Analysis of Breast Cancer detection using different Machine learning techniques

Abstract. *Data mining algorithms play an important role in the prediction of early-stage breast cancer. In this paper, we propose an approach that improves the accuracy and enhances the performance of three different classifiers: Decision Tree (J48), Naïve Bayes (NB), and Sequential Minimal Optimization (SMO). We also validate and compare the classifiers on two benchmark datasets: Wisconsin Breast Cancer (WBC) and Breast Cancer dataset. Data with imbalanced classes are a big problem in the classification phase since the probability of instances belonging to the majority class is significantly high, the algorithms are much more likely to classify new observations to the majority class. That is what we tried to deal with in this work. We tend to use the data level approach which consists of resampling the data in order to mitigate the effect caused by class imbalance. For evaluation, 10 fold cross-validation is performed to ensure the effectiveness of our models to mitigate overfitting. The efficiency of each classifier assesses in terms of accuracy, precision, and recall, true positive and false positive. Experiments show that using a resample filter enhances the classifier's performance where SMO outperforms others in the WBC dataset and J48 is superior to others in the Breast Cancer dataset*.

Keywords: breast cancer, classification, data mining.

1 Introduction

Breast cancer is the second leading cause of death among women worldwide [1]. In 2019, there are 268,600 new cases of invasive breast cancer are expected to be diagnosed in women in the U.S along with 62,930 new cases of non-invasive breast cancer [2]. Early detection is the best way to increase the chance of treatment and survivability. Data mining has become a popular tool for knowledge discovery which show good results in marketing, social science, finance and medicine. Recently, multiple classifiers algorithms are applied on medical datasets to perform predictive analyzes about patients and their medical diagnosis. For example, using machine learning techniques to assess tumor behavior for breast cancer patients. This paper introduces a comparison between three different classifiers: J48, NB, and SMO with respect to accuracy in detection of breast cancer. Our aim is to prepare the dataset by proposing a suitable method that can manage the imbalanced dataset and the missing values to enhance the classifier's performance. All tasks conducted using software Weka 3.8.3.

The remainder of this paper is organized as follows. Section 2 presents literature review. Section 3 introduces the datasets. Section 4 describes the research methodology

including pre-processing experiments, classification and performance evaluation criteria. The experimental results are presented in section 5 and we conclude in Section 6.

2 Literature review

In recent years, several studies have applied data mining algorithms on different medical datasets to classify Breast Cancer. These algorithms show good classification results, which motivated the use of this method in this work. Table 1, summarize few of literature survey.

3 Datasets

The datasets that are used in this paper are available at the UCI Machine Learning Repository [13].

3.1 WBC Dataset

WBC dataset contains 699 instances and 11 attributes in which 458 were benign and 241 were malignant cases [14]. In the WBC, the value of the attribute (Bare Nuclei) status was missing 16 records. Hence data preprocessing is essential and important phase for classification purpose we need to manage the imbalanced data and the missing values.

3.2 Breast Cancer Dataset

Dataset's features are computed from a digitized image of a fine needle aspirate (FNA) of a breast tumor. The target feature records the prognosis malignant or benign. The dataset contains 286 instances and 10 attributes in which 201 were no-recurrence-events and 85 were recurrence events. In the Breast Cancer dataset, the value of the attribute (node-caps) status was missing 8 records.

Table 1. Breast cancer detection captions using different machine learning algorithms.

