Feature and Classifier Selection for Respiratory Sound Classification

Zafer Yasir YILMAZ[1], Yasemin P. KAHYA[2]

[1] Department of Electrical and Electronics Engineering, Boğaziçi University, 34342 Istanbul, Turkey
yasir.yilmaz@boun.edu.tr
[2] Department of Electrical and Electronics Engineering, Boğaziçi University, 34342 Istanbul, Turkey
kahya@boun.edu.tr

Abstract

In this study, Mel Frequency Cepstal Coefficients, Autoregressive(AR) parameters and their combination are compared as features in classifiers for recognizing pathological and healthy subjects. Results show that AR parameters outperform both MFCCs and combined features. For fast and efficient classification, AR parameters of respiratory sounds are studied to characterize lung sounds for diagnosis of pathological subjects. Various time domain, frequency domain and time-frequency domain features are added to the feature set. After feature extraction step, feature selection based on feature importance scores and SVM-RFE are used as feature selection step. Experiments are conducted on a dataset of 30 subjects and several machine learning algorithms are used as for classification. For optimum computation time and classification accuracy, we propose a method based on random forests. The proposed method achieves an accuracy of 93.3 % for 30 subjects.

1. Introduction

The first tool for a physician to assess the health condition of the respiratory system is auscultation of lung sounds. However, this method is very subjective and depends strongly on the experience and expertise of the physician. Error caused by human factor can be reduced by technology. This is achieved by using computarized respiratory sound analysis on sound data. For this task, some researchers have used Autoregressive (AR) parameters [1] while some others have analyzed Mel Frequency Cepstal Coefficients (MFCCs) [2]. In this work, AR model, MFCCs and their combination are studied and their performances are compared.

To find out the best feature set correlating subjectś health condition, various classification algorithms were applied to the feature sets mentioned above. In the comparison of the three feature sets in terms of computational complexity and accuracy, one provided better results than the other two.

In addition, classification accuracy is improved by the addition of some time, frequency and time-frequency domain features to the feature set. After extracting these features, the subset having highest accuracy and lowest computational complexity is searched among the features. To find the optimum subset, dimensionality reduction with principal component analysis (PCA), support vector machine-recursive feature elimination (SVM-RFE), and tree-based feature selection algorithms were used. Eventually, one feature selection algorithm is chosen over the others for the proposed method.

In our proposed method for lung sound classification, 21 time domain, 4 frequency domain and 4 time-frequency domain features are used for each segment. Based on these features, KNN, SVM, Naive Bayes, Random Forest, Extra Trees and Gradient Boosting algorithms are used as classifiers. After classification of all segments, subjects are diagnosed based on majority voting of segments. Random Forest gives the best accuracy with 10 features, which is 93.3 % where its sensitivity is 93.75 % and its specifity is 92.85 %.

2. Data Preparation and Segmentation

Each subject has 4 recordings of 10 sec duration consisting of 96000 samples. The recordings are divided into 512-sample segments with 50 % overlapping [1]. Hamming window is used to reduce spectral leakage. For each recording, 375 segments are used to extract feature vectors, which are later classified as healthy or pathological. The decision on each subject is made using majority voting on all segment decisions.

3. Feature Extraction Methods

The order, p, of the AR model is selected as 10 and the number of Cepstal coefficients is selected as 13. As given in figure 1, there are 3 feature vectors and these vectors are fed to classifiers separately for segment classification. After segment classification, subjects are classified based on segment votes.

3.1. Mel Frequency Cepstal Analysis

The FFT-based Mel Frequency Cepstral Coefficients (MFCCs) constitute the feature set used in the detection of pathological respiratory sounds. For the extraction of MFCCs, power spectrum of segment is obtained and the short-term power spectrum is compressed using logarithm. The compressed power spectrum is converted to Mel scale after discrete cosine transform (DCT) computation. The detailed computation of MFCCs is explained in [3].

