Diagnosis of Breast Cancer Using Ensemble of Data Mining Classification Methods

GOKHAN ZORLUOGLU[1], MUSTAFA AGAOGLU[2]
Bioengineering Department[1], Computer Engineering Department[2]
Marmara University
Goztepe Campus 34722 / Kadikoy - Istanbul
TURKEY
[1]gokhanzorluoglu@marun.edu.tr, [2]agaoglu@marmara.edu.tr

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.

Key-Words: - Artificial Neural Network, breast cancer, cross validation, C5.0, data mining, decision tree, support vector machine

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. 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 a different 10% 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 modeling 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.

2 Related Works

There are many studies [2, 3, 4, 5, 6] 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 [8] (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, 8, 9].

While Olvi et.al (1995) and Wolberg et.al used only the MSM-Tree technique for the classification in their studies, Wolberg et.al in 1995 also used the Logistic Regression [6] technique. For the estimation of predictive accuracy, in all these studies 10-fold cross validation is chosen; however, Wolberget. 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, Wolberget. al. (1995c) gets the accuracy of 96.2% using the Logistic Regression technique.

2 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.

2.1 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 [10].

2.2 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 outcomes by mapping data to a higher-dimensional feature space so that data points can be categorized [10].

2.3 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 [10].

3 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.

3.1 Dataset

The Wisconsin Diagnostic Breast Cancer dataset from the UCI Machine Learning Repository [11, 12] is used in order to determine the input tuple saying that tumor is benign or malignant. In Table 1, description of the WDBC dataset 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.

3.2 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.

3.3 Data Preprocessing

Before the modeling 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.

3.4 Model Development

Each model is applied to the WDBC dataset 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:

