TEST Engineering & Management

November-December 2019
ISSN: 0193-4120 Page No. 4039 - 4043

Predicting Breast Cancer Survivability Rates a Differentiation of Three Data Mining Models

Ghofran Othoum
College of Engineering,
EFFAT University
AnNazlah Al Yamaniyyah,
Jeddah 22332, Saudi Arabia
gothoum@effatuniversity.edu.sa

Wadee Al-Halabi
College of Engineering,
EFFAT University
AnNazlah Al Yamaniyyah,
Jeddah 22332, Saudi Arabia
walhalabi@effatuniversity.edu.sa

Houria Oudghiri
College of Engineering,
EFFAT University
AnNazlah Al Yamaniyyah,
Jeddah 22332, Saudi Arabia
houdghiri@yahoo.com

Article Info
Volume 81
Page Number: 4039 - 4043
Publication Issue:
November-December 2019

Article History
Article Received: *5 March 2019*
Revised: *18 May 2019*
Accepted: *24 September 2019*
Publication: *19 December 2019*

Abstract:
The new approach in cancer research that shifted from pure long-term biological and clinical experiments to computer-generated experiments is the main inspiration for this project. Three Data mining techniques (Decision Trees, Neural Networks and Naïve Bayes) were built to compare their performance based on three main parameters: accuracy, sensitivity and specificity. The experiment was set up with multi-layer perceptron as the baseline scheme and with statistical significance of 0.05. The models were built using data collected from Saudi Arabia, more specifically, from King Faisal Specialist Hospital and Research Center. The prediction is based on 8 attributes: age, birth location, reason for no radiation, laterality, grade, sex, primary site and marital status. However, the data collected had around 680 instances which were not sufficient to build the models. Sampling with a random seed was completed to double the size of the training dataset. The results showed that decision tree had the highest accuracy and sensitivity with values (0.979, 0.988) respectively. Naïve bayes had the highest classification error (0.094) and neural networks had the highest specificity (0.896).

Keywords:- *Breast Cancer; data; mining; survival rates*

1. INTRODUCTION

In the last few decades, various treatments have decreased the number of fatalities as a direct result of breast cancer. This positive development can be attributed directly to clinical research advancement and to the complementary data usage in an advanced computational way [1].

According to the National Health Organization [2], breast cancer is the second most common reason for deaths for women. The Saudi Ministry of Health indicated in the latest published statistics [3] that around 500 to 700 women are prone to have breast cancer each year in Saudi Arabia. Moreover, 11.8 women between each 100000 Saudi women are diagnosed with breast cancer [4]. What is different in the statistics collected from Saudi Arabia and the Middle East is that breast cancer is diagnosed at earlier ages than at Western Countries [4].

Various researches, outside Saudi Arabia, have taken place to attempt to identify the factors behind this resistant epidemic and its results. Some research paper results indicate that genetic, lifestyle and medical history factors are the most influential factors in determining the onset probability and survivability rates of breast cancer patients [5].

However, as the number of breast cancers patients increases, the data to be analyzed also increase. In order to extract biologically significant information from redundant or insignificantcases, a structured approach must be adapted to reach an accurate generalization or conclusion. Data mining algorithms are considered one of these structured methodologies [6].

KDD (Knowledge Discovery in Databases) is the process of identifying statistically significant data from a group of datasets through sorting data, identifying patterns and establishing relationships. Part of this

November-December 2019
ISSN: 0193-4120 Page No. 4039 - 4043

process is data mining [7]. Data mining involves pattern analysis, data searching and data archeology [8]. The possible pathways that are commonly followed to discover significant knowledge are, data cleaning, data warehousing, data mining and pattern evaluation. More often, an iterative approach between these procedures is required to answer a KDD problem [9].

Data mining is a knowledge discovery technique that is used to generate new knowledge about events and phenomena from existing datasets. The end result could be manifested in classification of the dataset or in predicting future events. Data mining differs from any common statistical analysis process in different ways. First, data mining avoids the possibility of assigning patterns to randomly established associations. Data snooping, the corresponding term for data mining in statistics, use exhaustive search to identify patterns which might rise as a result of data randomness [10].