The Mel scale is related to perceived frequency of humans. The formula for mapping the true frequency to the perceived frequency is:

$$M(f) = 1125ln(1 + f/700) \tag{1}$$

3.2. Autoregressive Analysis

AR model represents a time series in which the next value in the sequence is predicted based on a certain number of previous values [4]. The p-order AR model is represented as follows:

$$x_t = a_1 x_{t-1} + a_2 x_{t-2} + \cdots + a_p x_{t-p} + \epsilon_t \tag{2}$$

where x_i represents the respiratory sound signal, a_i(i=1,2,...,p) is the i-th AR coefficient. For the extraction

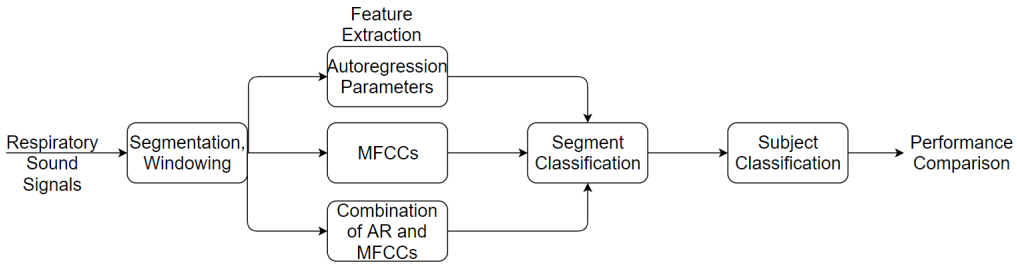

Figure 1. An Illustration of Feature Extraction Methods Comparison

of AR coefficients from respiratory sound signals, the model order p is selected and the AR coefficients are estimated. $(a_1,a_2,\cdots,a_p)$

The minimization of the Akaike Information Criteria (AIC) value is used for selecting the model order p of the AR model [5]. After selecting the best model order p based on this criteria, estimated p coefficients are used as representation of respiratory sound signals. In this study, AR model p order is chosen from the range of 4-14.

3.3. Comparison of Feature Extraction Methods

Feature extraction based on AR coefficients is much more efficient than the MFCCs based one; therefore, AR coefficients were chosen as features in this work.

Table 1. Results of Feature Extraction Methods on Subject Classification

	AR	**MFCCs**	**Combination**
KNN	73%	57%	68%
Naive Bayes	57%	52%	57%
Extra Trees	78%	68%	73%
SVM	73%	68%	73%
RF	84%	78%	84%
GB	78%	68%	73%

The three feature vector sets comprised of AR parameters (model order p is 10), Mel Frequency Cepstal Coefficients (13 coefficients) and the combination of AR parameters and MFCCs (23 features) were costructed. For segmentation of respiratory sound signals, the parameters (128,256,384,512,768,1024,1536) samples were experimented as frame size and the extracted features for these frames were fed into random forest classifier. The segment size providing the best accuracy was used. In AR feature vector, frame size was selected as 512 samples with 50 % overlap. In MFCC feature vector, frame size is selected as 1024 samples with 50 % overlap. In combination vector, frame size is 512 with 50 % overlap.

Leave-one-out cross-validation (LOOCV) was used for 19 subjects (10 healthy, 9 patient). The AR feature vector and the combination feature vector gave similar results in the subject classification. The best performances were achieved using these feature vectors and Random Forests as the classifier. Highest efficiency in computational complexity as well as highest accuracy are achieved using AR based feature vectors as can be seen in 1.

3.4. Additional Features to the AR Parameters

To increase the classification performance, additional features were added to the AR feature set so that resepiratory sounds would be represented better. These additional features comprised of time, frequency and time-frequency domain features and were calculated for each segment. In this work, the time-frequency features are constructed using wavelet packet decomposition (WPD). Bior3.7 wavelet was used in wavelet analysis and the short-time signal was decomposed up to 3 levels. These wavelet coefficients were used in computation of wavelet energy. Fast Fourier Transform (FFT) was employed to extract the frequency domain features. A total of 19 features grouped as 12 time domain, 3 frequency domain and 4 time-frequency domain from each signal were extracted for each frame. A summary of these features is presented in Table 2. These features were extracted as in [6]. The feature vector consists of AR parameters (10 parameters) and 19 additional features, so the number of features used in the proposed method is 29.

