Performance Improvement with Combining Multiple Approaches to Diagnosis of Thyroid Cancer

Ahmet Akbaş, Uğur Turhal, Sebahattin Babur, Cafer Avci
Department of Computer Engineering, Yalova University, Yalova, Turkey
Email: ahmetakbas@yalova.edu.tr, ugurturhal@hotmail.com, sebahattin_babur@hotmail.com, cafer.avci@yalova.edu.tr
Received June 2013

ABSTRACT

There are a lot of diseases that carry death risk when these diseases are infected to human body, if early measures are not taken. Thyroid cancer is one of them. In USA, number of thyroid cancer cases resulted in death in only 2013 shows necessity of early fight with this disease. This study aims performance improvement in diagnosis of thyroid cancer with machine learning techniques. Study consists of 3 phases. In the first phase, BayesNet, NaiveBayes, SMO, Ibk and Random Forest classifiers have been trained with thyroid cancer train dataset. In the second phase, trained classifiers have been tested with thyroid cancer test dataset and the obtained performance results have been compared. In the third and last phase, approaches named above have been integrated to algorithm AdaboostMI to show difference between of ensemble classifiers from conventional individual classifiers and first two phases have been repeated. With using ensemble approaches performance improvement has been achieved in diagnosis of thyroid cancer. Also, kappa, accuracy and MCC values obtained from these classifier models have been explained in tables and effects on diagnosis of the disease have been shown with ROC graphics. All of these operations have been carried out with WEKA data mining program.

Keywords: Thyroid Cancer; Classification; WEKA

1. Introduction

One of the most frequent cancer type is Thyroid cancer [1]. Tumors of the thyroid gland represent a variety of lesions from well-differentiated benign tumors to anaplastic malignant cancer. Approximately less than 5% - 10% of hyper functioning thyroid nodules develop thyroid cancer and the prevalence of these nodules is estimated to be 5 to more than 20% in humans [2]. According to 2013 records obtained in USA 60,220 thyroid cancer cases have been occurred and 1850 cases of them have been resulted in death [3]. The high death ratio necessitates study in this area.

In the previous studies, thyroid cancer dataset have been classified with various methods and pretty high accuracy values have been achieved [4]. The main purpose of this study is to increase accuracy of classifier made for diagnosis of the disease by combining different machine learning techniques with multi-approaches. In the study, thyroid cancer dataset has been classified with 5 individual classifiers and 1 ensemble classifier and performance improvement has been achieved as compared with previous studies. At the end of each classification, dominant method has been noted with boldface.

2. Methods

There are 21 samples belonging to 7200 subjects in the used thyroid cancer dataset. These samples split in 3 classes. Namely:
- Normal (166).
- Hyperthyroid (368).
- Hypothyroid (6600).

Also, dataset splits in two groups namely, train and test. While there are samples belonging to 3772 subjects in the used train dataset, there are samples belonging to 3428 subjects in the test dataset. Dataset used in the study can be reached from the related link [5].

2.1. Classifiers

1) BayesNet: It is one of the used methods for expressing data modeling and state transition. Properties of networks are their being statistical and branches that linked amongst the nodes being selected according to statistical decisions. BayesNet are directed acyclic Networks and each node express a different variable. Also, ordering between these variables can be shown with BayesNet [6].
\[ P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | PA_i) \quad (1) \]

2) Naïve Bayes: Naïve Bayes is the most basic form of Bayes Networks. All features are independent from given class variable values. This used method is called as conditional independence [7].

\[ f_{naive}(E) = \frac{p(C = +)}{p(C = -)} \prod_{i=1}^{n} p(x_i | C = +) \quad (2) \]

3) Sequential Minimal Optimization (SMO): This method has been improved as an alternative of support vector machine (SVM) and gives chance for making faster classifications. Without any need of forming a structure for classification it finds optimal values for every subset and applies to SVM. In order to train support vector classifier, it applies kernels of polynomial or radial based functions to John C. Platt’s minimal ordered optimization algorithm. In this application, all missing values are usually altered by transforming into lowly features. Coefficients obtained in the output consist of normalized dataset. Equality numbered as 3 has been used for normalization [8].

