Automatic Heart Sound Recording Classification using a Nested Set of Ensemble Algorithms

Masun Nabhan Homsi[1], Natasha Medina[1], Miguel Hernandez[1], Natacha Quintero[1], Gilberto Perpiñan[1], Andrea Quintana[1] Philip Warrick[2]

[1]Applied Biophysics and Bioengineering Group, Simon Bolivar University, Caracas, Venezuela
[2]PeriGen. Inc. Montreal, Canada

Abstract

Automated phonocardiogram (PCG) analysis may provide better clinical information to physicians for analyzing and diagnosing different heart abnormalities. However, despite recent advances in PCG analysis methods, it is still a challenging task to extract accurate and useful information from contaminated heart sound recordings. The main objective of this paper is to introduce a new approach for classification of normal and abnormal heart sound recordings using a nested ensemble of algorithms that includes Random Forest, LogitBoost and a Cost-Sensitive Classifier.

The approach consisted of three stages: preprocessing, classification and evaluation. In the preprocessing stage, PCG signals were first downsampled to 1 kHz using a polyphase antialiasing filter. Next, each heart sound was segmented using Springer's improved version of Schmidt's method to identify four states; S1, S2, systole and diastole. Thereafter, 131 features in time, frequency, wavelet and statistical domains were extracted from the entire signal and from the timings of the states. In the classification stage, the meta-classifier was cross validated on the entire training dataset provided by Physionet Challenge 2016. In the evaluation stage, the sensitivity and specificity of the trained algorithm was tested with unseen signals selected randomly by the Challenge testing environment. Experimental results showed that the proposed approach achieved an overall score of 84.48%, ranking fifth. The use of a nested set of ensemble classifier with a combined set of features extracted from different domains helped reduce overfitting and improved classification performance.

1. Introduction

Cardiovascular disease is the number one cause of death in the world. Phonocardiogram (PCG) signals are used for heart disease detection. They contain bioacoustic information reflecting the operation of the heart. Normally, they comprise two distinct activities, namely the first heart sounds S1 and S2, and may contain additional murmurs that indicate heart failure. These murmurs may overlap in time and frequency domains with S1 and S2, augmenting the difficulty of developing a robust PCG signal classifier to discriminate between normal and abnormal. To this end, PhysioNet organized the Challenge2016 to encourage the development of efficient algorithms that can identify whether the subject is healthy or suffers from heart disease[1].

Ensemble machine learning algorithms combine the predictions of several learning models into a single "ensemble" model, with the objective of improving their performance [2]. Common approaches to ensemble learning include bagging, boosting, and stacking, amongst others. Most studies of PCG classification to date have employed a single machine learning algorithm, such as Support Vector Machine (SVM)[3], Artificial Neural Networks (ANN)[4], etc. In this context, it was our aim to contribute a challenge entry based on a rather elementary decision tree classifier, but made more robust by virtue of its nesting within three ensemble classifiers: Random Forests (RF), LogitBoost (LB) and a Cost-Sensitive Classifier (CSC).

The paper is organized as follows. In section 2 the proposed methodology is described. In section 3 the results using PhysioNet-Computers in Cardiology Challenge 2016 datasets are presented and discussed. Finally, conclusions are presented in section 4.

2. Methods

The approach described was developed and validated using the PhysioNet Challenge 2016 datasets [1]. The challenge dataset consists of a collection of heart sound recordings at 2000Hz from 764 subjects/patients, lasting from 5s to just over 120s [1]. The proposed new system is built to classify PCG signals into normal or pathological. Figure 1 outlines the components and signal flow of our approach which consists of three major phases: Preprocessing, Classification and Evaluation.

