A Study On Prediction OfBreast Cancer RecurrenceUsingData Mining Techniques

Uma Ojha
Computer Science Department
ARSD College, Delhi University
Delhi-India
ojha.uma@gmail.com

Dr. Savita Goel
(Retired) Sr. System Programmer, CSC
IIT-Delhi
Delhi, India
savita@iitd.ac.in

***Abstract*—Breast cancer is the most common cancer in women and thus the early stage detection in breast cancer can provide potential advantage in the trea tment of this disease. Early treatment not only helps to cu re cancer but also helps in its prevention of its recurrence. Data mining algorithms can provide great assistance in prediction of earl y stage breast cancer that always has been a ch allenging research problem. The main objective of th is research is to find how precisely can these data mining algorithms predict the probability of recu rrence of th e disease among the patients on the basis of important stated parameters. The research highlights the performance of d ifferent clustering and classification algorithms on the dataset. Experiments show that classification algorithms are better predicto rs than clustering algorithms. The result indicates that the decision tree (C5.0) and SVM is the best predictor with 81% accuracy on the holdout sample and f uzzy c-means came with the lo west accuracy of 37% among the algorithms used in this paper**.

Keywords—Breast cancer predic tion; recurrance; data mining; classification models; clustering models; C5.0; adaptive boosting.

I. INTRODUCTION

Breast cancer is the most common cancer in the world among women according toWorld health organization's Globocan2012 report [1].As per the report, Indian women are most affected by this disease and,therefore, it is the most common cause of death too.Early detection of this cancer increases 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]. Different clustering and classification algorithms of data mining techniques have been used to find the performance of these prediction models. Four clustering algorithms (K means, EM, PAM and Fuzzy c-means) and four classification algorithms (SVM, C5.0, Naive bayes and KNN) are selected for this research. R programming tool is used for the implementation purpose that provides free software environment for data analysis [3].

In short, this research is toidentify the most successful data-mining algorithm that helps to predict those cases of cancer, which can recur. The objective here is also to find critical attributes which play major role in determining and predicting in advance the possibility of recurrence of breast cancer using C5.0 algorithm.

This paper is organized indifferent sections as follows. Section 2 highlights the already published literature in the area of breast cancer survivability prediction models using data mining. Section 3, explains the detail description of data, various prediction algorithms and measures for performance evaluation on the said models.The prediction results of all clustering and classification algorithms along with the accuracy, sensitivity and specificity are presented in section 4.Section 5 concludes with summary of results eventually leading to the future directions.

II. RELATED WORK

The past and current researchreports on medical data using data mining techniques have been studied. All these reports are taken as a base for this paper. 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 10-fold 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 (>5 years) and bad(<5 years)prognoses. Falk et al.[7] has explored

978-1-5090-3519-9/17/$31.00 ©2017 IEEE

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 earlydetection and preventionof this disease.

III. METHODOLOGY

A. 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 missingin 4 records. Since lymph node value is an important factor in determining the breast cancer status.Thus the records containing the missing data of this attribute were removed from the dataset rather than removing this attribute itself. Thus the final dataset contains 194 records in which 148 were non recurrent and 46 were recurrent cases.

B. Prediction Models

Data mining is the process of extracting interesting patterns and knowledge from data. This paper focuses on using some of the clustering and classification modelsto predict the chances of recurrence and survivability of the disease. A short description of these algorithms and their specific implementations for this research are given as follows.

1) Clustering Algorithms

In clustering process, data is partitioned in to sets of clusters or sub-classes. We have used four clustering algorithms namely K-means, EM, PAM, Fuzzy c-means. The K-means clustering algorithm works by partitioning n observations in to k sub-classes defined by centroids, where k is chosen before the algorithm starts. K-meansand EM are both iterative algorithms. EM (Expectation–maximization) is a statistical model that depends on unobserved latent variables to estimate the parameters using maximum-likelihood [9]. The PAM (Partitioning around Medoids)is similar to K-means except that here partitioning is based on K-medoids method that divides the data into a given number of disjoint clusters. In fuzzy clustering, data elements may belong to more than one cluster. This is also referred as soft clustering.

2) Classification Algorithms

Machine learning techniques of classification can be used to classify different objects on the basis of a training setof data whose outcome value is known. In this research, four classification algorithms used are KNN, SVM, Naive Bayes and C5.0. In KNN (K Nearest Neighbor), object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbours. In SVM (Support Vector Machines),data is first converted in to a set of points and then classified in to classes that can be separated linearly.The Naive Bayes model works by estimating the probability of a dataset that can belong to class using Bayes' rule.The C5.0 algorithm is a decision tree that recursively separates observations in branches to construct a tree for the purpose of improving the prediction accuracy. It is an improved version of C4.5 and ID3 algorithms [10]. It also offers the powerful boosting method to increase the accuracy of this classification algorithm [11].

