Nonnegative Matrix Factorization and Random Forest for Classification of Heart Sound Recordings in the Spectral Domain

Christoph Hoog Antink[1], Julian Becker[2], Steffen Leonhardt[1], Marian Walter[1]

[1]Philips Chair for Medical Information Technology, RWTH Aachen University, Aachen, Germany
[2]Institut für Nachrichtentechnik, RWTH Aachen University, Aachen, Germany

Abstract

Stimulating the development of robust algorithms for the automated classification of phonocardiograms (PCGs) is the goal of the PhysioNet/CinC challenge 2016. In this paper, an approach to classify PCGs in the spectral domain is presented. First, the magnitude spectrogram is calculated. Next, the spectral shapes of four states of the cardiac cycle ("S_1", "Systole", "S_2", "Diastole") are extracted using nonnegative matrix factorization, which is initialized with a time-domain segmentation algorithm. A Random Forest with 3000 trees is used for classification. Using 10-fold cross-validation on the unbalanced training data, a mean sensitivity of 0.92 at a specificity of 0.83 was achieved, resulting in an overall score of 0.88. On the complete hidden test data, a top score of 0.78 during phase II of the challenge with a sensitivity of 0.74 and a specificity of 0.83 was achieved.

1. Introduction

This year's PhysioNet/CinC challenge aims to stimulate the development of robust algorithms to accurately classify heart sound recordings automatically [1, 2]. Several groups have addressed the problem of frequency-domain classification of the phonocardiogram (PCG). In Bhatikar et al. [3], an artificial neural network (ANN) is used to differentiate between innocent and pathological murmurs based on spectral information. In their study, manual selection and segmentation of individual heart cycles of acceptable quality is performed. Sepehri et al. [4] recorded an Electrocardiogram (ECG) in parallel to the PCG. Five frequency bands in the systolic segment were identified as inputs to an ANN. De Vos et al [5] also recorded PCG and ECG in parallel. The PCG was segmented based on the ECG recording and an ANN was used for classification.

The algorithm presented in this paper combines frequency-domain analysis of the PCG with automated time-domain segmentation [6]. Thus, no signal except the PCG is necessary for its classification.

2. Materials and Method

An overview of the algorithm is given in Figure 1. The

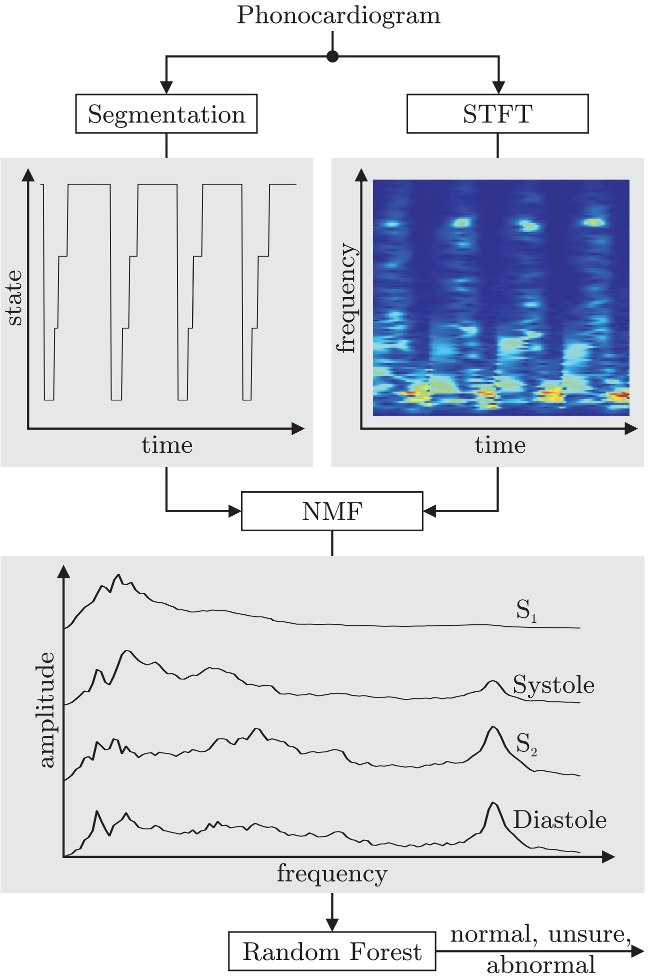

Figure 1. Schematic overview of the algorithm. Nonnegative matrix factorization is used to extract the spectral shapes of the four states of the PCG, which are obtained using Springer's segmentation algorithm. A Random Forest is used to classify the recording.