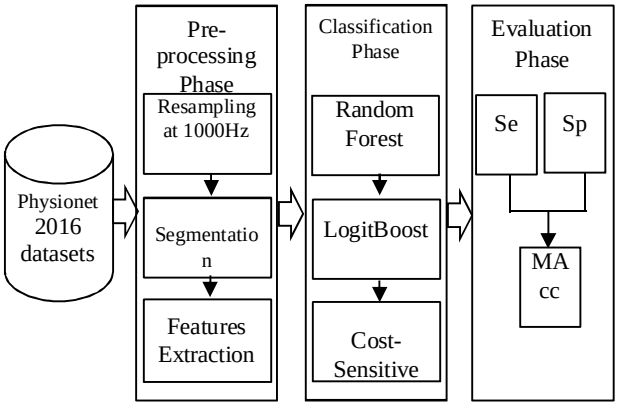

Figure 1. The block diagram of the proposed methodology employed to construct and test the proposed classifier

2.1. Preprocessing phase

During this phase, the PCG signals were first resampled at 1000Hz using a polyphase antialiasing filter. Next, each heart sound was segmented using Springer's improved version of Schmidt's method [5] to identify four 'states'; S1, S2, systole and diastole. Finally, several time domain, frequency domain, statistical features, and features generated by 5-level wavelet decomposition were extracted from the position information of the four states and from the entire signal data. A total of 131 features were obtained and summarized in table 1.

Table 1: List of Features Extracted for Classification

Qty	Feature(s)	Per S1,S2, Dia, Sys	Domain
20	m_RR, sd_RR, mean_IntS, sd_IntS1, mean_IntS2, sd_IntS2, mean_IntSys, sd_IntSys, mean_IntDia, sd_IntDia, m_Ratio_SysRR, sd_Ratio_SysRR m_Ratio_DiaRR sd_Ratio_DiaRR m_Ratio_SysDia sd_Ratio_SysDia, m_Amp_SysS1 sd_Amp_SysS1, m_Amp_DiaS2 and sd_Amp_DiaS2 [1]		Time and statistical
1	HR		Time
4	ZCR	✓	
4	TD	✓	
4	RMS	✓	
4	TotPowT	✓	
4	TotPowF	✓	Frequency
4	BW	✓	
4	Qf	✓	
4	Max	✓	Statistical
4	Mean	✓	
4	Variance	✓	
4	Skewness	✓	
4	Kurtosis	✓	
4	SampEn	✓	
4	SE1	✓	
6	SE2 (5-level wavelet)		Statistical and
24	SE3 (5-level wavelet)	✓	Wavelet
24	SE4 (5-level wavelet)	✓	
131	Total		

As can be observed from the above table, the first 20 features were the same used in [1], while the rest of the features were proposed in this research and explained as below. A checkmark in the third column indicates that the feature was calculated for each of the 4 cardiac cycles (S1, S2, systole and diastole):

- Heart Rate (HR) refers to the number of times a person's heart beats per minutes. An abnormal heart rhythm is when heart beats too fast, slow, or irregularly.
- Zero Crossing Rate (ZCR) represents the rate of sign-changes along an interval. Higher ZCR is expected more frequently for abnormal signals.
- Time Duration (TD) indicates the length of S1, S2, SYS or DIA in seconds. Considerably longer S1 and S2, and shorter SYS and DIA could be expected for abnormal signal than for the normal signal.
- RMS refers to the square-root of mean of square of the waveform.
- Total Power is the total power of the PCG signal. The abnormal signals tend to have higher power than the normal ones, because of the murmur's amplitude. This measure is calculated in time (TotPowT) and frequency domains (TotPowF). Frequency domain calculation is performed using the Fast Fourier Transform (FFT).
- Bandwidth (BW) is the difference between the upper and lower frequencies in an effective set of frequencies. As murmur signals are high in frequency, the upper frequencies of bandwidth will be affected.
- Q-Factor (Qf) describes how under-damped the oscillation in the signal is. Hence, in case of abnormality the Q-factor would increase[4].
- Max refers to the maximum value of the signal. Abnormal signals can get higher or lower maximums depending of the anomalies.
- Mean represents the average value of the signal. A

higher mean is expected for abnormal signals.

- Variance denotes the signal distribution variance.
- Skewness is a measure of the asymmetry of the probability distribution of the signal.
- Kurtosis reflects whether the signal distribution is flat or peaky.
- Sample Entropy (SampEn) is a useful tool for investigating the dynamics of heart rate and other time series, thereby diagnosing abnormal state.
- Shannon Entropy (SE) represents the randomness or unpredictable information present in a signal. This was calculated in several ways either on the PCG signal itself or on a five-level discrete wavelet signal decomposition (i.e., with 5 detail coefficients and 1 approximation coefficient):
  - SE1: mean SE per segment type.
  - SE2: SE of each wavelet coefficient (Daubechies db1 mother wavelet).
  - SE3: mean SE per segment type per wavelet coefficient (Daubechies db4 mother wavelet).
  - SE4: SE of all samples per segment type per wavelet coefficient (Daubechies db4 mother wavelet).

