindo L. Oliveira, “A data mining approach for the detection of high-risk breast cancer groups.” Advances in bioinformatics 4th International workshop on practical applications of computational biology and bioinformatics 2010 (IWPACBB 2010), Springer-Verlag Berlin Heidelberg, pp. 43–51, 2010.
[7] Bellaachia, Abdelghani, Erhan Guven," Predicting breast cancer survivability using data mining techniques", Vol. 58, Issue 13, pp. 10-110, 2006.
[8] D. Delen, G. Walker, and A. Kadam, “Predicting breast cancer survivability: A comparison of three data mining methods,” Artif. Intell. Med., vol. 34, no. 2, pp. 113–127, 2005.
[9] Rajesh K., Sheila Anand,“Analysis of SEER Dataset for Breast Cancer Diagnosis using C4.5 Classification Algorithm”, Int. Journal of Advanced Research in Computer and Communication Engineering, Vol. 1, Issue 2, pp. 72-77, 2012
[10] Haberman, S. J. (1976). Generalized Residuals for Log-Linear Models, Proceedings of the 9th International Biometrics Conference, Boston, pp. 104-122.
[11] [Online] Available: https://en.wikipedia.org/wiki/R_ (programming language)
[12] V. A. Kanimozhi and T. Karthikeyan, “A Survey on Machine Learning Algorithms in Data Mining for Prediction of Heart Disease,” Int. J. Adv. Res. Comput. Commun. Eng., vol. 5, no. 4, pp. 552–557, 2016.
[13] S. Gupta, D. Kumar, and A. Sharma, “Data Mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis,” J. Comput. Sci., vol. 2, no. 2, pp. 188–195, 2011.