Predicting Breast Cancer Recurrence using effective Classification and Feature Selection technique

Ahmed Iqbal Pritom
iqbal.cse@green.edu.bd
Lecturer, Dept. of CSE
Green University of Bangladesh
Bangladesh

Shahed Anzarus Sabab
sabab.iutcse@gmail.com
Lecturer, Dept. of CSE
Northern University Bangladesh
Bangladesh

Md. Ahadur Rahman Munshi
ahad_1114095@live.com
Lecturer, Dept. of CSE
Green University of Bangladesh
Bangladesh

Shihabuzzaman Shihab
shihabuzzaman.cse@green.edu.bd
Lecturer, Dept. of CSE
Green University of Bangladesh
Bangladesh

Abstract—**Breast cancer is a major threat for middle aged women throughout the world and currently this is the second most threatening cause of cancer death in women. But early detection and prevention can significantly reduce the chances of death. An important fact regarding breast cancer prognosis is to optimize the probability of cancer recurrence. This paper aims at finding breast cancer recurrence probability using different data mining techniques. We also provide a noble approach in order to improve the accuracy of those models. Cancer patient's data were collected from Wisconsin dataset of UCI machine learning Repository. This dataset contained total 35 attributes in which we applied Naive Bayes, C4.5 Decision Tree and Support Vector Machine (SVM) classification algorithms and calculated their prediction accuracy. An efficient feature selection algorithm helped us to improve the accuracy of each model by reducing some lower ranked attributes. Not only the contributions of these attributes are very less, but their addition also misguides the classification algorithms. After a careful selection of upper ranked attributes we found a much improved accuracy rate for all three algorithms.**

Keywords—breast cancer; recurrence; attribute selection; decision tree; SVM; Naïve Bayes; ROC curve; ranker.

I. INTRODUCTION

After increasing at an alarming rate for more than 20 years, breast cancer incidence rates in women began decreasing in 2000, and dropped by about 7% from 2002 to 2003[1]. But stats have shown that nearly 1.7 million new cases had been diagnosed in 2012 which is second most common cancer overall. This represents about 12% of all new cancer cases and 25% of all cancers in women [2]. Early detection and prediction is considered to be the best way to fight against this deadly disease. Most importantly, predicting the recurrence of cancer has become a real-world medical problem. Recurrent breast cancer is cancer that comes back in the same or opposite breast or chest wall after a period of time when the cancer couldn't be detected.

Recently, Data mining has become a popular and efficient tool for knowledge discovering and extracting hidden patterns from large datasets. It involves the use of sophisticated data manipulation tools to discover previously unknown, valid patterns and relationships in large dataset. We applied three strong data mining classification algorithms i.e. SVM, Naive Bayes and C4.5 Decision Tree on a medium sized dataset which contained 35 attributes and 198 cancer patient data. We classified the instances on the basis of 'output' attribute which can have only two nominal values i.e. 'R' for recursive cancer and 'N' for non-recursive cancer. Results have shown that SVM has higher prediction accuracy i.e. 75.75 % than Naïve Bayes (67.17 %) and C4.5 (73.73 %) methods.

Data sets having less attributes and higher instances can provide good result [3] than the result we have got using this data set where it has higher attributes and less instances. Too much attributes can miss guide a classifier from gaining its maximum result [4, 5], which gave us the idea of feature selection method. Feature selection algorithm [6] gave us upper ranked attributes as well as better result than the result we got without feature selection algorithm. In case of SVM top 11 upper ranked attributes gave us 1.5% improved result while Decision Tree gave 2.53% improved result for top 10 upper ranked attributes. But in case of Naïve Bayes improvement was 9.09% for top 8 upper ranked attributes.

The remaining section of this paper contains all related works that we have gone through. This includes background information on breast cancer research, prognosis factors, uses of ranking algorithm, several data mining techniques for breast cancer estimation and comparisons among their accuracies. This is followed by the method section which explains the proposed classification techniques. Finally, the results section is followed by a conclusion section. We tried to focus on all existing experiment's merits and demerits and targeted the

issues that the pervious solutions did not admit. We tried to come up with an optimized feature selection method which will not only improve the classifications, but also will be a key point to be remembered while using such heavily featured datasets.

