Predicting Breast Cancer Recurrence Using Machine Learning Techniques

Umesh D R
Department of Computer Science & Engineering
PESCE, Mandya, Karnataka, India

Dr. B Ramachandra
Department of Electrical and Electronics Engineering
PESCE, Mandya, Karnataka, India

Abstract- Breast cancer is the most common type of cancer in women in the developed countries including India. Breast cancer recurrence could recur anytime in the breast cancer survivors, but mostly it comes back in the first three to five years after the treatment. In this paper we investigate three Decision Tree (DT), Support Vector Machine (SVM), and Artificial Neural Network (ANN) machine learning Algorithms to predict whether or not breast cancer will recur for the breast cancer patient based on SEER (Surveillance, Epidemiology, and End Results) dataset.

Keywords – Breast cancer, Decision Tree, Support Vector Machine, Artificial Neural Network, SEER Dataset.

I. INTRODUCTION

Breast cancer is a malignant tumor that develops when cells in the breast tissue divide and grow without the normal controls on cell death and cell division [1, 2]. It is the most common cancer among women [3]. Early diagnosis and treatment helps to prevent the spread of cancer. Breast cancer begins in the cells of the lobules or the ducts [4]. 5-10% of cancers are due to an abnormality which is inherited from the parents and about 90% of breast cancers are due to genetic abnormalities that happen as a result of the aging process [5].

Treatments for breast cancer are classified into two main types, local and systematic. Surgery and radiation are examples of local treatments whereas chemotherapy and hormone therapy are examples of systematic therapies. Usually, these two types of treatment are used together for the best results [1]. Breast cancer is one among the leading cause of cancer death in women, the survival rate is high and with early diagnosis i.e. 97% of women survive for 5 years or more [2, 6].

Breast cancer can recur at any time or not at all, but most recurrences happen in the first 5 years after breast cancer treatment. Breast cancer can come back as a local recurrence/regional recurrence and the distant metastasis. Some of the most common sites of recurrence outside the breast are the lymph nodes, bones, liver, lungs, and brain. Therefore it is a most important research to predict the breast cancer recurrence.

Data mining techniques have been extensively used for breast cancer diagnosis. Diagnosis is used to predict the presence of cancer and differentiate between the existence of malignant and benign tumor. In our study we investigate three DT, SVM, and ANN machine learning Algorithms to predict whether or not breast cancer will recur for the breast cancer patient based on SEER dataset.

II. PROPOSED ALGORITHM

In this paper we investigate three DT, SVM, and ANN machine learning Algorithms to predict whether or not breast cancer will recur for the breast cancer patient based on SEER (Surveillance, Epidemiology, and End Results) dataset of Program of the National Cancer Institute (NCI). This dataset contained population characteristics and included 26 input variables. Our cases were collected from the random sample of SEER breast cancer dataset. We preprocessed the data to remove unsuitable cases. After using data cleansing and data preparation strategies, the final dataset was constructed. Finally, 547 cases were analyzed for breast cancer recurrences happen in the first 5 years after breast cancer treatment. The independent variables that we used are shown in Table 1. The dataset was cleaned by handling missing values, noise, identifying and correcting inconsistencies. Some fields, such as Her2,

age of menarche, and Npositive, contained missing values. Since these variables are important in predicting recurrence, the records containing the missing data were substituted using Expectation maximization (EM) method [7].

Sl. No.	Variable Name	Definition
1	Local Recurrence	Yes or No
2	Age at Diagnosis	≤ 35, 35 to 44, 44-45, 55 ≥ years old
3	Age at Menarche	≤ 12 to ≥ 12 years old
4	Age at Menopause	≤ 50 to ≥ 50 years old
5	Family History of Breast Cancer	Yes or No
6	History of other Cancer (CA)	Yes or No
7	Location	Upper outer Quadrant (UOQ), Upper inner Quadrant (UIQ), Lower outer Quadrant (LOQ), Lower inner Quadrant (LIQ), Central, Axilla, Upper half, Lateral half, Lower half
8	Side	Left, Right, Bilatera
9	CS Tumor Size	≤ 2cm to ≥ 5cm
10	CS LN/Nexion	Lymph node involvement/number of removed nodes after surgery
11	Metastasis	Bone, Liver, Lung, Brain, others
12	NPositive	Number of Positive lymph node involvement
13	B.Pathology	Results of Biopsy Pathology after
14	Type of surgery	Mastectomy (Preservative or Bilateral)
15	G (Grade)	1, 2 or 3
16	Margin of Involvement	Free or ≥ 2cm
17	Estrogen Receptor	Negative or Positive
18	Progesterone Receptor	Negative or Positive
19	Type of Chemotherapy	Adjuvant or Neoadjuvant
20	Radiotherapy	Yes or No
21	Hormone Therapy	Tamoxifen, Raloxifen, Femara, Aromasin or Megace
22	Death	Related to Breast Cancer or unrelated
23	Her2	Negative or Positive
24	S-phase fraction	Percentage of cells in Samples
25	DNA Index	Higher in more malignant tumors
26	CS Extension	Identifies contiguous growth (extension) of the primary tumor within the organ of origin or its direct extension into neighboring organs.

