Heart Sound Anomaly and Quality Detection using Ensemble of Neural Networks without Segmentation

Morteza Zabihi[1,*], Ali Bahrami Rad[2,*],
Serkan Kiranyaz[3], Moncef Gabbouj[1], Aggelos K. Katsaggelos[4]

[1]Tampere University of Technology, Tampere, Finland
[2]University of Stavanger, Stavanger, Norway
[3]Qatar University, Doha, Qatar
[4]Northwestern University, Evanston, USA.

Abstract

Phonocardiogram (PCG) signal is used as a diagnostic test in ambulatory monitoring in order to evaluate the heart hemodynamic status and to detect a cardiovascular disease. The objective of this study is to develop an automatic classification method for anomaly (normal vs. abnormal) and quality (good vs. bad) detection of PCG recordings without segmentation. For this purpose, a subset of 18 features is selected among 40 features based on a wrapper feature selection scheme. These features are extracted from time, frequency, and time-frequency domains without any segmentation. The selected features are fed into an ensemble of 20 feedforward neural networks for classification task. The proposed algorithm achieved the overall score of 91.50% (94.23% sensitivity and 88.76% specificity) and 85.90% (86.91% sensitivity and 84.90% specificity) on the train and unseen test datasets, respectively. The proposed method got the second best score in the PhysioNet/CinC Challenge 2016.

1. Introduction

Heart auscultation is one of the cursory and cost-effective diagnostic tests. It can provide primary evaluation of hemodynamic status and detect a cardiovascular disease, such as ventricular septal defects, and stenosis in aorta [1]. Heart sound (or phonocardiogram) can also offer additional diagnostic tests for further medical assessments.

However, the practical applications of heart sound highly depend on cognitive skills and expertise of the medical examiner. The limitation of audible frequency range, environmental noise, and variation in recording regions are other major shortcomings of this test. In order to address these shortcomings in a cost-effective diagnostic tests in ambulatory monitoring, several techniques [2-5] have been proposed for automatic analysis of heart sounds.

Phonocardiogram (PCG) signal analysis can fall into two major categories. The first type of approaches is based on temporal segmentation, i.e. identifying the cardiac cycles and localizing the position of the first (S1; beginning of the systole) and second (S2; end of the systole) primary heart sounds. The variation in the duration of S1 and S2, and their intensities are considered as the conclusive signs of cardiac anomalies.

Several studies have been conducted to PCG segmentation using different envelope extraction methods such as Shannon energy [2], Shannon entropy [3], Hilbert-Huang transform [4], and autocorrelation [5]. The envelope of signal attenuates the noise and amplifies the low-intensity components of the signal. Some segmentation approaches use envelope extraction based on wavelet transform to gain the frequency characteristics of S1 and S2 sound [6]. In the second type approaches, abnormal PCG records are detected without segmentation [7,8].

In this study we follow the second type approach towards PCG classification. The main motivation behind this is to remove the dependency on segmentation and reduce the computational burden. The main contributions of this study are the detailed investigation of time-frequency features (Section 2.1) and the design of effective neural network ensembles (Section 2.2). The proposed approach is evaluated on one of the largest public heart sound database [10]. The results are discussed in Section 3 and the Section 4 conclude the paper and suggests topics for future research.

2. Methodology

For this challenge, 3454 PCG labelled records (including training and validation sets) are provided by

[*] The first two authors have contributed equally to this paper.

Physionet/Computing in Cardiology Challenge 2016 [9]. More detailed Information can be found in [10]. The proposed feature extraction and classification approaches will be discussed next.

2.1. Feature extraction

In the initial phase of this work, 40 features in the time, frequency, and time-frequency domains were extracted. Then, a subset containing 18 features were selected using a wrapper-based feature selection scheme [11] in which sequential forward selection search algorithm [12] was used. The selected features can be categorized into 5 types as follows:

(1) Linear Predictive Coefficient (LPC): the first, third, sixth, eight, ninth, and tenth coefficients of 10th-order linear predictor are used as features.

(2) Entropy based features: Natural and Tsallis entropy of PCG signals are calculated as,

$$H(x) = -\sum_i p(x_i) \ln p(x_i) \tag{1}$$

$$S_q(x) = \frac{k}{q-1}(1 - \sum_i p(x_i)^q) \tag{2}$$

where $p(x_i)$ is the probability of i^{th} samples of PCG signal, x. k and q are real parameters equal to 1 and 2, respectively.

