Survey of Data Mining Techniques for Prediction of Breast Cancer Recurrence

Desta Mulatu.
CS Dept., *Symbiosis Institute of Technology*
Symbiosis International University

Rupali R. Gangarde.
CS Dept., *Symbiosis Institute of Technology*
Symbiosis International University

Abstract - Breast cancer is the major known disease in this world. This disease is not only causing women, sometimes male has also the probability to be caused by this disease. Recurrence is when breast cancer comes back. Breast cancer can recur at any time or not at all, but sometimes this disease can be re available in human body after five years of its treatment. Breast cancer can reoccur as a recurrence or occur outside of the breast. Those occurring outside of the breast may occur in the lymph nodes, liver, bones, brain and lungs. Early stage treatment not only helps to cure breast cancer but also help in preventing its recurrence. Data mining algorithm can provide great assistance in the prediction of early-stage breast cancer that always has been challenging research problem. The proposed research will identify the best algorithm that is used to predict the recurrence of the breast cancer and improve the accuracy the algorithms.

***Keywords*- recurrence, prediction, classification, breast, non-recurrence.**

I. INTRODUCTION

Data mining is become playing a great role in computing applications in the domain area of medicine. Availability of data mining applications and its techniques are shown in the areas of healthcare administrations, patient care, management, and intensive care system. Breast cancer is the second most common and leading cause of death among all women available in this world. According to published statics breast cancer is affecting both developed and developing countries. Still today there is no more effective ways to prevent the breast cancer because its cause is unknown. Even if there is no direct cause of these diseases early stage treatment of this disease can give full recovery from this disease. Currently, data mining is used for retrieving efficient document from larger documents. There are several data mining applications and techniques which are used to analyze huge data, those data mining techniques are Clustering, Classification, Association Rules, Prediction and Neural Networks Decisions Trees. Among these, some classification algorithms and much known algorithm such as naïve Bayes, support vector machine, artificial neural network, decision tree (c5.0) and k nearest neighbor algorithm are giving most accurate result in a lot of research paper.

Data mining in health care

The term data mining has a different meaning when different peoples are describing it. But the actual and basic definition is analyzing a large data in order to predict the future events. Currently data mining is playing great role in health industry to make healthy industry more efficient than before.

Data mining for breast cancer recurrence

Data mining is currently solving a lot of real world problems. Because the main use of data mining technique is to change raw data into more meaningful information. Both male and female have probability to be caused by breast cancer. But from the world breast cancer statics the occurrence of this disease is higher in female than males. Both patients and doctors should have to take symptoms of this disease seriously. Recurrence of breast cancer is when the cancer is come back after treated. There are three types of breast cancer recurrences, these are local recurrence, distance recurrence and regional recurrence [14].

II. LITERATURE REVIEW

Wisconsin University breast cancer database was analyzed by naïve Bayes prediction algorithm and naïve Bayes classification algorithms. So that algorithms are used to predict and classify whether the tumor is malignant or benign[1]. So data sets were chosen randomly. At the final naïve Bayes classification algorithm was shown that 10-15 percent is wrongly classified and 85-95 percent is correctly classified.

Two various data sets from Wisconsin breast cancer have been evaluated by different data mining algorithms. The outcome that Rotation Forest model shows the highest classification accuracy (99.48 %) and when compared with the previous works, the new approach and methodology have come with highest performance and accuracy[2].

Jimin Guo1, Benjamin C. M. Fung, Farkhund Iqbal implemented decision tree algorithm with breast cancer data sets that get from Leiden University Medical Center [3].The data sets have 574 patients who have got surgery at that hospital. So they generate the recurrence of breast cancer by a decision tree algorithm within three years of initial diagnosis. The classifier predicted 70% accuracy. For the independent classifier of 65 patients the classifier exactly predicts the recurrence of the disease in 55 patients. The classifier also separates patient into two based on their disease characteristic and their relevance of early relapse

Research paper done by Ahmed Iqbal Pritom,shahed anzarus sababa, Md. Ahadur Rahman Munshi, Shihabuzzaman Shihab predicts whether the breast cancer is recurrent or not[4]. They have used data sets from Wisconsin data sets of the UCI machine learning repository that have 35 attributes. After implementation of algorithms like C4.5 Decision Tree, Naive Bayes and Support Vector Machine (SVM) classification algorithm was implemented. The outcome of these SVM, Naïve Bayes and C4.5 has 75.75 %, (67.17 %) and (73.73 %) respectively.