ISSN: 2325-887X DOI:10.22489/CinC.2016.235-109

PCG is downsampled to 1 kHz and denoted $x(n)$, with n being the discrete time. Springer's segmentation algorithm [6] as provided in the sample code is used to assign the four states 1$\hat{=}$“S$_1$”, 2$\hat{=}$“Systole”, 3$\hat{=}$“S$_2$”, and 4$\hat{=}$“Diastole” to the state vector $x(n)$.

To calculate the spectrogram $S(i, \omega)$ of the PCG, the short term Fourier transform (STFT) is used, in which the Fourier transform is applied to a moving window of $x(n)$. Here, i is the index of the window and ω is the respective frequency bin. To extract the dominant spectral components of the four states, nonnegative matrix factorization (NMF) [7] is used.

NMF approximates a nonnegative matrix $X \in \mathrm{I\!R}_{+}^{I\times\Omega}$ by a product of two nonnegative matrices $W \in \mathrm{I\!R}_{+}^{I\times R}$ and $H \in \mathrm{I\!R}_{+}^{R\times\Omega}$,

$$X \approx \hat{X} = WH. \tag{1}$$

R is a user defined parameter which defines the rank of the approximation $\hat{X}$. W and H are calculated iteratively, minimizing the distance between X and $\hat{X}$. When applied on the magnitude spectrogram, i.e. $X = |S|$, the rows of H can be interpreted as R spectral shapes, that are active at different time instances defined by the R columns of W. Thus, to estimate the spectral shape of the four states (“S$_1$”,“Systole”, “S$_2$”, “Diastole”), R was set to four and W was initialized with the results of the segmentation algorithm.

For machine learning, the ensemble learning method Random Forest (RF) was used [8]. Here, a number of n_{tree} unpruned binary classification trees is learned, each on the basis of a random subset of the training data and a random subset of available features. In the prediction stage, decisions are made by a majority vote. If n_{tree} is large enough, the RF is known to be relatively insensitive to overfitting. This method offers the advantage that the quality of the learned classifier can be evaluated without an additional cross-validation dataset by analyzing the “out-of-bag” error (OOBE). It signifies the misclassification probability and is calculated by evaluating each tree with that fraction of the training data *not* used for its training, i.e. the out-of-bag data. Moreover, the feature importance can be evaluated by randomly permuting data for each feature and measuring the increase in error, ΔOOBE. To train the algorithm, all data available was used. As the training data is heavily biased, countermeasures have to be taken. Instead of creating a balanced training set by excluding data, the assumed prior distribution was manually optimized to maximize the score on the hidden test set.

3. Results and Discussion

To tune the algorithm, the OOBE was optimized. In the final implementation, a Hann-window of 50 milliseconds length and a hop-size of 13 samples is used. At 1 kHz sampling rate, this results in 26 frequency bins per state. As an additional feature, the residual of the NMF is used. Thus, a total of $4 \times 26 + 1 = 105$ features is used. In Figure 2, the OOBE over the number of trees n_{tree} is shown when the complete training set (subsets $a - f$) are used. One can see that the OOBE actually increases for $n_{\text{tree}} < 200$ indicating an inappropriate forest size. For $n_{\text{tree}} > 300$, however, a decrease in OOBE can be observed. To balance run time and accuracy, $n_{\text{tree}} = 3000$ was manually selected in the submitted entry.

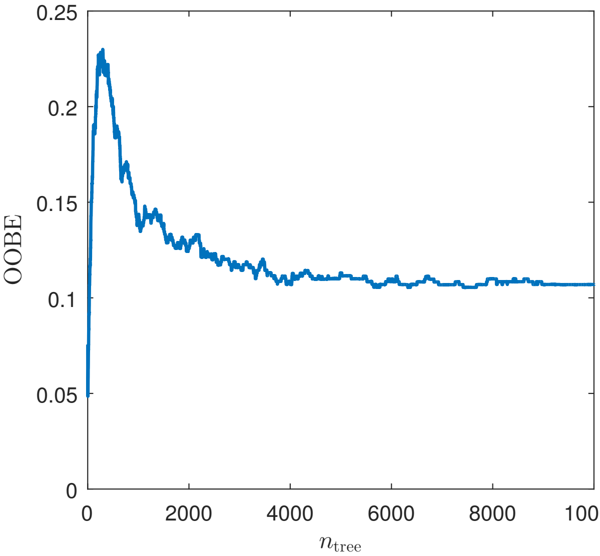

Figure 2. Out-of-bag error over number of trees when using the full training set (subsets $a - f$).

