International Journal of Bioinformatics and Biomedical Engineering
Vol. 1, No. 3, 2015, pp. 318-322
http://www.aiscience.org/journal/ijbbe
ISSN: 2381-7399 (Print); ISSN: 2381-7402 (Online)

Diagnosis of Breast Cancer Using Ensemble of Data Mining Classification Methods

Gokhan Zorluoglu[1, *], Mustafa Agaoglu[2]

[1]Bioengineering Department, Marmara University, Kadikoy, Istanbul, Turkey
[2]Computer Engineering Department, Marmara University, Kadikoy, Istanbul, Turkey

Abstract

Breast cancer is a very serious malignant tumor originating from the breast cells. The disease occurs generally in women, but also men can rarely have it. During the prognosis of breast cancer, abnormal growth of cells in breast takes place and this growth can be in two types which are benign (non-cancerous) and malignant (cancerous). In this study, the aim is to diagnose the breast cancer using various intelligent techniques including Decision Trees (DT), Support Vector Machines (SVM), Artificial Neural Network (ANN) and also the ensemble of these techniques. Experimental studies were done using SPSS Clementine software and the results show that the ensemble model is better than the individual models according to the evaluation metric which is the accuracy. In order to increase the efficiency of the models, feature selection technique is applied. Moreover, models are also analyzed in terms of other error measures like sensitivity and specificity.

Keywords

Artificial Neural Network, Breast Cancer, Cross Validation, C5.0, Data Mining, Decision Tree, Support Vector Machine

Received: September 15, 2015 / Accepted: December 5, 2015 / Published online: December 29, 2015

1. Introduction

Nowadays, computer science and medical area are nested in order to provide proper prognosis or diagnosis of the human diseases. Many computational techniques are used for the identification of the health problems. Data mining has turned into a crucial procedure for registering applications in the space region of medicine. In this study, it is aimed to identify the breast cancer with the help of data mining classification methods. The dataset named Wisconsin Diagnostic Breast Cancer Database (WDBC) is obtained from Wisconsin Madison University [1, 2]. The classification techniques used on WDBC are Decision Trees (DT), Support Vector Machines (SVM), Artificial Neural Network (ANN) and also the ensemble of them. SPSS Clementine software was used for the experimental studies. The models used in Clementine are support vector machine model, C5.0 model and neural net model. Furthermore, feature selection algorithm is used in order to reduce the dimensionality of the dataset. In order to measure the performance, 10-fold cross validation technique is used on WDBC dataset. That is, the data are partitioned by the ratio 90:10% for training and testing. This is done ten times by adifferent10% being tested each time.

The paper organizes as follows; in the next section, related works are indicated. Classification methods section describes the classification methods used for the modelling and the next section named experiments and model development presents the model development, the preprocessing steps including details of the datasets and feature selection. The results and the performance evaluations are discussed in discussion and results section. Finally, the last section introduces the conclusion of this study.

* Corresponding author
E-mail address: gokhanzorluoglu@marun.edu.tr (G. Zorluoglu), agaoglu@marmara.edu.tr (M. Agaoglu)

2. Breast Cancer Overview

Consistently, cells in your body divide grow and die on in an organized way. Breast tumor is an infection where cells in the breast tissue develop and isolate without ordinary control. Breast cancer is the most widely recognized disease among women, but men can get breast cancer, too. Around the world, it is the most widely recognized type of tumor in females that is influencing roughly 10% of all ladies at some phase of their life. [3]. Almost all of women may survive for a long time, with an early detection of cancer. In this study, we try to develop a computational approach using several data mining techniques and also ensemble of these methods for the diagnosis of this leading cancer for women.

3. Related Works

There are many studies [2, 4, 5, 6, 7] done for the diagnosis or the prognosis of the breast cancer using University of California Irvine (UCI) dataset called Wisconsin Diagnostic or Prognostic Breast Cancer Databases [9] (WDBC or WPBC). The frequently used method in these studies is a linear programming-based classification method which is called MSM-Tree (MSM-T) [2, 9, 10].

While Olvi et al. (1995) and Wolberg et al. used only the MSM-Tree technique for the classification in their studies, Wolberg et al. in1995 also used the Logistic Regression [7] technique. For the estimation of predictive accuracy, in all these studies10-fold cross validation is chosen; however, Wolberg et al. (1995a) used leave-one-out cross validation for the prognosis of the breast cancer. All these relevant studies get the same accuracy which is 97.5%with the help of MSM-T; whereas, Wolberg et al. (1995c) gets the accuracy of 96.2% using the Logistic Regression technique.

