RESEARCH ARTICLE

Automatic Classification on Bio Medical Prognosisof Invasive Breast Cancer

Sountharrajan S[1], Karthiga M[1], Suganya E[1], Rajan C[2]*

Abstract

Breast Cancer one of the appalling diseases among the middle-aged women and it is a foremost threatening death possibility cancer in women throughout the world. Earlier prognosis and preclusion reduces the conceivability of death. The proposed system beseech various data mining techniques together with a real-time input data from a biosensor device to determine the disease development proportion. Surface acoustic waves (SAW) biosensor empowers a label-free, worthwhile and straight detection of HER-2/neu cancer biomarker. The output from the biosensor is fed into the proposed system as an input along with data collected from Winconsin dataset. The complete dataset are processed using data mining classification algorithms to predict the accuracy. The exactness of the proposed model is improved by ranking attributes by Ranker algorithm. The results of the proposed model are highly gifted with an accuracy of 79.25% with SVM classifier and an ROC area of 0.754 which is better than other existing systems. The results are used in designing the proper drug thereby improving the survivability of the patients.

Keywords: Support vector machine- receiver operating curve- surface acoustic wave- human epidermal growth receptor

Asian Pac J Cancer Prev, **18 (9)**, 2541-2544

Introduction

Cancer, one of the appalling diseases in the world. Though lot of treatment facilities available for cancer these days the survival rate is very poor. The reason behind poor survival is identifying cancer only at their final stage. Breast Cancer is one of the appalling diseases among women. The influence of the disease depends upon the type of cancer, age of the patient and the disease extent. Worldwide breast cancer is the most common means of cancer among 25% of women. The diagnosis techniques are examining the breast physically by a health care specialist, biopsy in the tumor region and mammography. The various aftercare techniques of breast cancer include manual surgery of the affected part, chemotherapeutic treatments and radiotherapy in which surgery is always the best and first option. There is an urgent need of diagnosing and detecting the disease in advance effectively and accurately. The various genetic factors and environmental factors have to be considered while developing a novel method for cancer diagnosis. Therefore data mining techniques like clustering and decision tree are combined for building a multilayered novel platform for breast cancer risk prediction system.

The proposed system helps to identify the risk of breast cancer among the patients in advance using data mining techniques. The predicted results are compared with patient's prior medical data for validation. A biosensor device is used to detect the breast cancer biomarkers more accurately. This biosensor data is one of an input for the proposed system. The main aim of the proposed system is to provide advance warning to the patients and diagnosing the treatment cost effectively. This system helps in detection of a person's predisposition for cancer before going for clinical and lab tests which is cost and time consuming. The machine learning algorithms are used to learn and classify cancer. The classifier is tested with training examples and the accuracy of predicting the disease is evaluated.

In recent year's lot of new techniques have been proposed to develop Point-of-Care devices to detect cancer growth in advance. The future of healthcare industry is transformed by Point-of-care technologies (POCT) (Joshi et al., 2016). College of American Pathologists (COA) defined "POCT is a means of testing which go about near or at the site of a patient which results in a big revolution in patients care". With the remarkable growth and advancements in POCT, the future of health care industry is imagined as patients diagnosing their disease themselves based on the corresponding bio-marker using a small and handy biosensor kit. Mohanty et al., (2014) proposed a fabrication method and usage of silicon nanochannel FET devices as biosensors for diagnosing and monitoring the breast cancer. For successful treatment the disease has to be diagnosed earlier. Identifying the appropriate disease biomarker which differentiates healthy and affected states

*[1]Department of Computer Science and Engineering, Bannari Amman Institute of Technology, [2]Department of Information Technology, K S Rangasamy College of Technology, India. *For Correspondence: rajancsg@gmail.com*

in humans is a burgeoning research field.