2.2. Classification phase

In the classification phase, a nested set of ensemble classifiers was employed: Cost-Sensitive Classifier (CSC), LogitBoost(LB) and Random Forest(RF). It was trained and tested separately on both Physionet2016 datasets using 10-fold stratified cross-validation.

RF is a meta-learning approach that uses multiple random decision trees as base learners and aggregates them to compute the final ensemble prediction [2, 6]. RF involves sampling of the input data with replacement (bootstrap sampling). In this sampling, about one third of the data is used for testing. These are called the out of bag samples. Error estimated on these out of bag samples is called the Out of Bag error (OOB). An RF has three parameters that can affect its performance:

- Number of features to choose at each node for splitting (NF).
- Number of trees to grow in the forest (NT): Increasing the number of trees in a RF does not result in overfitting.
- Maximum depth of tree (MDT): Higher values generally increase the quality of the prediction, but can lead to overfitting. High values also increase the training and prediction time. If there is no depth limit, the tree is split until each node contains a single target value.

LogitBoost (LB) is a boosting algorithm that was first used with DNA microarray data [6] by Dettling and Bühlmann. LB is less sensitive to outliers and generally gives lower error rates than the commonly used AdaBoost algorithm, which is attributed to its use of logistic regression as the cost functional.

CSC incorporates arbitrary misclassification costs into the learning process. Misclassification penalties are associated with each of the four outcomes of a (binary) confusion matrix, referred to as C_{TP}, C_{FP}, C_{FN} and C_{TN}.

TP (True Positive) is the total number of correct positive classifications, TN (True Negative) is the total number of correct rejections, FP (False Positive) represents the total number of misclassified instances that were incorrectly classified as positive, and FN (False Negative) is the proportion of positive instances that are wrongly diagnosed as negative.

No costs are usually assigned to correct classifications, so C_{TP} and C_{TN} are set to 0. Since the positive class is often (and in our case) more interesting than the negative class, so C_{FN} is generally greater than C_{FP} [6]. Thus, the main objective of CSC is to minimize the expected overall cost as a function of the two error types, given by: $Cost=C_{FP}*FP+C_{FN}*FN$.

2.3. Evaluation phase

The proposed classifier performance was evaluated by using two new metrics proposed by the challenge: modified Sensitivity (Se) and modified Specificity (Sp). The overall score is given by $MAcc=(Se+Sp)/2$ which represents the average of the values of Se and Sp[1].

3. Results and discussions

The Physionet dataset contains 3153 recordings, out of which 2488 recordings are labelled as Normal and remaining 665 recordings are labelled as abnormal. The imbalance ratio between the two classes is therefore 2488/665=3.74. The dataset was first preprocessed to extract 131 features for the proposed classifier. Classifiers were constructed using stratified 10-fold cross validation. Table 2 displays the training and test MAcc scores of 7 entries in ascending order.

The first two entries addressed the problem of class imbalance in two ways. In the first one, the oversampling method SMOTE (Synthetic Minority Over-sampling Technique) [2, 6] was employed, while in the second one, the signal was divided into 3 parts. Each new sub-signal was preprocessed and added as a new instance to the training dataset. In both entries, a LB based on RF classifier was trained and tested using the default parameters given in Weka [6], which are: NF= 8, NT=100, MDT=unlimited and the number of LogitBoost iterations (LB-IT=10) that minimized the root mean squared error. Although these classifiers exhibited the highest sensitivity, specificity and MAcc in training and testing, they yielded the lowest entries scores of 75.7% and 75.6% respectively. Additionally, it can be noted that

there was a large discrepancy between high specificity and lower sensitivity, yielding high false negative rates.