Nowadays, it is still being challenged on data mining classification techniques applied for breast cancer diagnosis or prognosis. For instance, a framework for diagnosis and prognosis of tumor utilizing FP (frequent pattern mining)-growth algorithm is introduced and Decision Tree algorithm to anticipate the likelihood of disease in setting to age is utilized in [11]. Also this study [11] shows that when contrasted with apriori algorithm, fp-growth is more effective as time required to execute is not exactly apriori and likewise memory use is less in fp-growth. So they find that fp-growth is more proficient to utilize. Furthermore, Sumbaly et al. (2014) have picked J48 Decision Tree algorithm [13] to build up the model of the study called Diagnosis of Breast Cancer using Decision Tree Data Mining Technique. In this study [12] 10-fold cross validation [14] is utilized to specify test and training information. The J48 algorithm is utilized on the dataset utilizing WEKA (Java Toolbox for different information mining method) [15] after data pre-processing (CSV group). As a result, they gain an accuracy of 94.5637% with the performance of J48 Decision Tree. The last but not the least, Gupta et al. (2011) gives an investigation of different review and technical papers on breast cancer prognosis and diagnosis issues and investigates that data mining methods offer extraordinary guarantee to reveal examples covered up in the information that can help the clinicians in making of decision.

4. Classification Methods

The classification models of Clementine used in this study are C5.0, SVM and Neural Net which are briefly described below. The ensemble model will be explained in Section 4.

4.1. C5.0

Decision Tree (DT) is one of the supervised learning methods used generally for classification and also regression in a tree structure form. The aim is to construct this tree structure that predicts the label of a target variable by using the created model. The C5.0 is one of the rule induction algorithms of Clementine in order to generate a decision tree. It allows you to view the rules in two different formats which are the decision tree presentation and the rule set presentation [17].

4.2. SVM

Another classification method used in this study is the Support Vector Machine which is a supervised learning model with associated learning algorithms.

It attempts to classify out comes by mapping data to a higher-dimensional feature space so that data points can be categorized [17].

4.3. Neural Net

The third classifier of the study is the Neural Net which is the artificial neural network model of the Clementine. A typical neural network consists of several connected neurons arranged in layers to create networks. The connections between the neurons provide the network's having the ability to learn patterns and interrelationships in data [17].

5. Experiments and Model Development

In this section; the development of the model, the pre-processing steps including details of the datasets and feature selection are described.

5.1. Dataset

The Wisconsin Diagnostic Breast Cancer dataset from the

UCI Machine Learning Repository [18, 19] is used in order to determine the input tuple saying that tumor is benign or malignant. In Table 1, description of the WDBC data set is shown.

Table 1. Description of the Dataset.

Dataset Characteristics:	**Multivariate**	**Number of Attributes:**	**32**
Attribute Characteristics:	Real	Number of Instances:	569
Associated Tasks:	Classification	Number of Classes:	2

The first feature is the ID number and the second one is the class label which is the diagnosis (B = benign or M=malignant). Other 30 attributes are the mean, standard error and the worst value of the radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and the fractional dimension for each cell nucleus.

5.2. Evaluation Metrics

The effectiveness of the models is measured by the three well-known evaluation metrics which are shown in equations (1, 2 and 3). These given equations are calculated using the true positive (TP), false positive (FP), true negative (TN) and false negative (FN) cases.

$$\text{Accuracy\%} = (TP+TN)/(P+N)*100 \quad (1)$$

$$\text{Specificity\%} = TN/(TN+FP)*100 \quad (2)$$

$$\text{Sensitivity\%} = TP/(TP+FN)*100 \quad (3)$$

These metrics are calculated after obtaining the confusion matrix which includes the values of TP, FP, TN and FN cases.

5.3. Data Preprocessing

Before the modelling phase, some data preprocesses were done in Clementine. Initially, the data was filtered by excluding the first attribute including ID number. Then, type of the class label was changed into flag type. Finally, feature selection model is applied to the dataset. As a result of the feature selection technique, unimportant features having the value (calculated in feature selection model) less than 0.9 are extracted.