Chandra (2013) from IIT have proposed a biosensor for detecting the breast cancer biomarker namely Human Epidermal Growth Receptor (HER2) by using electrochemical immunosensor combined with hydrazine and aptamer-conjugated gold nanoparticles (AuNPs). Similarly, Malhotra (2016) from Delhi Technological University have discussed about the prospects of CP based biosensors for cancer biomarker detection which was proved to be flexible, cost effective, light weight and can be easily disposed. Vikas et al., (2014) et al proposed a novel approach for detecting the cancer using data mining techniques. The results are evaluated by comparing with three different classification algorithms like IBK, BF Tree and Sequential Minimal Optimization (SMP). He also suggested the most important attributes for breast cancer survival. Goreti et al., (2015) have proposed an effective toolkit for diagnosing the breast cancer disease by intending a decision support system for identifying the correct and important features on mammograms. This toolkit serves as a learning platform for practitioners and students for better understanding of the disease.

Materials and Methods

The dataset used in our proposed work is collected from Wisconsin Repository. Totally 198 findings with two probable results (Re-Recur (151 distribution) and Nr-Non Recur (47 distributions)) are there. 30 features describes about the uniqueness of nucleus cell of the cancer tissue. Using that ten features are calculated for each cell-nuclei. Some of the calculated features are radius, softness, mean, concavity, symmetry, perimeter, firmness, fractal dimension and contour's concave portions. Then for each feature mean, standard deviation and comparative mean value (best mean value for four attributes) are computed. Thus 30 features are obtained as result. The other features considered are ID, Outcome result (Re, Ne), size of the tumour, lymph node appearance. Totally 35 features are collected along with the result from biosensor. The dataset is divided into two subsets: training set and test set. In training set there are 65% and in test set there are 34% of our dataset. Both training set and test set are independent to each other. The following Figure-1 illustrates the proposed system.

The biosensor taken was designed to detect the HER-2 protein which is a major biomarker for breast cancer. SAW (Surface Acoustic Wave) biosensor which was developed in real time by Gruhul et al., (2010) is used in our proposed system. Breast serum is given as input to the biosensor. The target molecules binding in the biosensor are determined by measuring the variation in velocity of the surface wave. This helps in direct detection of HER-2 protein which is a major biomarker for breast cancer. The output from the biosensor is also fed into the proposed system as an input and it decode, transform and format the data for proper analysis. The biosensor data along with other attributes are processed and stored in the database as Intelligent data. The Data Cleaning and Pre-processing Module automatically processes the data stored in the database in order to correct and remove any inconsistencies between the data records. After this operation, the data are ready for analysis by the IDA (Intelligence Data Analysis) databases. The Intelligent Analysis Module is the core of the system. Its job is to interpret the sensor and test data of the database using the system expert knowledge.

For creating a training dataset, our proposed system uses three data mining classification algorithms, SVM, C4.5 Decision Tree and Naïve Bayes. Weka tool has been used for classification. The smart performance of the classifier to retrieve the intelligent training data is increased by using 10 fold cross validation methodology for each classification algorithms. In this cross validation methodology, our original data samples are partitioned erratically into k equal size sub samples. Out of the k subsamples, 1 subsample is used as a validation data for testing and the remaining k-1 samples are used for training.

The accuracy of the proposed model is increased by selecting the best feature and by removing the irrelevant attributes and redundant attributes. Ranker algorithm is used to select the good feature among the 35 features available. For Decision Tree method and Naïve Bayes' method the attribute assessor chosen is InfoGainEval. The attribute accessor chosen for SVM is SVMAttributeEval which evaluates the attribute worth for using in a classifier. Attributes are ranked in order to overcome the problem of choosing attributes for multi-class problem. High ranked attributes like lymph node growth, softness etc. plays a vital role in controlling the accuracy of the disease prediction whereas low-ranked attributes like id, concavity etc. contribute lesser in disease prediction.