Machine learning is used in artificial intelligence to refer to a program's ability to predict patterns based on experience. Machine learning is used in many fields including search engines, bioinformatics, and handwriting recognition and robot locomotion [11]. Tools that implement machine learning algorithms, adjust their behavior based on studied examples. For a machine learning model to be effective, it needs to be trained using a training dataset, and its classification capabilities need a testing dataset to verify and use the model. If the model is trained by labeled example then the process is commonly known as supervised learning. When labeled examples are not available, then an unsupervised approach to machine leaning must be adapted [12].

Thus in this work was done to find the most accurate model for the breast cancer survivability problem with respect to three main classification models: neural networks, decision trees and naïve bayes (C4.5, Neural Network and Naïve Bayes) based on accuracy, specificity and sensitivity.

2. METHODOLOGY

2.1 Prediction Model

The three data mining algorithms (Artificial Neural Network: Multi-layer Perceptron Back Propagation), naïve Bayes and decision tree (C4.5), correspond to the following classes in WEKA"s java library respectively:

WEKA.classifiers.functions.multiLayerPerceptron,
WEKA.classifiers.Bayes.naieveBayes,
WEKA.classifiiers.trees.J48 [30].These models fit the non linear nature of the problem, which requires machine learning applications like neural networks. Table 1 is a summary description of the algorithms.

Table 1: Description of the Prediction models

Name	Brief Description	Group	WEKA class
Naïve Bayesian	Probabilistic induction	Bayes	WEKA.classifiers.Bayes.naieveBayes
Decision Tree	Extends C4.5 algorithm	Tree	WEKA.classifiiers.trees.J48
Neural Network	Multi-layer perceptron back propagation	Functions	WEKA.classifiers.functions.multiLayerPerceptron

2.2 Data Collection

King Faisal Specialist Hospital and Research Centre, Jeddah branch, was the source from where the data was collected. Information gain index was used to classify the survivability attributes. Initial trials were made to get the data by mail. However, after much efforts and personal visits to more than one hospital and research center, King Faisal Hospital, Jeddah, agreed to grant 680 records of breast cancer diagnosed cases (See appendix D for official request). The coding system that's used for nominal attributes is the International Classification of Diseases for Oncology (ICDO).

2.3 Age at Diagnosis

The patient's age with highest incidence was 41 with 99 diagnosed cases and the highest mortality was at age 39with 78 diagnosed cases.

2.4 Primary Site

Primary site is the location of the tumor origination in the affected organ. (Following the International Classification of Oncology (IDC) 2007 Edition [13]

2.5 Differentiation Grade

Grade of tumor is a measure of the abnormality and level of the tumor by measuring the level of differentiation of the tumor cell with three main levels: well differentiated cells, moderately differentiated cells, and poorly differentiated cells

November-December 2019
ISSN: 0193-4120 Page No. 4039 - 4043

2.6 Performance Measurement

Three main performance measurements are used: accuracy, sensitivity and specificity. Sensitivity indicates the ratio of how many cases were truly classified as survived out of those who have not been classified as survived (true positive and false negative). Specificity indicates the ratio of how many cases were truly classified as not survived out of those who have not been classified as survived (true negative and false positive). Accuracy indicates the ratio of truly classified instances out of all instances (true positive, true negative, false positive and false negative). Sensitivity is referred to as the true positive rate (TPR) and specificity the true negative rate (TNR). Thus, the sum of the TPR and the TNR should equal 1.

2.7 Tools Used

The tools that were used t are:

1. WEKA Explorer: A module in WEKA with preprocessing and classification model building capabilities.

2. WEKA Experimenter A module in WEKA used to test different classifiers performance using statistical testing.

2.8 WEKA (Waikato Environment for Knowledge Analysis) Experiment

After the acquiring of the dataset from King Faisal Specialist Hospital, the data had to be preprocessed and then used to build the three models: Multi-layer Perceptron neural network (MLPNN), decision tree (C4.5) and Naïve Bayes. This process represents the crucial step of building the models in order to compare their performances with respect to accuracy, specificity and sensitivity. The first step was data preprocessing in which the data is cleaned and prepared appropriately in order for the algorithms to process the data correctly. The second step was feature selection, in which the attributes that contribute strongly in the classification are selected; this process requires the use of a search algorithm and an evaluation function. The third step is the processing step in which the data is fed into the three algorithms. The fourth step is the evaluation of the models using 10-fold cross validation with an iteration control of 10. This control assures that each algorithm is accessed 100 times (10 times for the folds* 10 times for the iteration control). The final step is the experiment validation using paired-T test with asignificance value of 0.05

3. RESULT AND DISCUSSION

The experiment results were verified using 10-fold cross validation. WEKA"s Experiment environment was customized to conduct the 10-fold cross validation on the collected data collected.