(3) Mel Frequency Cepstral Coefficients (MFCCs) based features: The MFCCs of each PCG signal are computed based on the parameter of 14 coefficients for frame duration of 25 ms with 10 ms overlap. The extraction of MFCCs results in 14 coefficients for each frame, $C_{i,j}$, where i and j are the number of features and frames, respectively. Once the $C_{i,j}$ is calculated, three features are extracted as follows:

$$AV_{MFCC} = \frac{1}{N}\sum_{j=1}^{N} \min_{i \in I} C_{i,j}, \qquad j = 1,2, \dots, N \tag{3}$$

$$\mu_r^{max,MFCC} = E(\max_{i \in I} C_{i,j} - \mu_j)^r \tag{4}$$

$$\mu_r^{skew,MFCC} = E\left(\underset{i \in I}{Skew}\, C_{i,j} - \mu_j\right)^r \tag{5}$$

where $\min_{i \in I} C_{i,j}$, $\max_{i \in I} C_{i,j}$, and $\underset{i \in I}{\text{Skew}}\, C_{i,j}$ are the minimum, maximum and the skewness of each column of matrix C. μ is the average and $r = 2$.

(4) Wavelet transform based features: Discrete wavelet transform (Daubechies 4) is applied to each PCG signal and the approximation coefficients of level 5 (a_5) and the detail coefficients of level 3 to 5 (d_3, d_4, and d_5) are used for feature extraction as follow:

$$H\,(a_5) = -\sum_i p(a_{5_i})\, ln\, p(a_{5_i}) \tag{6}$$

$$H_q\,(d_5) = \frac{1}{q-1} ln(\sum_i p(d_{5_i})^q) \tag{7}$$

$$H\,(d_4) = -\sum_i p(d_{4_i})\, ln\, p(d_{4_i}) \tag{8}$$

$$\lambda(d_3) = log_2(\sigma^2(d_3)) \tag{9}$$

where σ^2 is the variance. H_q is known as Rényi entropy in which $q = 2$.

(5) Features extracted over power spectral density: The power spectral density of each signal is calculated based on the normalized frequency (i.e. between 0 and 1), and then the following features are extracted as follows:

$$MPSD_{Centroid} = \frac{\int f P(f)^2 df}{\int P(f)^2\, df} \tag{10}$$

$$AUC_1 = \int_{0.7}^{0.8} P(f) df \tag{11}$$

$$AUC_2 = \int_{0.9}^{1} P(f) df \tag{12}$$

where $p(f)$ represents the power spectral density and $MPSD_{Centroid}$ is the modified power spectral density centroid. AUC_1 and AUC_2 show the areas under the curve over the two specified frequency intervals, i.e. 0.7-0.8, and 0.9-1.

2.2. Classification

In this work, the proposed classification algorithm has two main steps. In the first step, the bad quality recordings (class 0) are detected. In the second step, among the good quality signals, the normal (class -1) and abnormal (class 1) PCGs are classified. This process includes two different classifiers: one for good/bad quality recordings, and the other for normal/abnormal recordings (Fig. 1). In the following the structure of these classifiers are described in details.

2.2.1. Ensemble of neural networks

Ensemble based classification systems construct a set of classifiers and then classify new samples by integrating the results of those classifiers to obtain a better classification performance. In this work, we used an ensemble of 20 feedforward Artificial Neural Networks (ANNs) with two hidden layers in each, and 25 hidden neurons at each layer. The number of neurons in the output layer is 4 for the