Uma Ojha and Dr. Savita Goel were also discussed about the study on the prediction of breast cancer recurrence using data mining techniques. The research was applied by both clustering and classification algorithms. The results show that decision tree and Support vector machine (SVM) came out with the best predictor 80% accuracy.

Bojana R. Andjelkovic Cirkovic, Aleksandar M. Cvetkovic, Srdjan M. Ninkovic, and Nenad D. Filipovi presents the application of data mining on estimation of survival rate and disease relapse for breast cancer patients. A data set that was taken from the Clinical Center of Kragujevac is evaluated by some classification algorithm. Based on selected data sets naïve Bayes algorithm was selected as an algorithm which have higher accuracy on the basis of the 5 year survival rate.

The research paper done by Joana Diz Goreti Marreiros & Alberto Freitas presents new computer based diagnosis system[5]. By using this technique false positive diagnosis test can be reduced. After data sets analyzed nave bayes algorithm come with higher accuracy than Random forest.

Qi Fan,change-jie zhu and liu yin used different types of data mining techniques in order to predict the recurrence of breast[6]. In this paper they researcher uses SEER data sets and applied a new classification method in order to predict the recurrence of this disease. After preprocessing of data sets the researcher applied several algorithms, so that the decision tree (c5) algorithms come with better performance.

Dursun delen, Glenn walker And Amit Kadam used diferent data mining techniques for prediction of survivability of breast cancer. Data mining, classification algorithms such as artificial neural network and decision tree along with logistic regression to develop a model for breast cancer survivability. Based on this paper decision tree algorithm (c5) was coming with better. performance and predicted by more accuracy 93.6% and artificial neural network shows second performance 91.2% and logistic regression come to the worst of the three 89.2%.

A stage predictive model for breast cancer survivability presented by Rohit j and Ramya Nadig. In this paper they were used different algorithm in order predict the breast cancer solvability. The evaluation was done based the stage of the breast cancer. Three machine learning algorithms were applied in order to predict breast cancer survivability. These data sets was evaluated by classification algorithms such as naïve bayes, logistic regression and decision tree to predict breast cancer survivability.

Table1. Review of Algorithm & Approaches

Title	Methodology	Results
Analysis of Efficiency of Classification and Prediction Algorithms (Naïve Bayes) for Breast Cancer Dataset[1]	Naïve Bayes classification algorithm and naïve Bayes prediction algorithms	The Naïve Bayes classification algorithm has highest accuracy value 89 %-95%
Breast cancer diagnosis using GA feature selection and Rotation Forest[2]	Rotation forest model	Rotation forest shows highest classification accuracy (99.48%)
Revealing determinant factors for early breast cancer recurrence by a decision tree [3]	Decision tree	The classifier predicts for whether a patient developed early disease recurrence; and is estimated to be about 70% accurate
Predicting Breast Cancer Recurrence using effective Classification and Feature Selection technique[4]IEEE 2016	Decision tree c4.5,support vector machine, naïve Bayes algorithms	Support vector machine provide better performance before and after attribute selection.
[6] Predicting breast cancer recurrence using data mining techniques	C5.0, CHAID, QUEST , ANN	After evaluating decision tree(c5.0) algorithm gives 71% more accurate value than others

Prediction model for estimation of survival rate and relapse for breast cancer patient[7]	ANN, SVM ,
	LG, DT, NB
		ANN= 0.9315
		SVM = 0.952
		LR = 0.911
		DT= 0.9657
		NB= 0.856 so that decision tree algorithm has achieved the best performance in a classification task considering following parameters AC =0.9657,SENS =0.991 SPECI =0.889

Predicting breast cancer survivability [8]	Decision tree,(DT) Artificial neural network , Logistic regression	Decision tree =93.6 ANN = 91.2 Logistic regression = 89.2

Using three machine learning techniques for predicting breast cancer recurrence [9]	Decision tree,
	Artificial neural network,
	Support vector machine
		DT =0.936,
		ANN =0.947,
		Svm = 0.957. so that SVM have higher accuracy than other algorithms