Results

Accuracy for three classification algorithms is calculated first without applying any ranking to the attributes. The accuracy obtained is 76% for SVM and 72% for C4.5 and 68% for Naïve Bayes classification Algorithm. This shows that classifying the attributes without any feature selection will provide approximate results. Among the other classifiers SVM produces the

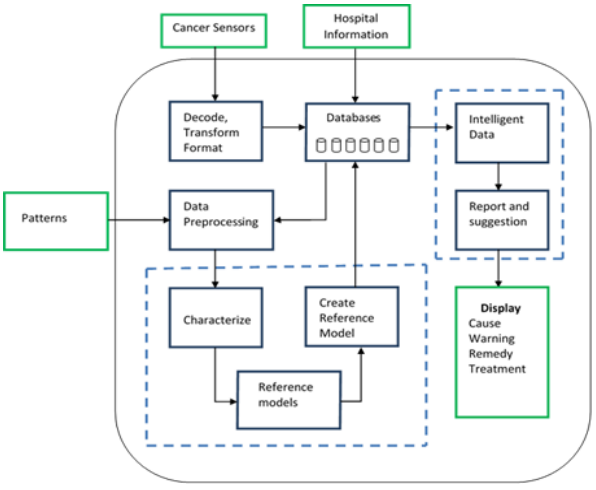

Figure 1. Representation of Proposed Model to Predict the Breast Cancer

Table 1. Classification Accuracy of 3 Classifiers

	SVM		Naïve Bayes		C4.5 Decision Tree	
Criteria for Evaluation	With Ranker	With Ranker	With Ranker	With Ranker	With Ranker	With Ranker
Time (in sec)	23.4	0.0955	24.7	0.0855	24.6	0.045
Instannces that are correctly classified	154	150	152	147	152	134
Instannces that are in correctly classified	45	48	47	53	47	66
Classification Accuracy	79.25%	76%	77.25%	68%	77.25%	72%

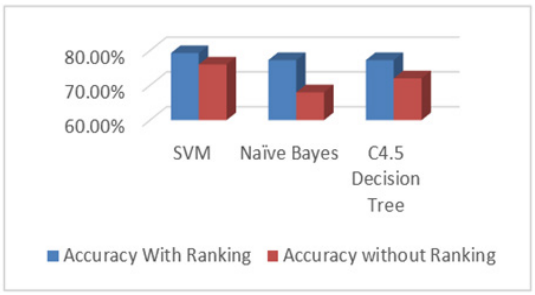

Figure 2. Comparative Study of the Classification Accuracy

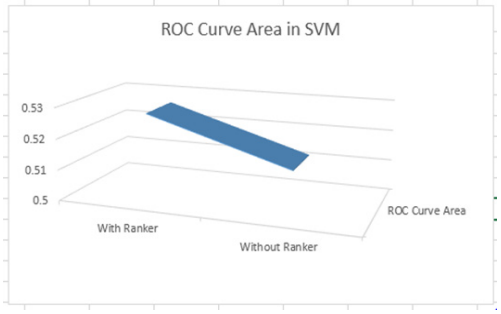

Figure 3. Comparative Result of ROC in SVM Classifier (With and Without Ranking of Attributes)

best result.

The accuracy of the result is improved by selecting the attributes more appropriately and by removing the redundant attributes. So, Ranker algorithm is used to rank the attributes and N best chosen attributes is given as input to each classifiers. The aim of ranking the attributes is to choose best attributes to maximize the classification accuracy. From the 35 attributes, 11 best attributes are chosen and remaining are considered as less valuable. Classification with best 11 attributes are performed in SVM and the classification accuracy gained is 79.25%. For Naïve Bayes and C4.5 Decision tree the accuracy improved by 9.2%. For C4.5 classifier 10 best attributes are provided as input whereas for Naïve Bayes 8 best attributes are provided. The following Table 1 shows the classification accuracy obtained from three different classifiers.