5.4. Model Development

Each model is applied to the WDBC data set after the preprocessing steps and the evaluation metrics are analyzed. Moreover, the ensemble model which is the combination of SVM, C5.0 and Neural Net are generated and analyzed. The Figure 2 shows the representation of combining these three techniques in order to develop the ensemble model. The model of the study is shown below:

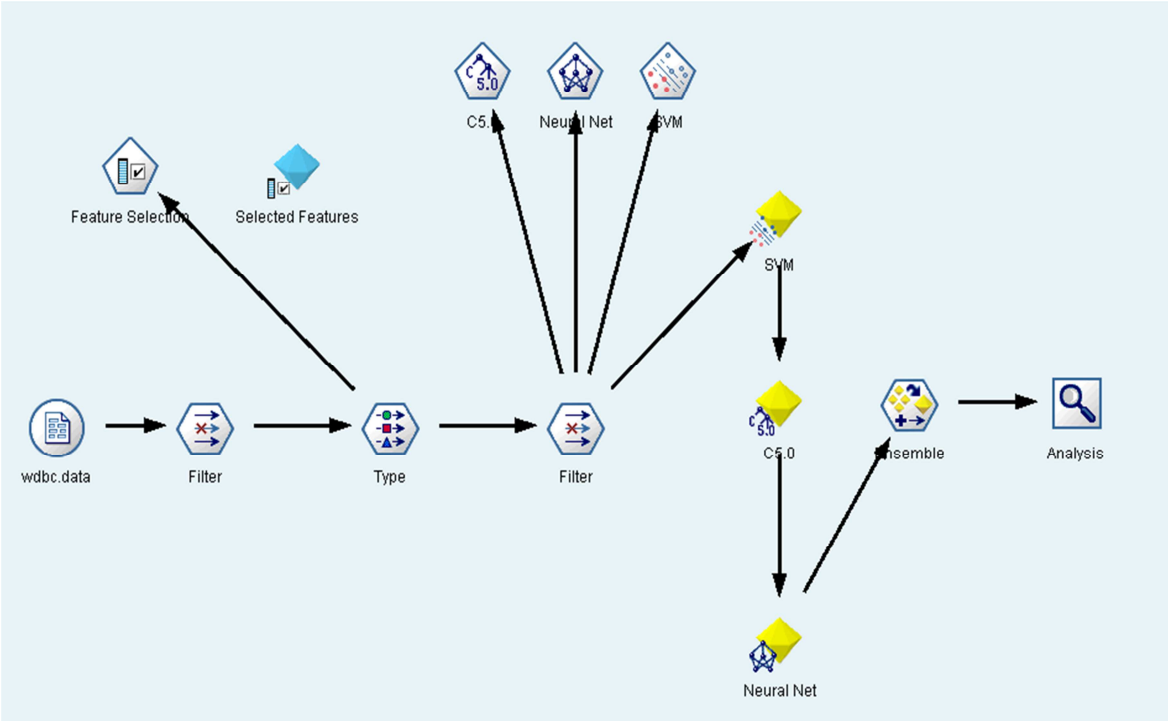

Fig. 1. The Ensemble Model Developed using SPSS Clementine Data Mining Tool.

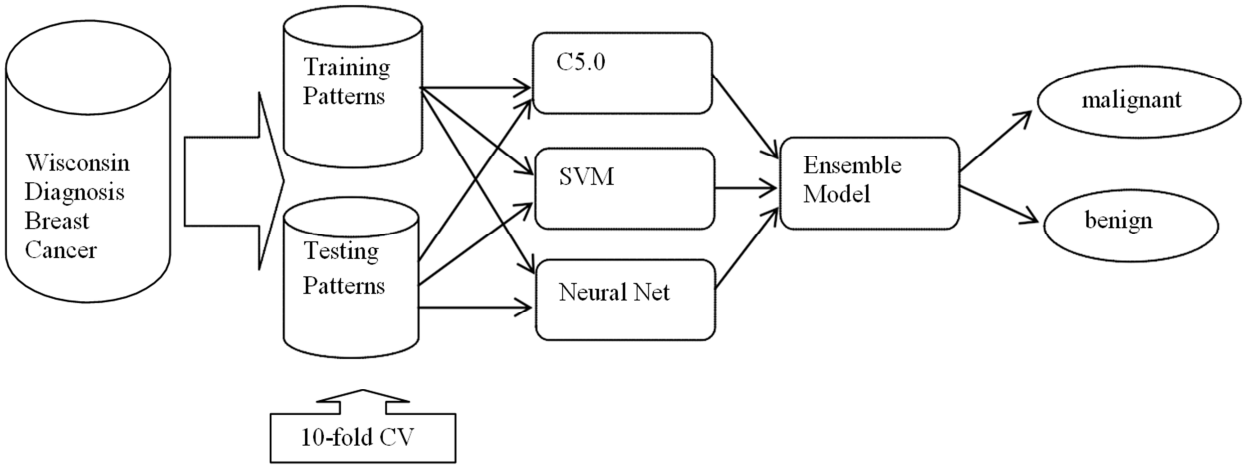

Fig. 2. Representation of Ensemble Mode l for the Breast Cancer Diagnosis.