Section II discuss about the related works. Section III discuss about the dataset. Section IV describes the evaluation method. Section V shows the experimental results. Section VI is the future scope.

II. Related Works

Bellaachia and Guven (2006) [7] investigated the accuracy of Naïve Bayes, the back-propagated neural network, and the C4.5 decision tree algorithms on SEER dataset. This dataset contained 16 attributes and 482,052 records. This is an ideal dataset which has a huge amount of patient data and a moderate number of attributes. According to their experiment, C4.5 algorithm gave the best performance of 86.7% accuracy. Jahanvi Joshi et al. [8] provided sophisticated evidence that KNN gives better accuracy than Expectation Maximization (EM) classification algorithm. Using Farthest First (FF) algorithm they declared 80% patient were healthy and 20% patient were sick, which is very near to the KNN algorithm result. Vikas Chaurasia and Saurabh Pal [9] claimed that Simple Logistic can be used for reducing the dimension of feature space and their proposed Rep Tree and RBF Network model can be used to obtain fast automatic diagnostic systems for other diseases. Correct classification rate of their proposed system is 74.5%. Pan wen [10] conducted experiments on ECG data to identify abnormal high frequency electrocardiograph using decision tree algorithm C4.5 with bagging. Delen et al. [11] among neural networks, decision trees and logistic regression, decision trees proved to be the best classifier for cancer prognosis using SEER data.

Huan Liu and Lei Yu [12] introduced active feature selection which promotes the idea to actively selecting instances for feature selection. S.Vanaja et al. [13] showed that each feature selection methodology has its own advantages and disadvantages inclusion of larger attribute causes the reduction of accuracy. Dong-Sheng Cao's [14] proposed a new decision tree based method combined with feature selection method backward elimination strategy with bagging to find the structure activity relationships in the area of chemo metrics related to pharmaceutical industry. Liu Ya-Qin's [15] experimented on breast cancer data using C5 algorithm with bagging to predict breast cancer survivability. Medhat et al. [16] found promising and superior result for SVM over Decision tree witnessed by minimum error rate and maximum average gain.

III. Breast cancer wisconsin data set summary

The data used in this study are provided by the UC Irvine machine learning repository located in breast-cancer Wisconsin sub-directory, filenames root: breast-cancer-Wisconsin having 198 instances, 1 class attribute named 'Outcome' with two possible results (R = recur, N = non-recur) and 34 other attributes. Attribute 'Lymph node status' is missing in 4 cases. Class distribution: 151 non-recur, 47 recur. Each record represents follow-up data for one breast cancer case. These are consecutive patients seen by Dr. Wolberg since 1984, and include only those cases exhibiting invasive breast cancer and no evidence of distant metastases at the time of diagnosis. The first 30 features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Ten real-valued features were computed for each cell nucleus. Which are, radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness (perimeter^2 / area - 1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry and fractal dimension ("coastline approximation" - 1). The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. The other attributes are ID number, Outcome (R = recursive cancer, N = non-recursive cancer), Time (recurrence time if field 2 = R, disease-free time if field 2= N), Tumor size (diameter of the excised tumor in centimeters) and Lymph node status (number of positive axillary lymph nodes observed at time of surgery). This summarizes all 35 attributes of our dataset.

IV. Evaluation Method

In this paper, we investigated three data mining techniques: Support Vector Machine, C4.5 Decision Tree and Naïve Bayes to predict the recurrence of breast cancer from Wisconsin breast cancer data set. We used Weka machine learning tool for all our classifications. Weka [6] is a collection of machine learning algorithms for data mining tasks. We used Weka version 3.6.9 for all our preprocessing and classifying. In order to maintain a fair measure of the performance of the classifier, we used 10 fold cross validation technique for all three algorithms. In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data. We used 66% of our dataset as Training data and remaining 34% were used as test data.