3.1 Decision Tree Classifier Performance

After the models were trained using an 80% split (i.e. 1084 instances of the dataset), they were tested on 272 instances (20%). The C4.5 classified 268 instances correctly and misclassified only 4 instances as shown in Table 2. The classifier had a 0.87(>0) Kappa agreement, which assures the classification accuracy was not obtained by chance. Kappa statistics have not shown to be accurate in indicating the level of accuracy of the classifier [30]. Kappa is often used as indicators that the classification agreements are higher than the agreement generated by chance

Table 2: Decision Tree classifier Performance

Correctly Classified Instances	268	98.5294
Incorrectly Classified Instances	4	1.4706 %
Kappa statistic	0.8745	
Mean absolute error	0.0234	
Root mean squared error	0.1173	
Relative absolute error	18.5419 %	
Root relative squared error	49.7304 %	
Total Number of Instances	272	

For each class; survive (0) and not survive (1), a collection of measurements are used to compare the accuracy of the classifiers with respect to the two different classes. These are the true positive rate, the false positive rate, the recall, the f- measure and the area under the receiver operator curve. For the survive class, the TP rate measures the number of instances that were correctly classified as survived out of all the instances that have a class label equivalent to survived. For the (not survive) class TP rate measures the number of instances that have (not survive) as class label that were correctly classified as not survive. The FP rate is a measurement of incorrectly classified instances out of the total number of instances that belong to that class. Table 3 shows Decision Tree Classifier Detailed Accuracy by Class. The accuracy, sensitive and specificity computed were 0.98, 0.99 and 0.83 respectively

November-December 2019
ISSN: 0193-4120 Page No. 4039 - 4043

Table 3: Decision Tree Classifier Detailed Accuracy by Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.988	0.063	0.996	0.988	0.992	0.966	0
	0.938	0.012	0.833	0.938	0.882	0.966	1
Weighted Average	0.985	0.06	0.986	0.985	0.986	0.966	

Table 5: MLP NN Detailed Accuracy by Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.98	0.063	0.996	0.98	0.988	0.97	0
	0.938	0.02	0.75	0.938	0.833	0.97	1
Weighted Average	0.978	0.06	0.982	0.978	0.979	0.97	

3.2 MLP Neural Network Classifier Performance

The multi-layer perceptron NN classifier classified 266 instances correctly and misclassified only 6 instances. The classifier had a 0.82 (>0) Kappa agreement, which assures the classification accuracy was not obtained by chance as shown in Table 4. The accuracy by class showed a ROC area of 0.97 for both classes of the survivability attribute. Table 5 is a summary of the accuracy of each class (0 and 1) with respect to: precision, recall, f-measure and ROC area. The MLP model had an accuracy of 0.98, sensitivity of 0.99 and specificity of 075.

Table 4: C4.5 Classifier Performance

Correctly Classified Instances	266	97.7941 %
Incorrectly Classified Instances	6	2.2059%
Kappa statistic	0.8217	
Mean absolute error	0.0283	
Root mean squared error	0.1396	
Relative absolute error	22.4394 %	
Root relative squared error	59.1876 %	
Total Number of Instances	272	

3.3 Naïve Bayes

The naïve bayes classifier classified 260 instances correctly and misclassified 12 instances. The classifier had a 0.47 (>0) Kappa agreement, which assures the classification accuracy was not obtained by chance as shown in Table 6. The accuracy by class showed a ROC area of 0.94 for both classes of the survivability attribute. Table 7 is summery of the accuracy each class (0 and 1) had with respect to: precision, recall, f-measure and ROC area.