Table 1: Variables used for Breast Cancer recurrence modeling.

For DT, C5.0 are sophisticated data mining tools for discovering patterns that delineate categories, assembling them into classifiers, and using them to make predictions. Each tree node is either a leaf node or decision node. All decision nodes have splits, testing the values of some functions of data attributes. Each branch from the decision node corresponds to a different outcome of the test. Each leaf node has a class label attached to it. Weka software was implemented to analyze the data with C5.0. It is an open source data mining tool and offers many data mining algorithms including AdaBoost, Bagging, C5.0 and SVM. It is a collection of tools for data classification, regression, clustering, association rules, and visualization [9].

Support vector machine (SVM) is an emerging powerful machine learning technique to classify cases. SVM has been used in a range of problems and they have already been successful in pattern recognition in bioinformatics, cancer diagnosis [10], and more. Figure 1 shows SVM topology in hyperspace.

SVM is a maximum margin classification algorithm rooted in statistical learning theory. It is the method for classifying both linear and non-linear data. It uses a non-linear mapping technique to transform the original training data into a higher dimension. It performs classification tasks by maximizing the margin separating both classes while minimizing the classification errors [11].

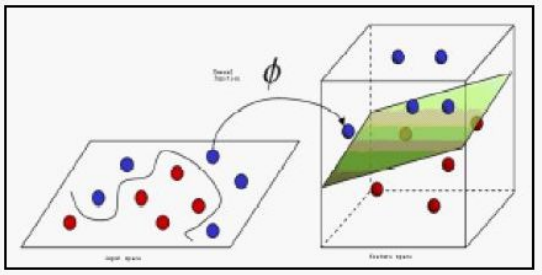

Figure 1: SVM topology

Neural networks are a large number of interconnected nodes that perform summation and thresholding in loose analogy with the neurons of the brain. The multi-layer perceptron (MLP) model is capable of mapping set of input data into a set of appropriate output data. The primary task of neurons in input layer is the division of input signal Xi among neurons in hidden layer. The output of neurons in the output layer is determined in an identical fashion [12]. Figure 2 shows MLP feed forward Neural Network.

The back-propagation algorithm can be employed effectively to train neural networks; it is widely recognized for applications to layered feed-forward networks, or multi-layer perceptrons. MLP is the most commonly used algorithm and performs better than other ANN architectures for this type of classification problems [13].

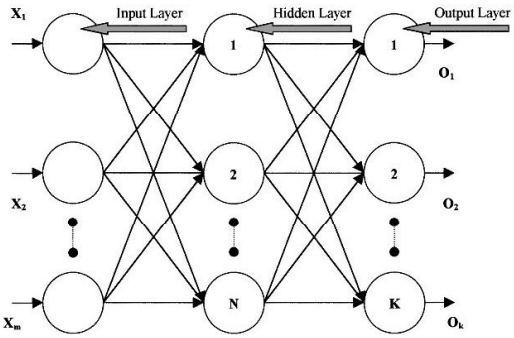

Figure 2: MLP feed forward Neural Network

We have used Weka software tool to experiment with these algorithms. The algorithms can either be applied directly to a dataset or called from a Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. It is open source software issued under the GNU General Public License.

We have adopted an Expectation maximization (EM) algorithm for efficient estimation from incomplete data. In any incomplete dataset, there is indirect evidence about the likely values of the unobserved values. This evidence, when combined with some assumptions, comprises a predictive probability distribution for the missing values that should be averaged in the statistical analysis. The EM algorithm is a common technique for matching models to incomplete data. EM is important on the relationship between missing data and unknown parameters of a model. When the parameters are known, then it is possible to obtain impartial predictions for the missing values [7, 8].

In order to evaluate the validity of present results for making predictions regarding new data, 10-fold cross-validation was implemented in model building, evaluation, and comparison (We performed experiments using Weka, an open source data mining tool and the comparison is based on 10-fold cross-validation). In this method, each of the 10 subsets acts as an independent holdout test set for the model trained with the rest of the subsets. A pair of testing and training sets is called a ‘’fold’’.

III. Experiment and Result

To compare the models, the data from the NCI for Breast Cancer dataset were analyzed. We selected three random sets of 547 records from the SEER breast cancer dataset. This was used as the training dataset to predict the

recurrence of a breast cancer. The classification error rate obtained for the three sets of samples is given in Table 2. It can be noted from the Table 2 that the lowest error rate of 0.0724 was obtained for random Sample 1.

Sample	Sample 1	Sample 2	Sample 3
Error Rate	0.0724	0.0945	0.0847

Table 2: C5.0 Training Phase Error Rates

As a further verification process, we applied the other classification algorithms to Sample 1 set of 547 records. The graphical representation of the classification results obtained is given in Figure 3.