To improve the accuracy of recurrence prediction, we used Ranker algorithm for best feature selection and for the removal of redundant and irrelevant attributes. InfoGainAttributeEval was selected as attribute evaluator for Naïve Bayes and C4.5 Decision Tree. It evaluates the worth of an attribute by measuring the information gain with respect to the class. On the other hand, for SVM, we choose SVMAttributeEval as attribute evaluator which evaluates the worth of an attribute by using an SVM classifier. Here, attributes are ranked by the square of the weight assigned by the SVM. Attribute selection for multiclass problems is

handled by ranking attributes for each class separately using a one-vs.-all method and then "dealing" from the top of each pile to give a final ranking.

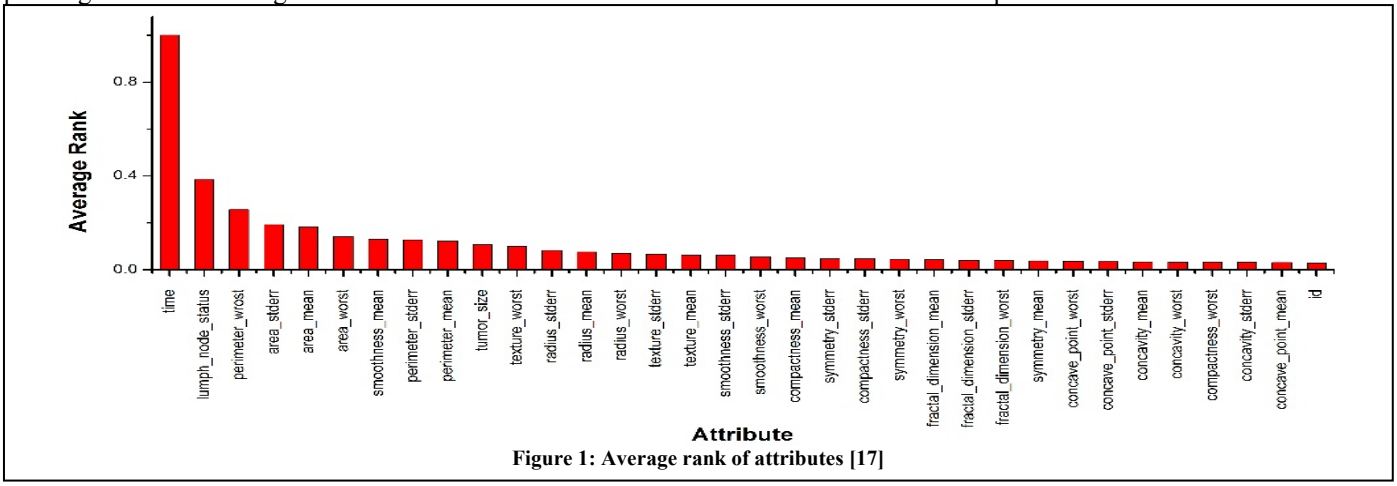

Figure 1: Average rank of attributes [17]

Figure 1 shows the average rank of all 34 attributes (except the 'outcome' which was set as class attribute). This is a clear indication that while classifying, heavily ranked attributes like 'time', 'lymph node status' and 'perimeter worst' will contribute significantly and will control the accuracy of the prediction. Again, features like 'id' and 'concave point mean' can be considered as light weight attributes who have very little to contribute in building this prediction models.

V. Experimental Results

First we experimented the recurrence prediction accuracy for all 3 classification algorithms without applying any ranker. Among them, the best result was recorded for SMO (John Platt's sequential minimal optimization algorithm for training a support vector classifier) which provided 75.75% accuracy. J48 (generates a pruned or unpruned C4.5 decision tree) and Naïve Bayes gave 73.73% and 67.17% accuracies respectively.

We believed that through proper attribute selection for classification and removal of redundant attributes, this result could be improved by a fare margin. So, after ranking all attributes, we selected 'N best attributes' for all 3 classifiers. Our target was to find the combination of N best features that predicts with maximum accuracy. Here, value of N can be any integer value from 1 to 35 as we have 35 classifying features in our dataset.