\[ Z = \frac{X - \mu}{\sigma} \quad (3) \]

Here, \( X \) denotes dataset \((x_i; i = 1, 2, 3, \ldots, N)\), \( \mu \) denotes arithmetic mean, \( \sigma \) denotes standard deviation, \( Z \) denotes normalized dataset. Multi class problems have been solved by using binary classes. Alternatives that are suitable to logistics regression models are used in the outputs obtained from (SVM) to achieve suitable and possible predictions. Logistics regression is a categorical type of regression analysis that used for predicting dependent variable results based on one or more determining variable. Probability estimation in multi classes is performed by combining binary methods of Hastie and Tibishiani [9,10].

4) IBK: It is an algorithm used in WEKA Data Mining Program corresponding to \( k \)-Nearest Neighbour (kNN) Algorithm [11]. It has a lot of disadvantages besides its advantages. Because Ibk makes classification process with mathematical calculations without any need for a structure, Ibk produces results in a very short time. Euclidean distance has been used in the algorithm as a distance function.

\[ d(p, q) = \sqrt{\sum_{i \in \mathbb{R}} (q_i - p_i)^2} \quad (4) \]

Euclidean distance between any two points \((p, q)\) is obtained with equality (4) [12].

5) Random Forest: Breiman has suggested combining decision of numerous multivariate trees that each of them trained with different train sets instead of producing just one decision tree. Different train sets constitute original teaching set with bootstrap and random feature selection. Multivariate decision trees are obtained with CART algorithm. Initially, every decision tree gives itself decision. Class that takes maximum vote in the decision forest is accepted as the last decision and coming test data is included in that class [13]. Random forest algorithm consist of 3 phases [14].

- Draw \( n_{tree} \) bootstrap samples from the original data
- For each of the bootstrap samples, grow \( \rightarrow \) grows an unpruned classification or regression tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample \( m_{pr} \) of the predictors and choose the best split from among those variables. (Bagging can be thought of as the special case of random forests obtained when \( m_{pr} = p \), the number of predictors.)
- Predict new data by aggregating the predictions of the \( n_{tree} \) trees (i.e., majority votes for classification, average for regression).

6) AdaBoostM1: AdaBoost. M1 was developed in 1997 (Freund and Schapire, 1997). AdaBoost is a general version of boosting algorithm. AdaBoost. M1 and AdaBoost.R1 are most used ones for multi class problems and regression problems between its variations, respectively [15,16]. It is a machine learning algorithm developed for reducing drift in boosting learning with instructor [17].

Input: \( m \) number sample series \((x_1, y_1), \ldots, (x_m, y_m)\) and class data \( y_j \in Y = \{1, \ldots, k\} \), learning algorithm (LA), iteration number \( T \)

Start: for every, \( D_i(\hat{e}) = 1/m \) is done.

Do for these values: \( t = 1, 2, \ldots, T \):

- Call for LA by carrying out distributions with obtained \( D_i \) values.
- Form hypothesis \( h_t : X \rightarrow Y \)
- Calculate error for the hypothesis \( \hat{e}_t = \sum_{i \in \mathbb{R}} D_i(\hat{e}) \)
  
  If \( \hat{e}_t > 0.5 \) then calibrate \( T \) to \( T = T - 1 \) and exit loop.

- Calibrate \( \beta_t \) value to \( \beta_t = \hat{e}_t / (1 - \hat{e}_t) \).
- Update distribution value \( D_t \):
  
  \( D_t(i) = \frac{D_i(i)}{Z_t} \times \begin{cases} \beta_t h_t(x_i) = y_i, & \text{1, other cases} \\ \end{cases} \)

\( Z_t \) : Normalization constant.

Output: The last hypothesis \( h_{max} (x) = \arg \arg \max_{y \in Y} \sum_{i \in \mathbb{R}} \log \frac{1}{\beta_t} \)

2.2. Algorithm Steps

Classification process has been made according to fol-
lowing algorithm.
- Classifiers (BayesNet, NaiveBayes, SMO, Ibkve Random Forest) were thought with train dataset that has 3772 samples and 21 features.
- Trained dataset were tested with test dataset that has 3428 samples and 21 features.
- Classification processes were made by ensemble methods used in the previous step with AdaBoostMI algorithm and new results were gathered.