The results of the experiments are expressed in the next section. According to the analysis it is proved that the ensemble model for the diagnosis of the breast cancer gives the best result.

6. Discussion and Results

Experimental studies were carried out with the help of SPSS Clementine software and the results are shown below:

Table 2. Evaluation Metrics of Each Model.

	Accuracy	Sensitivity	Specificity
SVM	98.07%	97.79%	98.55%
C5.0	98.07%	97.01%	100%
Neural Net	97.54%	96.73%	98.52%
Ensemble	98.77%	98.05%	100%

The results revealed that the measures of the ensemble model are more accurate than the individual models. Anyway, SVM, C5.0 and Neural Net models are also quite distinctive for the diagnosis of the breast cancer. Ensemble method has the highest accuracy which is 98.77%. Other individual classifiers SVM, C5.0 and Neural Net have the accuracy of 98.07%, 98.07% and 97.54%, respectively.

7. Conclusion

Diagnosis or prognosis of any serious disease such as breast cancer is a very challenging problem and it requires many preprocesses experiments and significant dataset. In this study, in order to identify the breast cancer, three different intelligent machine learning techniques which are SVM, DT (C5.0) and ANN (Neural Net) are used with the help of SPSS Clementine software. Besides, these three techniques are combined for creating the ensemble model to enhance the effectiveness of the models. Therefore, the ensemble model is the best classifier which gives 98.77% accuracy, 98.05% sensitivity and 100% specificity.

References

[1] Makinacı, M., Güneşer, C. (no date). Göğüs Kanseri Verilerinin Sınıflandırılması.

[2] Mangasarian, O. L., Street, W. N., Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), pp. 570-577.

[3] Kharya, S. (2012). Using Data Mining Techniques for Diagnosis And Prognosis of Cancer Disease, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol. 2, No. 2。

[4] Wolberg, W. H., Street, W. N, Mangasarian, O. L. (1994). Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. Cancer Letters vol. 77, 163-171.

[5] Wolberg, W. H., Street, W. N., Mangasarian, O. L. (1995a). Image analysis and machine learning applied to breast cancer diagnosis and prognosis. Analytical and Quantitative Cytology and Histology, Vol. 17 No. 2, pp. 77-87.

[6] Wolberg, W. H., Street, W. N., Heisey, D. M., Mangasarian, O. L. (1995b). Computerized breast cancer diagnosis and prognosis from fine needle aspirates. Archives of Surgery; 130:511-516.

[7] Wolberg, W. H., Street, W. N., Heisey, D. M., Mangasarian, O. L. (1995c) Computer-derived nuclear features distinguish malignant from benign breast cytology. Human Pathology, 26: 792-796.

[8] Bache, K., Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[9] Bennett, K. P. (1992). Decision tree construction via linear programming. In Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society Conference, pp. 97-101.

[10] Bennett, K. P., Mangasarian, O. L (1992). Robust linear programming discrimination of two linearly inseparable sets. Optimization Methods and Software, 1: 23-34.

[11] Majali, J., Niranjan, R., Phatak, V. and Tadakhe, O. (2014). Data Mining Techniques For Diagnosis And Prognosis of Breast Cancer. International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 5 (5), 2014, 6487-6490.

[12] Sumbaly, R., N. Vishnusri, S. Jeyalatha (2014). Diagnosis of Breast Cancer using Decision Tree Data Mining Technique. International Journal of Computer Applications (0975 – 8887) Volume 98 – No. 10.

[13] Bhargava, N., Sharma, G., Bhargava, R. and Mathuria, M. (2013). Decision Tree Analysis on J48 Algorithm for Data Mining. Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6.

[14] H. Blockeel and J. Struyf (2001). Efficient algorithms for decision tree cross-validation. Proceedings of the Eighteenth International Conference on Machine Learning (C. Brodley and A. Danyluk, eds.), Morgan Kaufmann, pp. 11-18.

[15] Feld, M., Kipp, M., Ndiaye, A. and Heckmann, D. (no date). Weka: Practical machine learning tools and techniques with Java implementations.

[16] Gupta, S., Kumar, D. and Sharma, A. (2011). Data Mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis. Indian Journal of Computer Science and Engineering (IJCSE), Vol. 2 No. 2.

[17] Nicholas, E. (2008). Introduction to Clementine and Data Mining. Brigham Young University.

[18] Salama, G. I., Abdelhalim, M. B., Zeid, M. A. (2012). Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers. International Journal of Computer and Information Technology (2277 – 0764), Volume 01– Issue 01.

[19] Frank, A., Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.