Paper title	Datasets	Algorithms	Results
Integration of Data mining Classification Techniques and Ensemble Learning for Predicting the Type of Breast Cancer Recurrence [3], 2019	Breast Cancer	NB, SVM, GRNN and J48	GRNN & J48 accuracy: 91% NB & SVM: 89%
A Study on Prediction of Breast Cancer Recurrence Using Data mining techniques [4], 2017	WPBC	Classification: KNN, SVM, NB and C5.0 Clustering K-means, EM, PAM and Fuzzy c-means	Classification accuracy is better than clustering, SVM & C5.0: 81%
Predicting Breast Cancer Recurrence using effective Classification and Feature Selection technique [5], 2016	WPBM	NB, C4.5, SVM	NB: 67.17%, C4.5: 73.73%, SVM: 75.75%
Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis [6], 2016	WBC	SVM, C4.5, NB, KNN	SVM outperform others: 97.13%
Study and Analysis of Breast Cancer Cell Detection Using Naïve Bayes, SVM and Ensemble Algorithms [7], 2016	WDBC	NB, SVM, Ensemble	SVM: 98.5% NB & Ensemble: 97.3%
Analysis of Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection [8], 2015	WDBC	NB, J48	NB: 97.51% J48: 96.5%
Comparative Study on Different Classification Techniques for Breast Cancer Dataset [9], 2014	Breast Cancer	J48, MLP, rough set	J48:79.97%, MLP:75.35%, rough set:71.36%
A Novel Approach for Breast Cancer Detection using Data mining Techniques [10], 2014	WBC	SMO, IBK, BF Tree	SMO: 96.19%, IBK: 95.90%, BF Tree: 95.46%
Experiment Comparison of Classification for Breast Cancer Diagnosis [11], 2012	WBC WDBC WPBC	J48, SMO, MLP, NB, IBK	In WBC: MLP & J48 : 97.2818% In WDBC: SMO: 97.7% or fusion on SMO & MLP:97.7% In WPBC: fusion of MLP, J48, SMO and IBK: 77%
Analysis of Feature Selection with Classification: Breast Cancer Datasets [12], 2011	WBC WDBC Breast Cancer	Decision Tree with and without feature selection	Feature selection enhance the results WBC: 96.99% WDBC: 94.77% Breast Cancer:71.32%

4 Research Methodology

The two datasets used in this work are vulnerable to missing and imbalanced data therefore, before performing the experiments, a large fraction of this work will be for preprocessing the data in order to enhance the classifier's performance. Preprocessing will

focus on managing the missing values and the imbalanced data. To manage the missing attributes, all the instances with missing values are removed. The imbalance data problem needs to adjust either the classifier or the training set balance. To do so, we re-balance them artificially using the resample filter that implemented in Weka. Then a comparison between the three classifiers is applied.

4.1 Preprocessing phase

First, the data were discretized using discretize filter in Weka, then instances were resampled using the resample filter in order to maintain the class distribution in the subsample and to bias the class distribution toward a uniform distribution. Section 5 will show that this idea is improving the classifier's performance. Second, removing the missing values from the dataset. The new total number of instances become 683 records in the WBC dataset and 278 cases in the Breast Cancer dataset. Third, 10 fold cross validation was applied then experiments applied over three classifiers Naïve Bayes, SMO and J48 as illustrated in Fig 1.

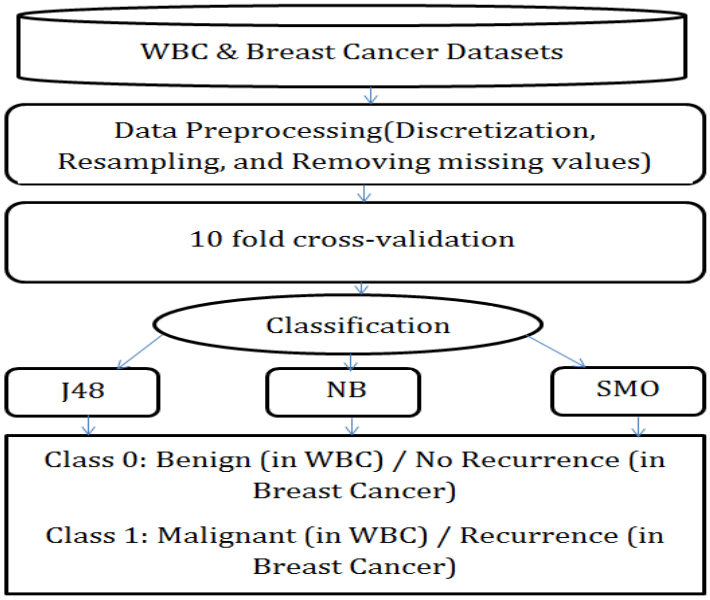

Fig. 1. Proposed breast cancer detection model using Breast Cancer and WBC datasets.

In Fig.1 at first data preprocessing technique has been applied including three processes discretization, instances resampling and removing the missing values. After that, 10 fold cross validation has been applied, then three classifiers has been applied over the prepared datasets.