Table 2. Additional Features Extracted from each Short-Time Signal

Time Domain	Skewness	Power Spectral Density of FFT
RMS	Line Integral	**Time-Frequency Domain**
Variance	Shannon Entropy	Energy of WPD Detail Coefficient One
Peak Value	Shape Factor	Energy of WPD Detail Coefficient Two
Crest Factor	**Frequency Domain**	Energy of WPD Detail Coefficient Three
Kurtosis	Peak to Peak Value	Energy of WPD Approximate Coefficient Three
Clearance Factor	Peak Value of FFT	
Impulse Factor	Energy of FFT	

4. Feature Selection Methods

It is very important to remove redundant features in a classification algorithm. Therefore, the minimum number of features for optimum correct classification rate and efficient computation should be determined. Feature selection provides the classification task with three critical properties. Firstly, selecting optimum features reduces redundancy in the data, weakening

the effect of noise on classification. Secondly, optimum feature selection enhances classification accuracy by eliminating the data that misleads the classification task. Finally, optimum feature selection reduces the data size, decreasing the training time.

In this work, three types of feature selection methods were tested. These are Principal Componet Analysis (PCA), Recursive Feature Elimination and lastly Feature Importance Ranking.

4.1. Principal Component Analysis

Principal Component Analysis (PCA) is a method which reduces the size of the feature space by using advanced numerical methods in data analysis [7]. PCA transforms various potentially related factors into fewer factors called principal components. In other words, PCA obtains vectors of different sizes by changing the directions of multi-dimensional cluster coordinates and these vectors are eigenvectors. Consequently, PCA accelerates the process of the training of classification algorithm by diminishing the dimensionality of vast informational vectors. However, the data loose some properties while the multidimensional vector space maps to a smaller vector space. Therefore, PCA aims to reduce the loss of data by keeping variance high. PCA eliminates the redundant characteristics of the data and enables to distinguish patterns and anomalies in the data, much more effortlessly than without PCA. It is also used for easy visualization because PCA's multidimensional data shrinks.

4.2. SVM-Recursive Feature Elimination

Recursive feature elimination is a method that lists features in importance order by setting up a model repeatedly. This recursive method selects the feature that performs best or performs poorly at the most remarkable level, puts this feature in an edge, and restarts the procedure with the rest of the features. This procedure continues until all features in the dataset are depleted. Features are then positioned by when they were removed. Regarding SVM-RFE, it is a application of RFE which establishes the SVM model to rank features. This process is terminated when the margin of separating hyperplane is reached to the maximum level and the selected features comprise the set of features that leads to the largest margin [8].

4.3. Feature Selection Based on Extra-Trees

The Extra-Tree strategy (extremely randomized trees) was proposed in [9], with the fundamental target of further randomizing tree working with regards to numerical info features, where the decision of the ideal cut-point is in charge of a huge extent of the change of the actuated tree. Concerning random forests [10], the strategy drops utilizing bootstrap duplicates of the learning test, and as opposed to endeavoring to locate an ideal cut-point for every last one of the K randomly picked features at every hub, it chooses a cut-point at random. One of the key favorable benefits of Extra-Trees is that Extra-Trees can gauge significance score of each feature to take in the effect of each feature with respect to the forecast of the classes. The feature selection calculation, specifically random forest, uses the significance scores from an Extra-Tree to choose least number of profoundly discriminative features, consequently.

4.4. The Comparison of Feature Selection Methods

For high dimensional data, the number of features may be enormous that makes the manual examination of the feature significance scores and selection of the most important features for classification very difficult. In this respect, automatic feature selection strategy in light of significance scores can lead to choose important, minimized and discriminative features. Therefore, Feature Selection based on Extra-Trees (ET-FS) is better than the other two methods in terms of computational complexity of the selection of relevant features.