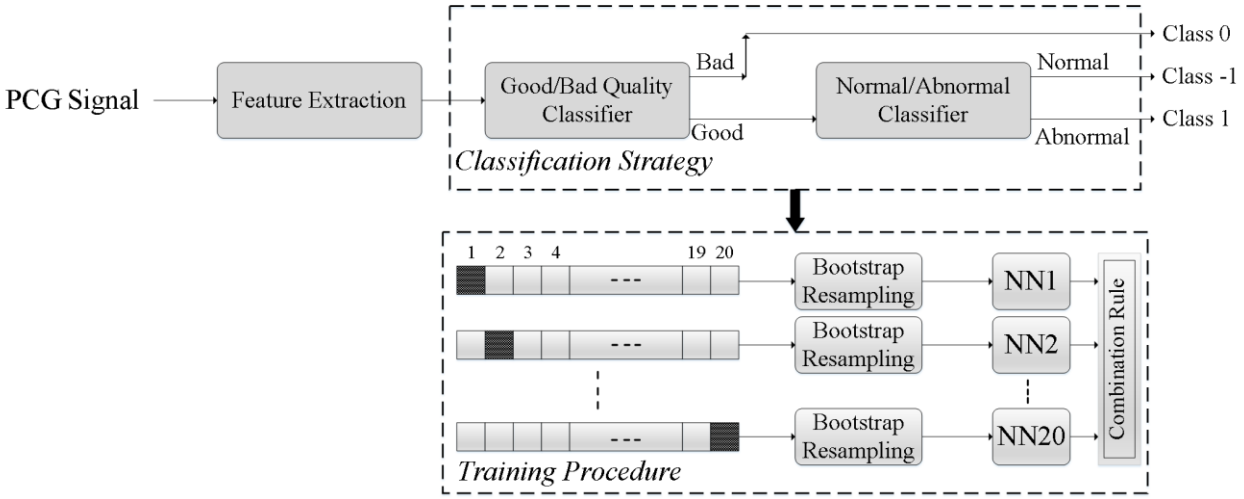

Figure 1. Schematic demonstration of classification strategy and training procedure.

purpose of two classification tasks simultaneously, i.e., classification of the signal type (normal vs. abnormal) and quality (good vs. bad). We used the hyperbolic tangent activation function as the transfer function. In addition, we used Levenberg–Marquardt optimization method [13] with Bayesian regularization backpropagation [14] training algorithm.

2.2.2. Training

To construct the training data for the proposed ensemble of ANNs, we used 20-fold cross-validation committee [15]. To do so, we generate 20 replicates from the original training data, and then by removing 5% disjoint random subset from each replica, 20 overlapping training sets were constructed. In addition, due to the data imbalance problem between normal and abnormal signals, in each training sets we used bootstrap resampling method to make the data balanced in the following way: First, we calculate the number of normal signals (which is higher than abnormal). Then, by random sampling with replacement from abnormal signals the size of the selected set becomes equal to the size of the normal set.

Although the data get balanced by using the aforementioned technique, caution is needed. In bootstrap resampling usually the size of the selected samples is equal to or smaller than the size of the original data, but in the proposed method the size of the selected samples is larger than the original data. This would mimic a situation where we have a larger dataset than what we actually had, i.e. we will get higher precision in the bootstrap resampling than what we have in our data, which in our case leads to an overfitting problem for abnormal recordings. Our impression was that to address this drawback we could use the so-called *jackknife* resampling (i.e. random sampling without replacement) of normal recordings instead of bootstrap resampling of abnormal recordings. However, this technique had inferior performance than the proposed technique. Thus we discarded it. Fig. 1 demonstrates both the classification strategy and the training procedure.

2.2.3 Combination rule

The last key factor for the proposed classification technique is the combination rule to integrate the results of 20 classifiers. In this work we used two approaches: 1) non-trainable rule and 2) trainable rule. In the first approach, we used unweighted average of class-specific outputs [16] of the ANNs. In the second approach, the combination is based on the voting system of the class labels which is learned during a 10-fold cross-validation scheme as follows: if at least 17 out of 20 classifiers recognize a signal as bad quality, our algorithm recognizes it as bad quality and assigns the label 0. For the remaining signals, which recognized as good quality, our algorithm decides whether it is normal or abnormal such that if at least 7 out of 20 classifiers recognize it as abnormal our algorithm detects the signal as abnormal (1) and otherwise as normal (-1).

3. Results and discussion

We have conducted experiments in order to compare the two proposed combination rules (Section 2.2.3) for PCG classification. The performance of each combination rule is evaluated using Sensitivity (Se), Specificity (Sp), and Score (Sc) based on the provided scoring mechanism of the PhysioNet/Computing in Cardiology Challenge 2016 [9,10] by running a 10-fold cross-validation procedure.

In Table 1, these results are shown. Although the performance of the two rules are fairly close (91.17% vs.

Table 1. The average (Ave) and standard deviation (Std) of Sensitivity (Se), Specificity (Sp), and Score (Sc) using 10-fold cross-validation procedure for the combination rules.