In the last 5 entries, sensitivity-specificity balance was improved and false negatives rates were reduced by assigning a higher cost to false negatives than false positives. Thus, the classifiers' outputs were adjusted by changing their probability thresholds (penalties) from the default value of pt=0.5 to pt= 6/(6+1)=0.8571 in Entry 3 and to pt=8/(8+1)=0.8888 in Entries 4, 5, 6 and 7. The threshold was calculated from the Cost Matrix (CM) as follows: pt=CFP/(CFP+CFN). Although, Entry 7 had greater cost (652) than Entry 6 (600), it yielded our best overall score of all the submitted entries of 84.48%, ranking fifth out of the 48 participants of the challenge, achieving the third highest Se of 88.48%, but with a relatively modest Sp of 80.48%.

Entries 4, 5, 6 and 7 tuned LB-IT to 3 and NF to 43 to address the lower value of OOB calculated for each tree and obtained in that iteration.

In Entry 5, normalized weight was assigned to each instance according to a misclassification cost matrix. That is, instances, which carried a higher misclassification cost, were assigned proportionally higher weights. Instances with higher weight can therefore be viewed as instance duplication. This entry yielded similar results to other CSC classifiers, with sensitivity and specificity of 81.2% and 85.2% respectively.

4. Conclusions

The proposed approach seems promising for classifying heart sound recordings collected from heterogeneous environments. Nevertheless, the performance of the detector strongly depends on the quality of the data. Furthermore, we suggest that the number of examples of the minority class may have been too small for classifiers to learn adequately; the training data may have been insufficient to represent well the plethora of abnormalities attributable to heart disease. Finally, future works will employ pre-processing methods to determine the most discriminating features from our large set and to gain insight into developing more improved features.

References

[1] Chengyu Liu, David Springer, et al, An open access database for the evaluation of heart sound algorithms, Physiological Measurement, 37, 9, 2016.

[2] Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier; 2011.

[3] Gur Emre Guraksin and Harun Uguz, Classification of Heart Sounds Based on the Least Squared Support Machine, ICIC International, 2011,Volume 7, 12, 7131-7144.

[4] Simarjot Kaur Randhawa, Mandeep Singh, Classification of Heart Sound Signals Using Multi-modal Features, Procedia Computer Science, 2015, Volume 58, , 165-171.

[5] DB Springer, L Tarassenko, GD Clifford, Logistic regression-hsmm-based heart sound segmentation, IEEE Transactions on Biomedical Engineering 63 (4), 822-832, 2016.

[6] Ian H. Witten, Eibe Frank and Mark A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers, 2011.

Address for correspondence.

Masun Nabhan Homsi, Universidad Simón Bolívar, Valle Sartenejas, Edo. Miranda, Caracas- Venezuela.
mnabhan@usb.ve

Philip Warrick, PeriGen. Inc. Montreal, Canada
philip.warrick@gmail.com

Table 2: Results of different experiments and entries

#	Classifier	NF	NT	LB-IT	CM	Cost	Training			Testing			Entries Results		
							Se %	Sp %	MAcc %	Se %	Sp %	MAcc %	Se %	Sp %	MAcc %
1	LB+RF	8	100	10	-	-	98.6	93.4	96.0	83.3	94.5	92.2	58.2	93.1	75.7
2	LB+RF	8	100	10	-	-	91.1	99.4	97.7	92.7	99.8	98.2	57.2	94.1	75.6
3	RF+LB+CSC	8	100	10	0,6 1,0	1049	80.0	95.1	91.9	80.2	95.6	92.6	76.5	85.2	80.8
4	RF+LB+CSC	43	120	3	0,8 1,0	1200	93.5	87.1	88.5	92.1	88.9	89.5	82.1	83.2	82.7
5	RF+LB+CSC	43	120	3	0,8 1,0	7052	91.8	88.4	88.9	89.7	89.7	89.7	81.2	85.2	83.2
6	RF+LB+CSC	43	150	3	0,8 1,0	600	93.9	85.5	87.3	94.4	86.3	88.0	84.2	83.2	83.7
7	**RF+LB+CSC**	**43**	**100**	**3**	**0,8 1,0**	**652**	**94.2**	**86.0**	**87.8**	**94.4**	**86.9**	**88.4**	**88.48**	**80.48**	**84.48**