C. Measures for performance evaluation

In this study, we used three performance measures: accuracy, sensitivity and specificity.

Accuracy=	(TP+TN)/ (TP + TN + FP + FN)	(1)
Sensitivity=	TP/ (TP + FN)	(2)
Specificity=	TN/(TN + FP)	(3)

Where TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives respectively.

IV. RESULTS & DISCUSSION

The performance of these algorithms ismeasured based on the accuracy, sensitivity and specificity. Probability of accuracy in results is measured in the range of 0 to 1 whereas 1 means 100% accuracy. The two classification algorithms C5.0 and SVM achieve 81% accuracy, which is better than all algorithms mentioned in this paper.

A. Clustering& Classification Results

The attribute named 'Outcome' classifies whether the disease was recurrent or not and thus it was removed from the dataset so that we find how accurately our data-mining algorithm can predict all such cases. Jacob et al. [5] has also used a 'Time' attribute that specifies time in months in which the disease had recurred or cured completely. Since the aim of this research is to predict by using prediction models whether the disease will recur among patients, therefore,'Time' attribute is ignored to get unbiased results. The complete dataset of WPBC is divided into the ratio of 70:30 in classification algorithms.The 70% of data is used for training purposes and 30% of the dataset is used for testing purposes. The two classifiers C5.0 and SVM were the best predictor algorithms with accuracy of 0.813 whereas fuzzy clustering mean algorithms came worst with the accuracy of 0.3711.Though the accuracy of classification algorithm KNN is

0.7068 which is close to the accuracy of EM clustering algorithm i.e. 0.6804. Thiscame out to be best accurate results of clustering algorithms as shown in fig. 1. Further,Table 1 shows the detailed set of results in the form of confusion matrix.

The average score of all classification and clustering algorithms were calculated to measure the overall performance of these algorithms. The comparison graphs of all algorithms are shown in fig. 2. The result shows that on comparison, classification algorithms are better predictors than clustering algorithms. The classification algorithms were 0.7154 accurate as compared to theaccuracy of 0.5257 of clustering algorithms.

TABLE I. OVERALL RESULT PERFORMANCE OF ALL ALGORITHMS

Clustering Algorithms					Classification Algorithms				
Algorithms	**Confusion Matrix**	**Accuracy**	**Sensitivity**	**Specificity**	**Algorithms**	**Confusion Matrix**	**Accuracy**	**Sensitivity**	**Specificity**
K-Means	N R N 100 48 R 23 23	0.6340	0.8130	0.3239	**C5.0**	N R N 47 0 R 11 0	0.8103	1.0	0.0
EM	N R N 117 31 R 31 15	0.6804	0.7905	0.3260	**KNN**	N R N 47 0 R 11 0	0.7068	0.8297	0.2
PAM	N R N 64 84 R 29 17	0.4175	0.4324	0.1683	**Naïve Bayes**	N R N 47 0 R 11 0	0.5344	0.5319	0.2142
Fuzzy c-means	N R N 50 98 R 24 22	0.3711	0.3378	0.1833	**SVM**	N R N 47 0 R 11 0	0.8103	1.0	0.0
Mean		0.5257	0.5934	0.2504	**Mean**		0.7154	0.8404	0.1036

Confusion matrix shows the classification of the cases in the test dataset. In confusion matrix, the columns denote the actual cases and the rows denote the predicted. Accuracy = (TP + TN)/(TP + FP + TN + FN); sensitivity = TP/(TP + FN); specificity = TN/(TN + FP).