We experienced that, SMO gave maximum accuracy when we carefully selected best 11 features from our dataset. So, we removed remaining 25 less important attributes and performed SVM classification again. This time, the accuracy was improved by 1.52%. C4.5 Decision Tree and Naive Bayes both gave improved maximum accuracy of 76.26%. Their accuracies improved by 9.09% and 2.53% respectively compared to their previous results when we din't apply any feature selection technique. We recorded this result for C4.5 Decision Tree using 10 best attributes while for Naïve Bayes, this result was achieved using best 8 attributes. Table 1 summarizes our experiment result for this section. In 'With Ranker' column, we included the result found using N best attribute selection technique.

Table 1: Performance of the Classifier

Evaluation Criteria	Classifiers					
	SMO		C4.5 Decision Tree		Naïve Bayes	
	Without Ranker	**With Ranker**	**Without Ranker**	**With Ranker**	**Without Ranker**	**With Ranker**
Timing to build model (in Sec)	0.095	22.72	0.085	23.74	0.045	23.74
Correctly classified instances	150	153	146	151	133	151
Incorrectly classified instances	48	45	52	47	65	47
Accuracy (%)	75.75%	77.27%	73.73%	76.26%	67.17%	76.26%

Undoubtedly, SMO provided best breast cancer recurrence accuracy with and without applying Ranker algorithm. In figure 2 we showed that every evaluation criteria achieves certain level of improvement when appropriate feature selection is confirmed. This is a strong evidence that medium sized dataset with large amount of attributes can easily be misguided by the additional amount of features with very less contribution in classification.

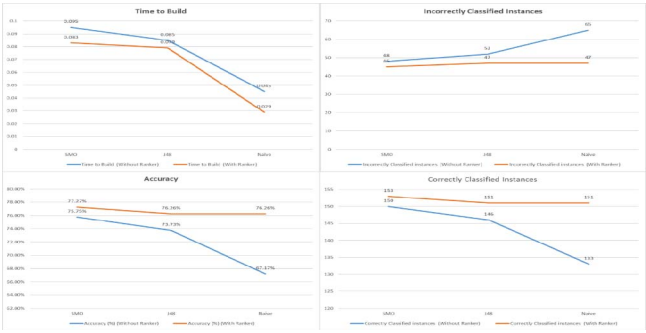

Figure 2: Comparative graph of different classifier showing at different evaluation criteria [9]

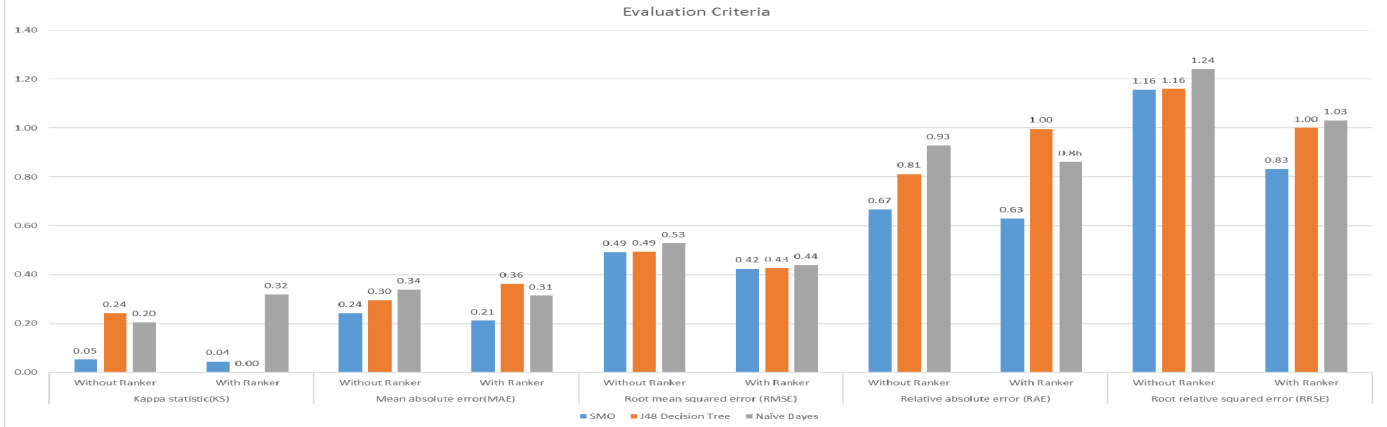

Figure 3: Comparison between Parameters [9]