4.2 Training & Classification

In order to minimize the bias associated with the random sampling of the training data, we tend to use 10 fold cross validation after the pre-processing phase. In k-fold cross-validation, the original dataset is randomly partitioned into k equal size subsets. The classification model is trained and tested k times. Each time, a single subset is retained as the validation data for testing the model, and the remaining k-1 subsets are used as training data. Three classification techniques were selected: a Naïve Bayes (NB), a Decision J48, and *a* Sequential Minimal Optimization (SMO). The NB classifier is a probabilistic classifier based on the Bayes rule. It works by estimating the portability of each class value that a given instance belongs to that class [15]. The J48 algorithm [16] it uses the concept of information entropy and works by splitting each data attributers into smaller datasets in order to examine entropy differences. It is an improved and enhanced version of C4.5 [17]. The SMO model implements John Platt's sequential minimal optimization algorithm for training a support vector classifier. This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default [18].

4.3 Performance evaluation criteria

In this study, we use four performance measures to evaluate all the classifiers: true positive, false positive, ROC curve and accuracy (AC).

$$AC = (TP+TN)/(TP+TN+FP+FN) \qquad (1)$$

Where TP, TN, FP and FN denote true positive, true negative, false positive and false negative, respectively.

5 Experimental Results

First, we experimented the WBC and the Breast Cancer datasets for all the three classification algorithms J48, NB and SMO without applying any preprocessing. Among them, the best result was recorded for J48:75.52% in the Breast Cancer dataset and for SMO: 94.56% in the WBC dataset. Next, after applying preprocessing techniques accuracy increases to 97.12% with J48 in the Breast Cancer dataset and 99.56% with SMO in the WBC dataset.

5.1 Experiment using the Breast Cancer Dataset

First, we test the three classifiers with their original values (without any preprocessing).The results show that J48 is the best one with 75.52% accuracy where the accuracy of NB and SMO are 71.67% and 69.58%, respectively. Next, we apply discretization filter and remove the records with missing values, results improved with NB and SMO as follows: NB: 74.82% and SMO: 72.30% where J48: 74.82%. After that, resample filter was applied for 9 times. The Performance of the classifiers are improved and enhanced as shown in Table 4.

Table 4. Performance of the classifiers in the Breast Cancer Dataset.

Experiments steps	Classifier accuracy / time to build classifier (sec)		
	J48	NB	SMO

Original without preprocessing	75.52% / 0.03	71.67% / 0	69.58% / 0.16
After removing missing values & discretization	74.82% / 0	75.53% / 0	72.66% / 0.03
After applying resample filter (first time)	79.49% / 0.02	77.33% / 0	80.93% / 0.06
applying resample filter (second time)	81.65% / 0	78.05% / 0	80.57% / 0.09
applying resample filter (third time)	87.41% / 0	78.41% / 0	82.73% / 0.03
applying resample filter (fourth time)	92.08% / 0	77.69% / 0	88.84% / 0.05
applying resample filter (fifth time)	95.68% / 0	79.13% / 0	91.72% / 0.03
applying resample filter (sixth time)	97.48% / 0	79.85% / 0	95.68% / 0.03
applying resample filter (seventh time)	98.20% / 0	76.61% / 0.02	95.32% / 0.03
applying resample filter (eightth time)	97.12% / 0	79.49% / 0	96.04% / 0.03
applying resample filter (nineth time)	97.12%/ 0	86.33% / 0	96.76%/ 0.03

As illustrated in Table 4, we can obviously notice that the more resample filter we apply, the improved accuracy we obtain. That is because the data is imbalanced and the filter maintains the class distribution. For the Breast cancer dataset, J48 outperforms others with 97.12%. Accuracy measures for J48 classifier is shown in Table 5. Roc curve of J48 is shown in Fig. 2.

Table 5. Accuracy measures for J48 in the Breast Cancer Dataset.