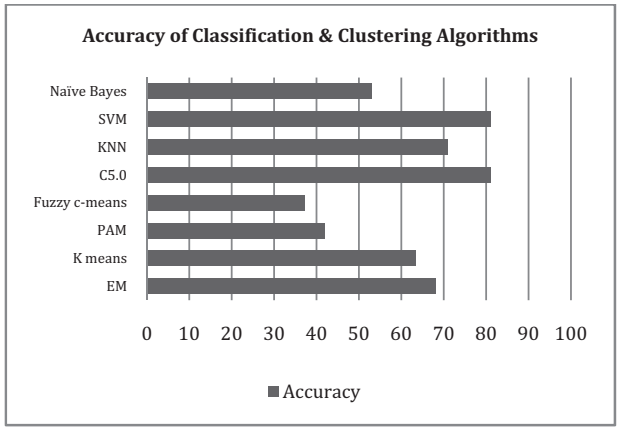

Fig. 1. Accuracy of Classification & Clustering Algorithms

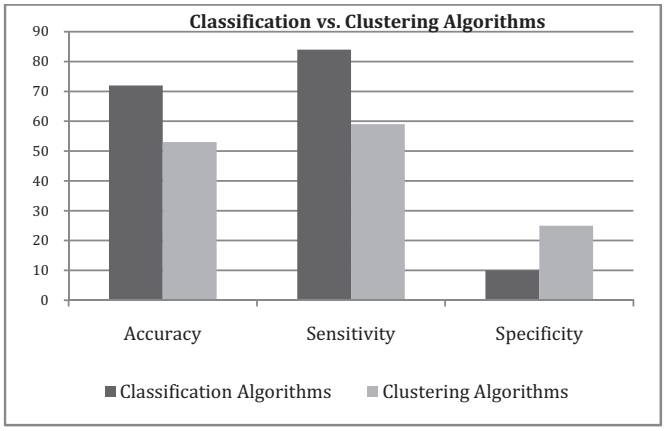

Fig. 2. Classification Vs. Clustering Algorithm

The behavior and performance of predicatevariables can be analyzed through sensitivity analysis, which helps to determine the output results.Sensitivity of classification algorithmswas 0.8404 as compared to 0.5934 of clustering algorithms. Specificity of classification algorithms was found to be 0.1036 as compared to specificity of 0.2504 of clustering algorithms. It can,therefore, be concluded that classification algorithms can lead to better results for predicting the cause of breast cancer.

B. Analysis on C5.0

C5.0,an iterative algorithm, keeps improving with trials.This algorithm has an added adaptiveboostingfeature thatworks by generating multiple classifiers (either decision trees or rule sets)

[11]. In this boosting feature, a new object is classified only by voting of all existing classifiers that predict the class of this object.Different trials were applied on the complete dataset as training data using rule based model to determine (i)the number of trials necessary to get 100% accuracy; and, (ii)to find critical parameters that should be given more importance in predicting the result. The default trial 1 achieved accuracy of 88.1%. It means the 'Perimeter' attribute itself is most critical factor for getting the result for predicting the recurrence of the disease. In successive trials we found that 100% accuracy was achieved in trial 7 although all the attributes were not used. After Trial 30, results remained constant as shown in Table II. This result can be used for diagnosis in medical practice to verify whether the findings about the important attributes that helps in predicting recurrent cases of cancer. This is an interesting study in the field of clinical investigation to find how accurate these algorithm works.

TABLE II. TRIAL RESULTS OF C5.0

Trials	Accuracy Rate	Error Rate	Attribute Usage	Time
Trial 1	88.1%	11.9%	Perimeter (90.21%)	0.0secs
Trial 5	97.42%	2.58%	Symmetry, Lymph Node (100%)	0.0secs
Trial 7	100%	0.0%	Symmetry, Lymph Node (100%)	0.0secs
Trial 15	100%	0.0%	Smoothness, Perimeter, Symmetry, Worst area, Lymph Node	0.1secs
Trial 30	100%	0.0%	Smoothness, compact, fractal Dimension, perimeter, symmetry, worst texture, worst area, worst concave points, tumor size, lymph nodes	0.1secs

V. CONCLUSION

Using prediction model to classify recurrent or non-recurrent cases of breast cancer is a researchthat 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. Eight popular data mining methods have been used,four from clustering algorithms (K-means, EM, PAM and Fuzzy c-means) and four from classification algorithms (SVM, C5.0, KNN and Naive Bayes). The results of these algorithms are clearly outlined in this paper with necessary results.The classification algorithms, C5.0 and SVM have shown 81% accuracy in classifying therecurrence of the disease.This is found to be bestamong all. On the other hand, EM was found to be the most promising clustering algorithm with the accuracy of 68%. The research shows that the classification algorithms are better predictor than clustering algorithms.The impact factors of various parameters responsible for predicting the occurrence/non-occurrence of the disease can be verified clinically. Further, the identified critical parameters should be verified by applying on larger medical dataset to predict the recurrence of the disease in future.

REFERENCES

[1] J. Ferlay, Globocan 2012 v1.0 Cancer Incidence and Mortality Worldwide: Iarc Cancerbase 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] R Core Team2013, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, [online] Available: http://www.R-project.org.

[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] P. C. Pendharkar, J. A. Rodger, G. J. Yaverbaum, N. Herman, M. Benner, "Association statistical mathematical and neural approaches for mining breast cancer patterns", Exper. Syst. with Applicat., vol. 17, pp. 223-232, 1999.

[9] A. P. Dempster, N. M. Laird, D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", J. Roy. Stat. Soc., vol. 39, no. 1, pp. 1-38, 1977.

[10] J. R. Quinlan, "Simplifying Decision Trees", International journal ofMan-Machine Studies, vol. 27, pp. 221-234, 1987.

[11] M.Kuhn, S.Weston, N.Coulter, M.Culp, Package C50. 2015 [online] Available: http://cran.r-project org/web/packages/C50/index.html.