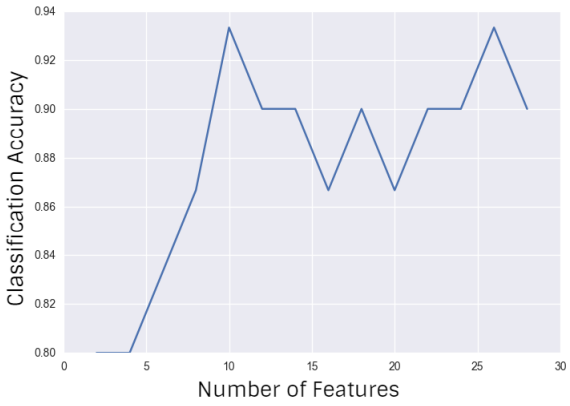

Figure 2. Feature Selection Based on Extra-Trees with Random Forest Classifier

The three feature selection methods are compared in terms of accuracy. In this sense, KNN, Naive Bayes, Random Forest, Extra-Trees and Gradient Boosting algorithms were tested with selected features. In PCA analysis, the best component number was found to be 10 and the most accurate classifier was Random Forest classifier with subject classification accuracy of 83.3%. As for SVM-RFE, the most accurate classification was achieved with 14 features and the accuracy was 90% with Random Forest classifier. Random Forest-Recursive Feature Elimination (RF-RFE) was also tested [11] and Random Forest algorithm gives the best result of 90% using 12 features that RF-RFE selected. The most accurate predictions were achieved by using ET-FS. Random Forest classified 93.3% of respiratory sounds correctly by using 10 important features. As seen in figure 2, the classification accuracy of 93.3% was obtained with 26 important features. However, the first 10 significant features were selected when taking computational burden of the extraction of features and training time of classifier algorithms into account.

The three feature selection methods are evaluated on two aspects, their capacity to find small subsets with a high discrimination capability, and computational burden of selecting the small subsets. Considering the former aspect, it was observed that both ET- FS and RFE methods outperform the PCA analysis. As for the latter, ET- FS has better computational performance.

5. Results and Proposed Method

In this paper, we show that AR analysis constitute better features than Mel Frequency analysis and Extra Tree based feature selection outperforms PCA and RFE. Therefore, it is decided to use AR parameters and ET-FS for the proposed method. After selecting the best feature subset, KNN, Naive Bayes, Random Forest, Extra-Trees and Gradient Boosting algorithms were tested with selected features. The most accurate resuts were obtained with Random Forest. Therefore, Random Forest

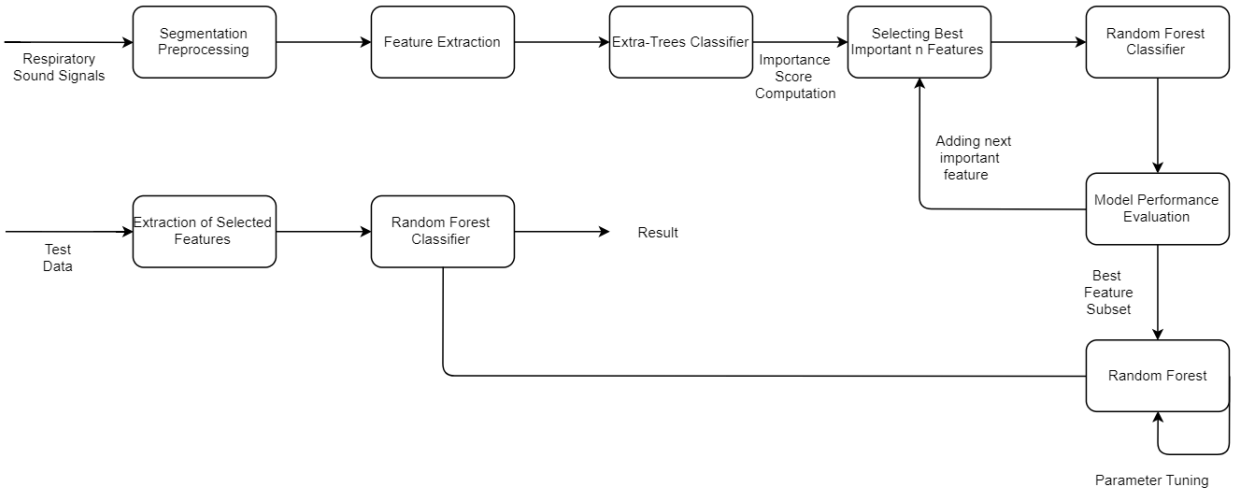

Figure 3. Block Diagram of Proposed Method

is used in the proposed method, which is shown in figure 3.