Table 6: Naïve Bayes Classifier Performance

Correctly Classified Instances	260	95.5882 %
Incorrectly Classified Instances	12	4.4118 %
Kappa statistic	0.4796	
Mean absolute error	0.0866	
Root mean squared error	0.1862	
Relative absolute error	68.6591 %	
Root relative squared error	78.9181 %	
Total Number of Instances	272	

November-December 2019
ISSN: 0193-4120 Page No. 4039 - 4043

Table 7: Naïve Bayes Accuracy by Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.992	0.625	0.962	0.992	0.977	0.947	0
	0.375	0.008	0.75	0.375	0.5	0.947	1
Weighted Average	0.956	0.589	0.95	0.956	0.949	0.947	

Naïve bayes had an accuracy of 96 %, sensitivity of 96 % and specificity of 75 % which verifies the hypothesis that this statistical model is less accurate than the previous two models.

Table 8 is a tabular result of the average performance for the 10 folds of the cross-validations. It is represented by accuracy, sensitivity, specificity, mean error, kappa statistics and ROC area. Decision tree had the highest accuracy and sensitivity with values (0.979, 0.988) respectively. Naïve bayes had the highest classification error (0.094) and neural networks had the highest specificity (0.896).

Table 8: Comparing Models' Performance using 10-fold Cross Validation

	Neural Network	Decision Tree	Naïve Bayes
Accuracy	0.978	0.979	0.949
Sensitivity	0.984	0.988	0.950
Specificity	0.896	0.857	0.895
Mean Error	0.0319	0.024	0.094
Kappa Statistic	0.831	0.846	0.479
ROC Area	0.901	0.952	0.873

4. CONCLUSION

In an effort to construct several predictive models for breast cancer survivability rates and to conduct an empirical comparison between the performances of three main data mining algorithms, a data mining project was implemented and tested. Two of the models are machine learning algorithms: decision trees and neural networks and the third is Naive Bayes. The performance of the three predictive models was examined based on three main performance criteria: accuracy, sensitivity and specificity.The results showed that decision tree had the highest accuracy and sensitivity with values (0.979, 0.988) respectively. Naïve bayes had the highest classification error (0.094) and neural networks had the highest specificity (0.896).

5. REFERENCES

[1] Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 34(2), 113-127.

[2] Siegel, R. L., Miller, K. D., &Jemal, A. (2017). Cancer statistics, 2017. *CA: a cancer journal for clinicians*, *67*(1), 7-30.

[3] Al Diab, A., Qureshi, S., Al Saleh, K. A., Al Qahtani, F. H., Aleem, A., Alghamdi, M. A., ... & Qureshi, M. R. (2013). Review on breast cancer in the Kingdom of Saudi Arabia. *Middle-East J Sci Res*, *14*(4), 532-543.

[4] Elkum, N., Al-Tweigeri, T., Ajarim, D., Al-Zahrani, A., Amer, S. M. B., &Aboussekhra, A. (2014). Obesity is a significant risk factor for breast cancer in Arab women. *BMC cancer*, *14*(1), 788.

[5] Belciug, S., & El-Darzi, E. (2010, July). A partially connected neural network-based approach with application to breast cancer detection and recurrence. In *Intelligent Systems (IS), 2010 5th IEEE International Conference* (pp. 191-196). IEEE.

[6] Han, J., Pei, J., &Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

[7] Rokach, L., &Maimon, O. Z. (2008). *Data mining with decision trees: theory and applications* (Vol. 69). World scientific.

[8] Rastogi, S. C., Rastogi, P., &Mendiratta, N. (2008). Bioinformatics Methods And Applications: Genomics Proteomics And Drug Discovery 3Rd Ed. PHI Learning Pvt. Ltd..

[9] Sarvestani, A. S., Safavi, A. A., Parandeh, N. M., &Salehi, M. (2010, October). Predicting Breast Cancer Survivability using data mining techniques. In *Software technology and Engineering (ICSTE), 2010 2nd international Conference on*(Vol. 2, pp. V2-227). IEEE.

[10] Epstein, I. (2009). Clinical data-mining: Integrating practice and research. Oxford University Press.

[11] Dunham, M. H. (2006). *Data mining: Introductory and advanced topics*. Pearson Education India.

[12] Hand, D. J. (2007). Principles of data mining. *Drug safety*, *30*(7), 621-622.