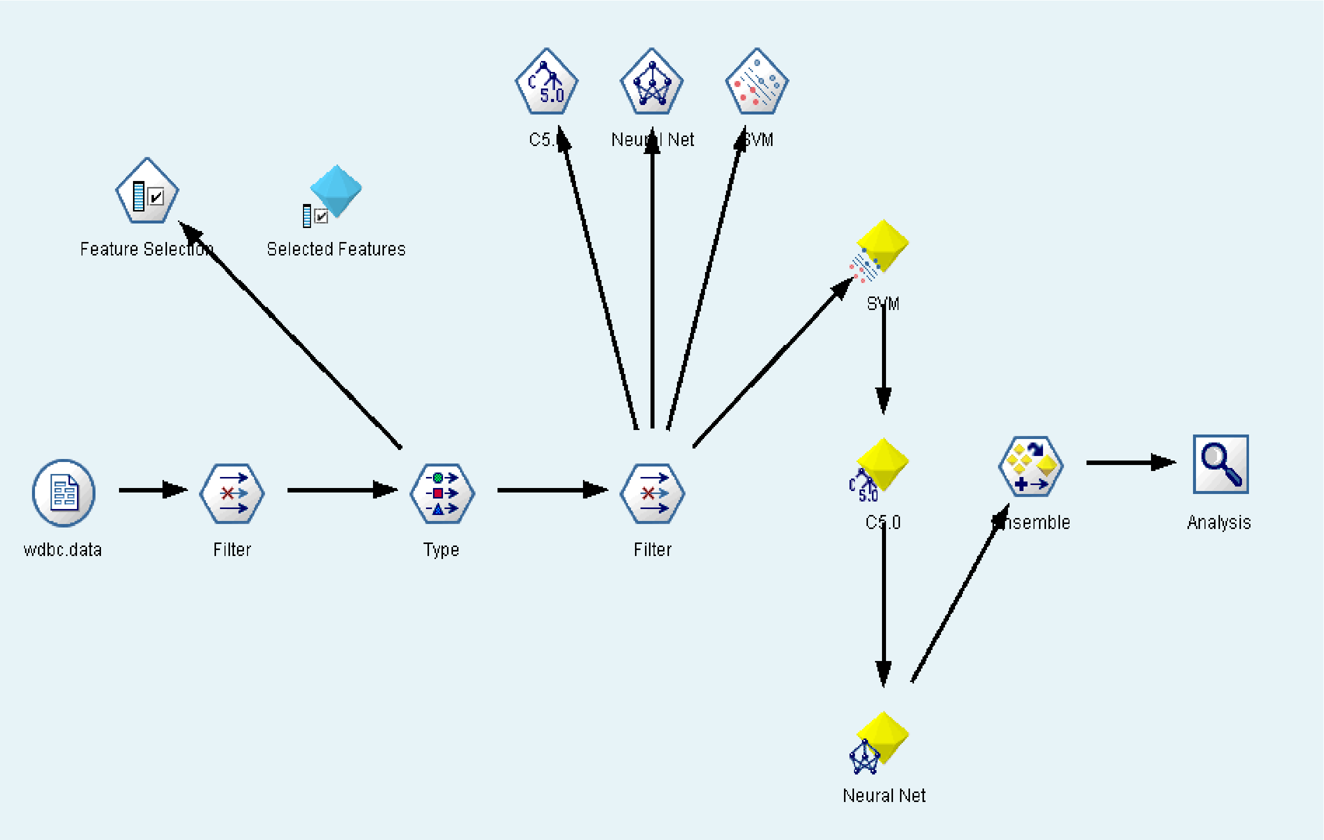

Fig. 1. The Ensemble Model Developed using SPSS Clementine Data Mining Tool

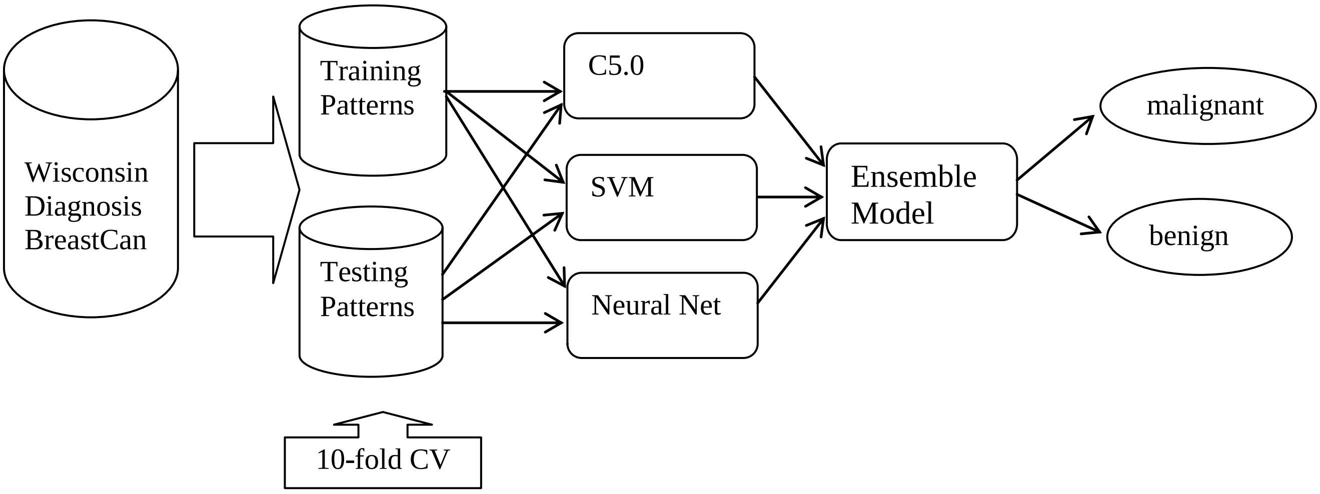

Fig. 2. Representation of Ensemble Model for the Breast Cancer Diagnosis

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.

4 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.

5 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). GöğüsKanseriVerilerininSınıflandırılması.

[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] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

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

[11] 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,

[12] 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.