Hybrid computer aided diagnosis system for prediction of breast cancer recurrence using optimized ensemble learning [10] Elsevier (2016)	Svm Decision tree multilayer perception (MLP)	Support vector machine give a more accurate value 78%
Comparison of Data Mining Classification Algorithms for Breast Cancer Prediction [11] IEEE 2013	k nearest neighbor algorithm Decision tree, and naïve ayes algorithm	Better performance was provided by naïve bayes algorithm 95.9943% [7]

III Proposed Research

Breast cancer starts to grow in the human body when cells in the breast are growing most in an unexpected manner. After these cells grow, it can be seen by x- ray. Basically, there are two types of breast cancer, cancer that spread into another area and cancer that can't spread into another area. Among the world women breast cancer is the first and the most leading of death of women and the accurate diagnosis have lots of advantage to prevent and detection of the disease. Data mining is a technique can support doctors in the decision making process. As breast cancer recurrence is high, good diagnosis is important. Many studies have been conducted to analyze Breast Cancer Data. This research is going to be implemented by different data mining algorithms like Bayes net, support vector machine and Decision tree (j48). So to get a more accurate value about the recurrence of breast cancer we are going to use data sets which were taken from the UCI machine learning repository and data was kept in. ARFF file and it will open by weak tools

IV Conclusion

As discussed in this survey breast cancer recurrence is the most challenge of researchers for a lot of years. The actual cause of breast cancer is unknown, but early treatment may good way for prevention and also detection of breast cancer. Data mining technology is the most simple and easy ways to predict whether that breast cancer is recurrent or non-recurrent. In a lot of papers listed above, data mining algorithms like decision tree, support vector machine naïve Bayes are giving more accurate results.

References

[1] G. D. Rashmi, A. Lekha, and N. Bawane, "Analysis of efficiency of classification and prediction algorithms (Naïve Bayes) for Breast Cancer dataset," *2015 Int. Conf. Emerg. Res. Electron. Comput. Sci. Technol.*, pp. 108–113, 2015.

[2] E. Aličković and A. Subasi, "Breast cancer diagnosis using GA feature selection and Rotation Forest," *Neural Comput. Appl.*, vol. 28, no. 4, pp. 753–763, 2017.

[3] J. Guo *et al.*, "Revealing determinant factors for early breast cancer recurrence by decision tree," *Inf. Syst. Front.*, 2017.

[4] A. I. Pritom, "Predicting Breast Cancer Recurrence using effective Classification and Feature Selection technique," pp. 310–314, 2016.

[5] J. Diz, G. Marreiros, and A. Freitas, "Applying Data Mining Techniques to Improve Breast Cancer Diagnosis," *J. Med. Syst.*, vol. 40, no. 9, 2016.

[6] Q. Fan, C. Zhu, and L. Yin, "Predicting Breast Cancer Recurrence Using Data Mining Techniques," pp. 310–311, 2010.

[7] B. R. A. Cirkovic, A. M. Cvetkovic, S. M. Ninkovic, and D. Nenad, "Prediction Models for Estimation of Survival Rate and Relapse for Breast Cancer Patients."

[8] R. J. Kate and R. Nadig, "Stage-specific predictive models for breast cancer survivability," *Int. J. Med. Inform.*, vol. 97, pp. 304–311, 2017.

[9] A. LG and E. AT, "Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence," *J. Heal. Med. Informatics*, vol. 4, no. 2, pp. 2–4, 2013.

[10] M. R. Mohebian, H. R. Marateb, M. Mansourian, M. A. Mañanas, and F. Mokarian, "A Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer Recurrence (HPBCR) Using Optimized Ensemble Learning," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 75–85, 2017.

[11] C. Shah and A. G. Jivani, "Comparison of Data Mining Classification Algorithms for Breast Cancer Prediction," *2013 Fourth Int. Conf. Comput. Commun. Netw. Technol.*, pp. 1–4, 2013.

[12] lulu wang. "Early Diagnosis of Breast Cancer", Sensors, 2017

[13] https://www.healthcatalyst.com/data-mining- in-healthcare