Table 3. Classification accuracies of machine learning algorithms with selected features

	Subject Classification	Segment Classification
KNN	66.7%	59.54%
Naive Bayes	46.7%	51.11%
Random Forest	93.3%	69.40%
Extra-Trees	86.7%	71.90%
Gradient Boosting	83.3%	71.51%

The proposed method was tested with 30 subjects by using leave-one-out method. One of the pathological subjects and one of the healthy subjects were misclassified and the results are shown in table 4. Statistical performance results are given in table 5

Table 4. Confusion matrix for the result of random forest classifier

		Prediction	
		Healthy	Pathological
Actual	Healthy	13	1
	Pathological	1	15

Table 5. Values of statistical performance parameters for the random forest classifier

Statistical performance parameters	Values (%)
Specificity	92.85
Sensitivity	93.75
Accuracy	93.3

6. Conclusion

In this study, we showed that AR parameters represents respiratory sound signals better than MFCCs. MFCCs are useful in adventious sound classification [12], but AR parameters are better than MFCCS in the subject classification. Also, computation of AR parameters is more efficient than computation of MFCCs. Therefore, we decided using AR parameters as features for each segment. Then, we added some time domain, frequency domain, and time-frequency domain features to the feature vector.

After constructing feature vector, the best feature subset was searched. In feature selection phase, we tested three different feature selection algorithms. We show that feature selection based on extra trees can select more compact feature subsets compared to the other two methods, while preserving classification accuracy. We applied random forests classifier to evaluate the effectiveness of the feature selection methods and demonstrate that random forests with feature subset selected by extra trees performs comparatively better than feature subsets selected by other feature selection methods. We decided to use random forest as base classification algorithm. The five classifier algorithms (KNN, Naive Bayes, Random Forest, Extra-Trees, Gradient Boosting) were trained with the selected features. Random Forest classifies the subjects most accurately. After selecting random forest, we optimized parameters of random forest.

We computed accuracy by using leave-one-out method. The data consists of 14 healthy subject and 16 pathological subject.The algorithm classifies 28 subject correctly and 2 subject incorrectly. Both of the number of false positives and true negatives are 1. The overall accuracy is 93.3% ,specifity is 92.85% and sensitivity is 93.75%. In the future, we plan to extend our dataset to include a higher number of subjects.

7. References

[1] Sankur, Bulent, et al. "Comparison of AR-based algorithms for respiratory sounds classification." *Computers in Biology and Medicine* 24.1 (1994): 67-76.

[2] Sengupta, Nandini, Md Sahidullah, and Goutam Saha. "Lung sound classification using cepstral-based statistical

features." *Computers in biology and medicine* 75 (2016): 118-129.

[3] Molau, Sirko, et al. "Computing mel-frequency cepstral coefficients on the power spectrum."*Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference* on. Vol. 1. IEEE, 2001.

[4] Akaike, Hirotugu. "Fitting autoregressive models for prediction." *Annals of the institute of Statistical Mathematics* 21.1 (1969): 243-247.

[5] Akaike, Hirotogu. "Information theory and an extension of the maximum likelihood principle." *Selected Papers of Hirotugu Akaike*. Springer New York 1998. 199-213.

[6] Kimotho, James Kuria, and Walter Sextro. "An approach for feature extraction and selection from non-trending data for machinery prognosis." *Proceedings of the second european conference of the prognostics and health management society.* 2014.

[7] Richardson, Mark. "Principal component analysis." *URL: http://people. maths. ox. ac. uk/richardsonm/SignalProcPCA. pdf (last access: 3.5. 2013). Aleš Hladnik Dr., Ass. Prof., Chair of Information and Graphic Arts Technology, Faculty of Natural Sciences and Engineering, University of Ljubljana, Slovenia ales. hladnik@ ntf. uni-lj. si* (2009).

[8] Xie, Zong-Xia, Qing-Hua Hu, and Da-Ren Yu. "Improved feature selection algorithm based on SVM and correlation." *Advances in Neural Networks-ISNN* 2006 (2006): 1373-1380.

[9] Geurts, Pierre, Damien Ernst, and Louis Wehenkel. "Extremely randomized trees." *Machine learning* 63.1 (2006): 3-42.

[10] Breiman, Leo. "Random forests." *Machine learning* 45.1 (2001): 5-32.

[11] Granitto, Pablo M., et al. "Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products." *Chemometrics and Intelligent Laboratory Systems* 83.2 (2006): 83-90.

[12] Sengupta, Nandini, Md Sahidullah, and Goutam Saha. "Optimization of cepstral features for robust lung sound classification." *India Conference (INDICON), 2015 Annual IEEE*. IEEE, 2015.