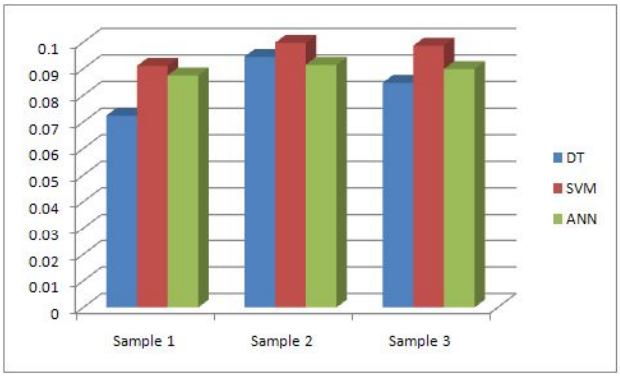

Figure 3: Comparison of error rates obtained for various classification techniques for Sample 1.

The results obtained are tabulated in Table 3. It is seen from the table and the graph that best results are obtained for C5.0 algorithm. This justifies and validates our choosing C5.0 for classification of SEER data set.

Sl. No.	Classification Technique	Error rate for Sample 1
1.	DT	0.0724
2.	SVM	0.0911
3.	ANN	0.0875

Table 3: Comparison of Classification Technique

In testing phase we applied the classification techniques for the random Sample 1 to the complete 547 records. The actual and predicted values obtained in the classification exercise are shown in the confusion matrix given in Table 4.

	DT		SVM		ANN	
	R	NR	R	NR	R	NR
Recurrence (R)	102	15	92	25	94	23
Non-Recurrence (NR)	17	413	19	411	27	403
Total = 547						

Table 4: Confusion Matrix of SEER Dataset

This paper has explored risk factors for predicting breast cancer by using data mining techniques. Each method has its own limitations and strengths specific to the type of application. Table 5 shows the accuracy, sensitivity, and specificity comparison of the data mining methods. Our results show that DT C5.0 outperforms both SVM and ANN in all the parameters of sensitivity, specificity and accuracy. DT is the best predictor of breast cancer recurrence. Table 5 shows the accuracy, sensitivity, and specificity comparison of decision tree C5.0, SVM and ANNs.

Algorithm	Accuracy	Sensitivity	Specificity
DT	**94.15 %**	**87.17%**	**96.04%**
SVM	91.95%	78.63%	95.58%
ANN	90.86%	80.34%	93.72%

Table 5: Comparison of data mining models.

There are some limitations in the current study. There were many cases lost in the follow-up and there were records with missing values that were omitted unfortunately. However, these obtained results were based on a SEER Dataset of Program of the National Cancer Institute (NCI) by comparison three different data mining methodology and also Weka toolkit.

IV.CONCLUSION

This study clearly shows that data mining techniques is a good method to predict breast cancer recurrence. We present an efficient pre-classification method and discover the information of breast cancer recurrence of SEER dataset.

Further studies should be conducted to improve performance of these classification techniques by using more variables and choosing for a longer follow-up duration. Besides, we could also try recurrence prediction of other certain cancer data where the recurrence is seriously high.

REFERENCES

[1] Breast cancer Q&A/facts and statistics (http://www.komen. org/bci/bhealth/QA/q_and_a.asp).
[2] Jerez-Aragone´s JM, Gomez-Ruiz JA, Ramos-Jimenez G, Munoz-Perez J, Alba-Conejo E. A combined neural network and decision trees model for prognosis of breast cancer relapse. Artificial Intelligence Medicine 2003; 27:45—63.
[3] Calle J. Breast cancer facts and figures 2003—2004. American Cancer Society 2004. p. 1—27.
[4] http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0001911.
[5] Breast Cancer statistics from Centers for Disease Control and Prevention, http://www.cdc.gov/cancer/breast/statistics/.
[6] O'Malley CD, Le GM, Glaser SL, Shema SJ, West DW. Socioeconomic status and breast carcinoma survival in four racial/ethnic groups: a population-based study. American Cancer Society 2003; 1303—11.
[7] Dempster AP, Laird NM, Rubin DB (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. J R Stat Soc Series B 39: 1-38.
[8] Rubin DB, Schenker N (1991) Multiple Imputation in Health-Care Databases - an overview and some applications. Stat Med 10: 585-598.
[9] Weka 3: Data Mining Software in Java.
[10] Cristianini N, Shawe-taylor J (2000) An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, London: Cambridge University Press.
[11] Joachims T (1998) Making large-scale support vector machine learning practical. Advances in Kernel Methods: Support Vector Learning. MIT Press, Cambridge, MA, 169-184.
[12] Haykin S (1998) Neural networks: a comprehensive foundation. New Jersey: Prentice Hall, New Jersey, USA.
[13] Hornik K, Stinchcombe M, White H (1990) Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. Neural Networks 3: 359-366.