Table 2: Training and Simulation Error

Evaluation Criteria	Classifiers					
	SMO		J48 Decision Tree		Naïve Bayes	
	Without Ranker	**With Ranker**	**Without Ranker**	**With Ranker**	**Without Ranker**	**With Ranker**
Kappa statistic (KS)	0.053	0.044	0.241	0	0.204	0.319
Mean absolute error (MAE)	0.242	0.211	0.295	0.362	0.338	0.313
Root mean squared error (RMSE)	0.492	0.423	0.494	0.426	0.529	0.439
Relative absolute error (RAE)	66.6%	62.9%	81.2 %	99.6 %	92.9%	86.1%
Root relative squared error (RRSE)	115.6%	83.3%	116.1 %	99.9%	124.2 %	103.1%

We acknowledged that error rate is reduced significantly for almost every evaluation criteria after including ranker algorithm. In SMO, all 5 error rate showed consistent improvement after applying Ranker. In case of J48 Decision Tree, Kappa statistic, Root mean squared error and Root relative squared error rate were improved. And Naïve Bayes showed improved error rate in all evaluation criteria except Kappa statistics. Table 2 summarizes our experiment result for this section.

Figure 3 shows the statistical comparison among three classification algorithms. Again it shows the improved error rate gain using 'N best attributes' selection technique. In every possible evaluation criteria, SMO is showing far too improved and reduced error rate.

Our final evaluation criteria was to find the true positive and false positive rate and also to generate a Receiver operating characteristic (ROC) curve. Without proper feature selection, Area under ROC curve was very poor for all three

Table 3: Comparison of Accuracy measure

Evaluation Criteria	Classifiers					
	SMO		J48 Decision Tree		Naïve Bayes	
	Without Ranker	**With Ranker**	**Without Ranker**	**With Ranker**	**Without Ranker**	**With Ranker**
True Positive Rate	0.758	0.766	0.737	0.763	0.672	0.763
False Positive Rate	0.72	0.731	0.507	0.751	0.439	0.455
Area Under ROC Curve	0.519	0.529	0.528	0.745	0.642	0.699

Some key points should be noted here. Even after the careful selection of attributes, area under roc curve is not up to the mark for any algorithm. We can consider the accuracy of J48 (AUC=0.745) 'Fair' while accuracies of Naïve Bayes (AUC = '0.699') and SMO (AUC = '0.529') should be considered as 'Poor'. The improvemrnt of ROC Curve shows the importance of best feature selection for datasets. Figure 4, 5 and 6 summarizes our result of this section.

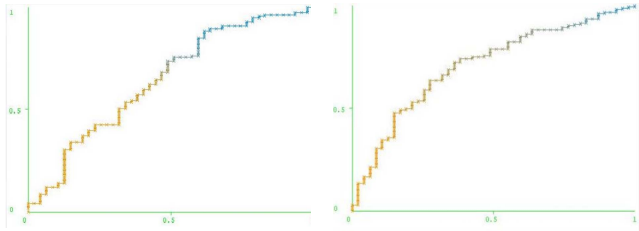

Figure 4: AUC improvement in Naïve Bayes (without Ranker in left and with Ranker in Right)

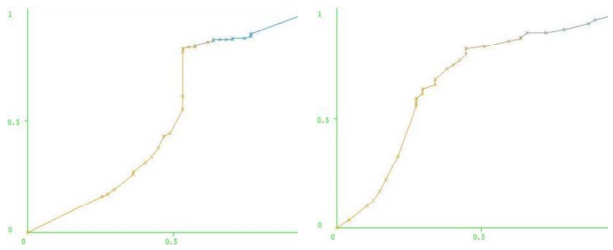

Figure 5: AUC improvement in J48 Decision Tree (without Ranker in left and with Ranker in Right)