TP	FP	Precision	Recall	Roc curve	Std	class
0.969	0.110	0.955	0.996	0.981	0.5678	no-recurrence-events
0.890	0.031	0.924	0.996	0.981		recurrence-events

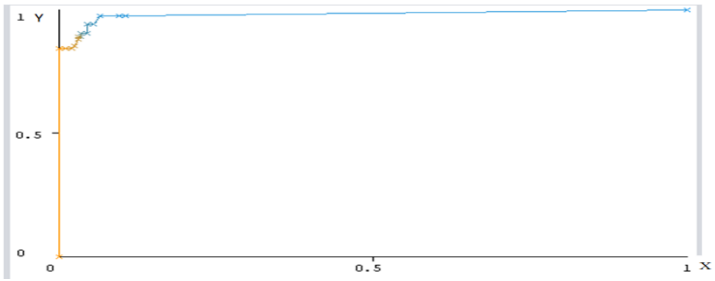

Fig. 2. J48 ROC curve in Breast Cancer Dataset.

Comparing these results with study proposed in [9], using the same dataset and three classifiers including J48 algorithm, we can obviously state that J48 classifier's accuracy is much more better using the resample filter for the pre-processing phase rather than feature selection technique as illustrated in Table 5.

Table 5. Compression of accuracy measures for the Breast Cancer Dataset.

Methodology	Study [9]	Proposed method
With out pre-processing	None	J48: 75.52% ,NB: 71.67% SMO: 69.58%
With pre-processing	Missing values were replaced with WEKA pre-processing techniques and feature selection was applied J48: 79.97%, MLP: 75.35% & rough set: 71.36%	Delete records of missing values and Descretization J48: 74.82% ,NB: 57.53% SMO: 72.66%
Using the resample filter	None	Applying the resample filter for 9 times J48: 97.12% ,NB: 86.33% SMO: 96.76%

5.2 Experiment using the WBC Dataset

Same experiments were applied with the WBC dataset. With respect to apply preprocessing techniques all algorithms present higher correct classification accuracy, the different lies in the fact that using the resample filter several times improves the classification accuracy. SMO classifier achieve 99.56% efficiency compared to 99.12% of the Naïve Bayes and 99.24% of the J48. Results are illustrated in Table 6.

Table 6. Performance of the classifiers in WBC dataset.

Experiments steps	Classifier accuracy / time to build classifier (sec)		
	J48	NB	SMO
Original Without preprocessing	94.56% / 0.12	95.99% / 0.01	96.99% / 0.08
After removing missing values & discretization	95.91% / 0	97.37%/ 0	96.78% / 0.03
After applying resample filter (first time)	95.91% .0	97.51% / 0	98.97% / 0.03
applying resample filter (second time)	97.95% / 0	98.10% / 0	99.41% / 0.03
applying resample filter (third time)	98.68% / 0	98.10% / 0	99.12% / 0.02
applying resample filter (fourth time)	99.24% / 0	99.12% / 0	99.56% / 0 .02

In the WBC dataset, SMO superior than others with 99.56%. Accuracy measures for SMO classifier is shown in Table 7. Roc curve of SMO is shown in Fig. 3.

Table 7. Accuracy measures for SMO in WBC Dataset.

TP	FP	Precision	Recall	Roc curve	Std	class
0.996	0.004	0.998	0.996	0.996	0.2220	benign
0.996	0.004	0.992	0.996	0.996		malignant

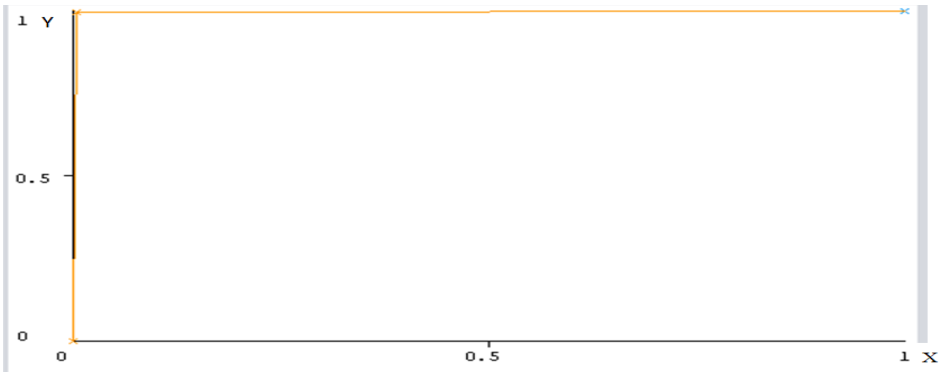

Fig. 3. SMO ROC curve in Breast Cancer Dataset.