Undoubtedly the maximum classification accuracy obtained is through SVM classifier with and without ranking. The Figure 2 depicts the comparative study of the classification accuracy for three classifiers. Different error rates are also reduced while applying ranking to the attributes. For SVM all error rates are improved unfailingly and for C4.5 Decision Tree Root mean squared error rate and relative absolute error rate is improved. For Naïve Bayes all error rates showed improvement. The Table 2 shows the error rate for all three classifiers. Finally, true positive and true negative rates are calculated

Table 2. Different Error Rates Obtained Using Three Classifiers

Evaluation Criteria	SVM		Naïve Bayes		C4.5 Decision Tree	
	Ranking	Without Raking	Ranking	Without Raking	Ranking	Without Raking
Mean Absolute Error Rate (MAER)	0.221	0.243	0.323	0.348	0.0342	0.0252
Relative Absolute Error Rate (RAER)	63.70%	66.30%	87%	91.50%	99.50%	82%
Root Relative Squared Error Rate (RRSE)	84%	116.40%	102.20%	123.10%	99.98%	114.10%
Root Mean Squared Error Rate (RMSE)	0.432	0.423	0.439	0.525	0.432	0.493

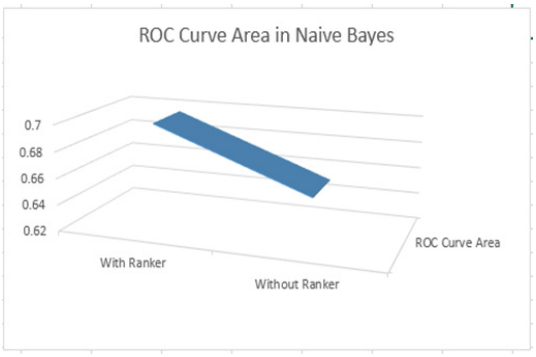

Figure 4. Comparative Result of ROC in Naive Bayes Classifier (With and Without Ranking of Attributes)

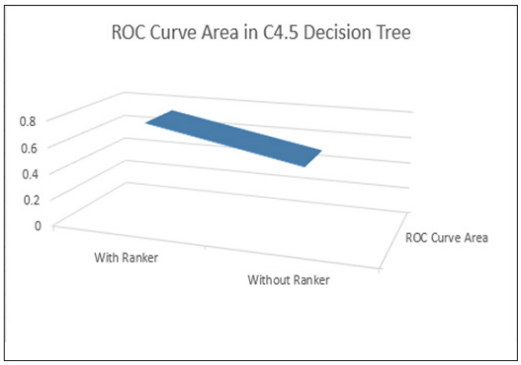

Figure 5. Comparative Result of ROC in C4.5 Decision Tree Classifier (With and without Ranking of Attributes)

Table 3. ROC Curve for Three Classifiers

	SVM		Naïve Bayes		C4.5 Decision Tree	
Criteria for Evaluation	With Ranker	With Ranker	With Ranker	With Ranker	With Ranker	With Ranker
True Positive Value	0.734	0.728	0.863	0.772	0.774	0.736
True Negative Value	0.728	0.71	0.453	0.442	0.752	0.506
ROC Curve Area	0.527	0.513	0.698	0.654	0.754	0.533

for all the three classifiers. Then a ROC curve is generated. ROC stands for Receiver Operating Curve and it is drawn against TP and TN rate. ROC is a diagnostic tool for predicting the disease accurately. The ROC curve is very meagre when it is plotted without ranking of the attributes.

After ranking the attributes the ROC curve holds good for C4.5 Decision Tree classifier. For Naïve Bayes and SVM, the ROC curve is not significantly good. Thus plotting of ROC curve reveals the importance of ranking the best features. Figure 3,4,5 clarifies the comparative results of the three classifiers.

Discussion

In conclusion, in this proposed system earlier prognosis of breast cancer to improve the survivability of patients is emphasized. Data collected from Wisconsin dataset along with biosensor output are stored in database and processed to produce Intelligent Data. Real-time output from a biosensor device improves the accuracy of disease prediction. The best possible features are chosen and fed into the three different types of classifiers like C4.5 Decision tree, Naïve Bayes and Support Vector Machine (SVM) to achieve a comparative accuracy result. Among the three classifiers the accuracy obtained from SVM is fairly better. A ROC Curve area holds better for C4.5 Decision Tree Classifier. Thus the trained intelligent data obtained from the proposed model is used by the drug designers to improve the treatment of the breast cancer patients.