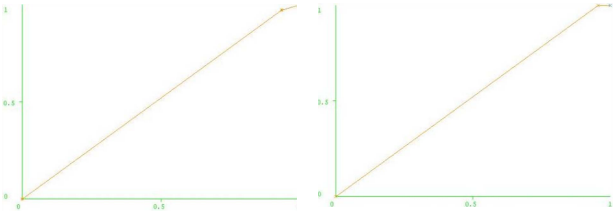

Figure 6: AUC improvement in SMO (without Ranker in left and with Ranker in Right)

VI. Conclusion and Future Scope

In this paper we tried to focus on the importance of feature selection in breast cancer prognosis. Using proper attribute selection technique, any classification algorithm can be improved significantly. Attributes with less contribution in dataset often misguides the classification and results in poor prediction. In our work, we found Support Vector Machine giving much better output both before and after attribute selection. Area under ROC curve analysis showed results in our favor where Naïve Bayes and Decision Tree showed much better improvement after feature selection method. In future we will try to evaluate some newer algorithms with better feature selection technique. In this paper we only focused on whether breast cancer is recursive or not. In addition of this work, we will try to predict the time of recurrence of cancer which is classified as recursive.

VII. Reference

[1]http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-key-statistics

[2] http://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/breast-cancer-statistics

[3]https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)

[4] Wolberg, William H., and Olvi L. Mangasarian. "Multisurface method of pattern separation for medical diagnosis applied to breast cytology." Proceedings of the national academy of sciences 87.23 (1990): 9193-9196.

[5] Zhang, Jianping. "Selecting typical instances in instance-based learning." Proceedings of the ninth international conference on machine learning. 1992.

[6] Weka 3.5.6, An open source data mining software tool developed at university of Waikato, New Zealand, http://www.cs.waikato.ac.nz/ml/weka/ 2009.

[7] Bellaachia, Abdelghani, and Erhan Guven. "Predicting breast cancer survivability using data mining techniques." Age 58.13 (2006): 10-110.

[8] Joshi, Jahanvi, Rinal Doshi, and Jigar Patel. "Diagnosis of breast cancer using clustering data mining approach." International Journal of Computer Applications 101.10 (2014).

[9] Chaurasia, Vikas, and Saurabh Pal. "A novel approach for breast cancer detection using data mining techniques." International Journal of Innovative Research in Computer and Communication Engineering 2.1 (2014): 2456-2465.

[10] Wen, Pan. "Application of decision tree to identify abnormal high frequency electro-cardiograph." Physics Experimentation 11 (2009): 011.

[11] D. Delen, G.Walker, A.kadam, "Predicting breast cancer survivability: a comparison of three data mining methods", Artificial Intelligence in Medicine (2005) 34, 113—127

[12] Liu, Huan, and Lei Yu. "Toward integrating feature selection algorithms for classification and clustering." IEEE Transactions on knowledge and data engineering 17.4 (2005): 491-502.

[13] Vanaja, S., and K. Ramesh Kumar. "Analysis of feature selection algorithms on classification: a survey." International Journal of Computer Applications 96.17 (2014).

[14] Cao, Dong-Sheng, et al. "Automatic feature subset selection for decision tree-based ensemble methods in the prediction of bioactivity." Chemometrics and Intelligent Laboratory Systems 103.2 (2010): 129-136.

[15] Ya-Qin, Liu, Wang Cheng, and Zhang Lu. "Decision tree based predictive models for breast cancer survivability on imbalanced data." 2009 3rd International Conference on Bioinformatics and Biomedical Engineering. 2009.

[16] Abdelaal, Medhat Mohamed Ahmed, et al. "Using data mining for assessing diagnosis of breast cancer." Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on. IEEE, 2010.

[17] Khan, Muhammad Umer, et al. "Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare." 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, 2008.