In terms of the WBC dataset, our proposed method is compared with two studies [6, 10]. Results shows that the performance of SMO classifier is better since our model employs pre-processing, and resampling approaches. Thus, utilizing pre-processing, and resampling techniques play an important role in increasing the SMO accuracy comparable to the other techniques in [6] & [10]. Details are shown below in Table 8.

Table 8. Compression of accuracy measures for the WBC Dataset.

Methodology	Study [6]	Study [10]	Proposed method
Without pre-processing	C4.5: 95% NB: 95.9% SVM: 97.3%	SMO: 96.19% IBK: 95.90% BF Tree: 95.46%	J48: 94.56% ,NB: 95.99% SMO: 96.99%
With pre-processing	None	None	Delete records of missing values and Descretization J48: 95.91% ,NB: 97.37% and SMO: 96.78%
Using the resample filter	None	None	Applying the resample filter for 4 times J48: 99.24% ,NB: 99.12%, SMO: 99.56%

6 Conclusion

Breast cancer is one of the major causes of death in women. Early detection of breast cancer is essential to save women's life. Breast cancer detection can be done with the help of modern machine learning algorithms. In this paper, we tried to focus on how to deal with data that is imbalanced and has missing values using resampling techniques implemented in Weka software tool in order to enhance the classification accuracy of detecting the breast cancer. In our work, three classifiers algorithms have been applied J48, NB and SMO on two different breast cancer datasets. Results show that using the resample filter in the preprocessing phase enhance the classifier's performance. In the future, same experiments can be applied on different classifiers and different datasets.

References

1. U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control
2. http://www.breastcancer.org/symptoms/understand_bc/statistics
3. Silva, J., Lezama, O. B. P., Varela N., and Borrero L. A, Integration of Data Mining Classification Techniques and Ensemble Learning for Predicting the Type of Breast Cancer Recurrence, In International Conference on Green, Pervasive, and Cloud Computing, Springer, pp. 18-30, 2019.
4. Ojha U., and Goel, S., A study on prediction of breast cancer recurrence using data mining techniques, In 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence, IEEE, pp. 527-530, 2017.
5. Pritom A. I., Munshi M. A. R., Sabab S. A., and Shihab S., Predicting breast cancer recurrence using effective classification and feature selection technique, In 19th International Conference on Computer and Information Technology (ICCIT), IEEE, pp. 310-314, 2016.
6. Asri H., Mousannif H., Al Moatassime H., and Noel T., Using machine learning algorithms for breast cancer risk prediction and diagnosis, Procedia Computer Science 83, pp.1064-1069, 2016.
7. Hazra A., Mandal S. K., and Gupta, A., Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms, International Journal of Computer Applications 145, pp. 0975-8887, 2016.
8. Borges, L. R., Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection, 1989.
9. Saabith, A. L. S., Sundararajan, E., & Bakar, A. A., Comparative study on different classification techniques for breast cancer dataset, *International Journal of Computer Science and Mobile Computing*, 2014.
10. Chaurasia V., and Pal, S., A Novel Approach for Breast Cancer Detection using Data Mining Techniques, International Journal of Innovative Research in Computer and Communication Engineering, *An ISO 3297: 2007 Certified Organization*, 2017
11. Salama G. I., Abdelhalim M. B., and Zeid, M. A. E., Experimental comparison of classifiers for breast cancer diagnosis, In 2012 Seventh International Conference on Computer Engineering & Systems (ICCES), IEEE, pp. 180-185, 2012.
12. Lavanya, D., & Rani, D. K. U., Analysis of feature selection with classification: Breast cancer datasets, Indian Journal of Computer Science and Engineering (IJCSE) pp.756-763, 2011.

13. Breast Cancer Wisconsin Dataset. Available at: UCI Machine Learning Repository.
14. Dataset Description. Available at: UCI Machine Learning Repository.
15. Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.
16. Quinlan, R. C., 4.5: Programs for machine learning, morgan kaufmann publishers inc. *1993.*
17. Quinlan, J. R., Simplifying Decision Trees", International journal of Man-Machine Studies, pp. 221-234, 1987.
18. Piatt, J., Fast Training of Support Vector Machines using Sequential Minimal Optimization, Advances in Kernel Methods-Support Vector Learning, 1998.