2.3. Performance Measuring

In this study, comparisons have been made by using evaluation methods that are accepted in literature to measure reliability of results.
- Acc: Accuracy is closeness degree of measurement value of one quantity to its real value [18]. More closeness to 1 shows better results.

\[
\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN}
\]  

\((5)\)

TP: Number of true positives.
TN: Number of true negatives.
FP: Number of false positives.
FN: Number of false negatives.

- Kappa: It is a method that measures reliability of comparative cohesion between two data [19]. More bigness from 0 means better results.

\[
K = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}
\]  

\((6)\)

\(Pr(a)\); Summing ratio of cohesions observed for two data.
\(Pr(e)\); Probability of emerging this cohesion by coincidence.

\(K\); Kappa result.

- MCC: It is a method, called as Matthews Correlation Coefficient that used for measuring quality of binary classifiers [20]. More bigness from 0 means better results.

\[
\text{MCC} = \frac{TPxTN - FPxFN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}
\]  

\((7)\)

- ROC: It is a method used for showing performance of binary classifiers with graphics [21].

\[
\text{ROC} = \frac{\text{sensitivity}}{1 - \text{specificity}}
\]  

\((8)\)

2.4. Classification

Obained figures as a result of classification process of data with above-stated methods by using WEKA data mining program have been given in Table 1.

Table 1. Individual classification results.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Performance Values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc</td>
</tr>
<tr>
<td>BayesNet</td>
<td>0.976</td>
</tr>
<tr>
<td>NaiveBayes</td>
<td>0.949</td>
</tr>
<tr>
<td>SMO</td>
<td>0.938</td>
</tr>
<tr>
<td>Ibk</td>
<td>0.912</td>
</tr>
<tr>
<td>RandomForest</td>
<td>0.990</td>
</tr>
</tbody>
</table>

Random forest classifier, that produced the highest accuracy and ROC figures in the previous step, produced the best results in here, too. ROC performance graphics obtained as a result of classification process have been shown in Figures 1 and 2.

As it can be seen in Figures 1 and 2, between used methods, while random forest is the classifier that produces the highest ROC values in thyroid cancer diagnosis, the smallest values were produced by Ibk classifier. When viewed to these results, Ibk classifier falls short in thyroid cancer diagnosis. When the same graphics were analyzed again, Random Forest classifier has been observed without change in the performance. But SMO classifier, contrary to Random Forest classifier, when combined with AdaBoostMI classifier, performance increase has been observed in the ROC figure. Because of this, while making performance analysis, methods more than one were used. In this type of situations, MCC and Kappa figures play effective roles in determining the best method.

3. Results

When looked at classification accuracy results and ROC graphics, in diagnosis of thyroid cancer, random forest has been observed as more effective than other used methods. When also regarded previous studies, accuracy results of classification process in this dataset have been observed with their coming fairly close to %100. Making predictions with such high accuracy values makes study in this area hard in the subject of thyroid cancer. When taken into account obtained results and used datasets being old (1992), a need can be seen for a new dataset to obtain more accurate and more valid results. At this stage, new and original datasets can be obtained, a result of joint studies in hospitals, laboratories, medical centers and studies can be conducted over these datasets. Also, in the course of these studies, classifier effects directed to problem solving can be compared by using different per-
Table 2. Ensemble classification results.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Performance Values</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>BayesNet</td>
<td>0.987</td>
<td>0.995</td>
</tr>
<tr>
<td>NaiveBayes</td>
<td>0.949</td>
<td>0.566</td>
</tr>
<tr>
<td>SMO</td>
<td>0.942</td>
<td>0.479</td>
</tr>
<tr>
<td>lbk</td>
<td>0.912</td>
<td>0.315</td>
</tr>
<tr>
<td>RandomForest</td>
<td>0.991</td>
<td>0.940</td>
</tr>
</tbody>
</table>

**REFERENCES**