References

Bellaachia A, Erhan G (2005). Predicting breast cancer survivability using data mining techniques. *Age*, **58**, 10-110.

Chandra P, Suman P, Mukherjee M, et al (2013). HER2 protein biomarker based sensor systems for breast cancer diagnosis. *J Mol Biomark Diagn*, ISSN-2155-9929.

Chaurasia Vikas, Saurabh Pal (2014). A novel approach for breast cancer detection using data mining techniques. *IJIRCCE*, **2.1**, 2456-65.

Delen D, Walker G, Kadam A (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artif Intell Med*, **34**, 113-27.

Dong-Sheng Cao, Xu QS, Liang YZ, et al (2010). Automatic feature subset selection for decision tree-based ensemble methods in the prediction of bioactivity. *Chemometr Intell Lab*, **103**, 129-36.

Geetha K, Rajan C (2016). Automatic colorectal polyp detection in colonoscopy video frames. *Asian Pac J Cancer Prev*, **17**, 4869-73.

Geetha K, Rajan C (2017). Heuristic classifier for observe accuracy of cancer polyp using video capsule endoscopy. *Asian Pac J Cancer Prev*, **18**, 1681-8.

Goreti M, Alberto F (2015). Using data mining techniques to support breast cancer diagnosis. New contributions in information systems and technologies. *Adv Intel Syst Comput*, **353**, 689-700

Gruhul JF, Michael Rapp, Kerstin Lange (2010). Label-free detection of breast cancer marker HER-2/neu with anacoustic biosensor. *Proc Eurosensors*, **5**, 914-17.

Huan Liu, Lei Yu (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE T Knowl Data En*, **17.4**, 491-502.

Karim Khani Zand H (2015). A comparative survey on data mining techniques for breast cancer diagnosis and prediction. Indian *J Fund Appl Life Sci*, **5**, 4330-9.

Jahanvi J, Rinal D, Jigar P (2014). Diagnosis of breast cancer using clustering data mining approach. *Int J Comput Appl*, **101**, 10.

Jianping Z (1992). Selecting typical instances in instance based learning. Proceedings of the ninth international conference on machine learning. ACM Digital library, pp 470-9.

Joshi PN (2016). Nano-biosensors: Point of care devices for personalized cancer diagnosis. *J Chem*, e-ISSN: 2319-9849.

Liu YQ, Wang C, Zhang Lu (2009). Decision tree based predictive models for breast cancer survivability on imbalanced data. 2009 3rd international conference on bioinformatics and biomedical engineering, pp 1-4.

Malhotra DB, Saurabh K, Pandey CM (2016). Nanomaterials based biosensors for cancer biomarker detection. IOP Science. Conference Series 704, pp 11-18.

Mohanty P, Chen Y, Wang X, et al (2014). Field effect transistor nanosensor for breast cancer diagnostics. *Int J Mol Biol Cancer Diag*, e-ISSN, 1401-1168

Padmapriya B, Velmurugan T (2014). A survey on breast cancer analysis using data mining techniques. 2014 IEEE International conference computational intelligence and computing research (ICCIC), pp 1-4.

Pan W (2009). Application of decision tree to identify abnormal high frequency electro-cardiograph. *Physics Experiment*, **11**, 011.

Vanaja S, Ramesh Kumar K (2014). Analysis of feature selection algorithms on classification: a survey. *Int J Comput Appl*, **96**, 17.

Vikas C, Saurabh P (2014). A novel approach for breast cancer detection using data mining techniques. *IJIRCCE*, ISSN online (2320-9801) .

Wolberg WH, Mangasarian OL (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proc Natl Acad Sci U S A*, **87.23**, 9193-6.

Yuvarani S, Jothi V (2015). Breast cancer detection in data mining: A review. *Int J Comput Appl*, **7**, 45-8.