	Train (Rule 1)			Train (Rule 2)		
	Se (%)	Sp (%)	Sc (%)	Se (%)	Sp (%)	Sc (%)
Ave.	89.82	92.53	91.17	94.23	88.76	91.50
Std.	2.79	1.23	1.67	2.22	1.96	1.35

91.50%), the second rule is proposed for applying on the unseen test data. The proposed solution achieved the overall score of 85.90% (86.91% Se and 84.90% Sp) on the unseen test dataset, which is the second best score in the competition.

As discussed in Section 2.1, 18 proposed features were selected using a wrapper-based feature selection scheme. In that scheme, an internal "feature selection classifier" (FS classifier) was used to detect only normal/abnormal signals. This means that the features were not selected by considering the quality detection task into account and this was in accordance by the initial scoring strategy. Only during the final stage, the organizers changed the scoring strategy and the remaining time was not sufficient to redesign the proposed system accordingly. Consequently, this decreases the final score. Thus, we decided to adapt our former method to the new strategy. In the future, for further performance improvement, two independent classification scenarios will be designed with such features that will be selected accordingly.

4. Conclusion

This study proposes a solution for anomaly and quality detection of PCG recordings without segmentation. The proposed method got the second best score in the PhysioNet/CinC Challenge 2016. Many previous methods based on PCG analysis are relied on segmentation which potentially increases the computational burden. The achieved sensitivity (86.91%) and specificity (84.90%) on the unseen test dataset demonstrate the potential of improvement in the future. Designing specific features and additional classifier for quality detection may increase the system performance.

References

[1] McConnell ME, Pediatric Heart Sounds, Springer Science & Business Media, 2008.

[2] Balili CC, Sobrepena M, Naval PC, Classification of heart sounds using discrete and continuous wavelet transform and random forests. In 3rd IAPR Asian Conference on Pattern Recognition, Kuala Lumpur, 2015.

[3] Moukadem A, Dieterlen A, Brandt C. Shannon Entropy based on the S-Transform Spectrogram applied on the classification of heart sounds. In IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, 2013.

[4] Zhang D, He J, Jiang Y, Du M. Analysis and classification of heart sounds with mechanical prosthetic heart valves based on Hilbert-Huang transform. International Journal of Cardiology 2011;151(1):126–127.

[5] Kao WC, Wei CC. Automatic phonocardiograph signal analysis for detecting heart valve disorders. Expert Systems with Applications 2011;38(6):6458–68.

[6] Huiying L, Sakari L, Iiro H. A heart sound segmentation algorithm using wavelet. In 19th international Conference IEEE/EMBS, Chicago, IL, 1997.

[7] Yuenyong S, Nishihara A, Kongprawechnon W, Tungpimolrut K. A framework for automatic heart sound analysis without segmentation. In BioMedical Engineering OnLine, 2011; 10-13.

[8] Deng SW, Han JQ. Towards heart sound classification without segmentation via autocorrelation feature and diffusion maps. Future Generation Computer Systems 2016;60:13–21.

[9] Physionet. Physionet Challenge 2016. https://physionet.org/challenge/2016/. Accessed 8 September 2016.

[10] Liu C, Springer D, Li Q, et. al. An open access database for the evaluation of heart sound algorithms. Physiological Measurement 2016;37(11).

[11] Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell1997;97:273–324.

[12] Whitney AW. A direct method of nonparametric measurement selection. IEEETrans Comput 1971;20:1100–3.

[13] Hagan MT, Menhaj MB. Training feedforward networks with the Marquardt algorithm. IEEE transactions on Neural Networks 1994;5(6):989-93.

[14] MacKay DJC. Bayesian interpolation. Neural Computation 1992;4(3):415–447.

[15] Parmanto B, Munro PW, Doyle HR. Improving committee diagnosis with resampling techniques. Advances in neural information processing systems 1996:882-8.

[16] Polikar R. Ensemble based systems in decision making. Circuits and Systems Magazine, IEEE 2006;6(3):21–45.

Address for correspondence:

Morteza Zabihi, Ali Bahrami Rad
P.O. Box 527, FI-33101 Tampere, Finland.
P.O. Box 8600 Forus, N-4036 Stavanger, Norway.
morteza.zabihi@tut.fi, abahramir@gamil.com.