To compensate for biased training data, the RF as implemented in MATLAB allows to set an expected distribution of classes. However, assuming a balanced distribution led to a low sensitivity (Se) at a high specificity (Sp) on both the unbalanced training set as well as the balanced hidden test set. Thus, the priors for normal p(n), unsure p(u), and abnormal p(a) were manually modified as shown in Table 1. If the priors are set to $p(\text{n}) = 14\%$, $p(\text{u}) = 3\%$ and $p(\text{a}) = 83\%$, the maximum score on a balanced, random subset of the hidden dataset is achieved. These priors also resulted in the maximum score on the complete

selected priors			cross-val. training / hidden test set		
p(n)	p(u)	p(a)	Se	Sp	Overall
38%	4%	58%	0.84 / 0.64	**0.92 / 0.91**	0.88 / 0.77
32%	4%	64%	0.86 / 0.66	0.91 / 0.90	0.88 / 0.78
14%	3%	83%	**0.92 / 0.80**	0.83 / 0.83	0.88 / **0.81**

Table 1. Manually selected priors and results from cross-validation on the training data as well as a random subset of the hidden test data.

test set, 0.78, with a sensitivity of 0.74 and a specificity of 0.83. Note that the mean score when using 10-fold cross-validation of the training data is not influenced but that a trade-off between sensitivity and specificity can be observed. Thus, the priors that maximize the score on the hidden test are used in the following.

Table 2 shows the results from 10-fold cross-validation on the complete training set when no balancing is performed. Using this form of evaluation, a mean sensitivity of 0.92 and a mean specificity of 0.83 is achieved. This results in a mean overall score of 0.88.

In Table 3, the results for leave-one-out cross-validation of the training data is presented. Here, each subset of the training data is excluded from training and then predicted using the learned classifier. One can see that results are much lower compared to the 10-fold cross-validation. It is also interesting to note that excluding training-b resulted in a low OOBE. At the same time, the overall score when predicting training-b with a classifier trained with the rest of the data results in a very low overall score. This indicates that training-b is significantly different from the other training sets.

In Figure 3, ΔOOBE for each frequency bin of each state is shown. For the residual, ΔOOBE is -0.6. One can see that for every feature, ΔOOBE is negative and small (-1 to 0), indicating a weak learner. Moreover, it can be seen that frequency bins below approximately 250 Hz and, to a smaller degree, the residual as well as frequency bins above 450 Hz are relatively important for classification.

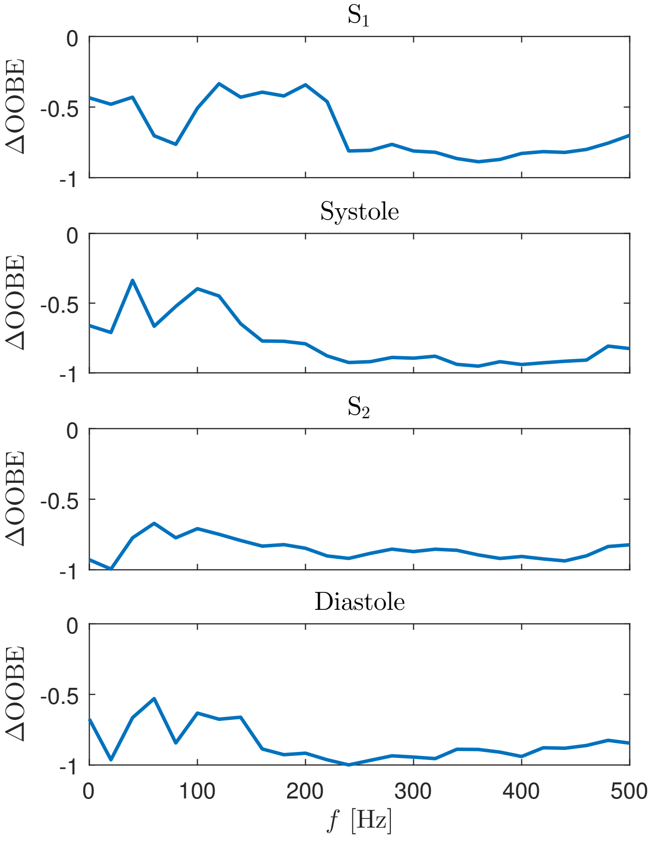

Figure 3. ΔOOBE for each frequency bin of each state.

4. Conclusion and Outlook

First, Tables 1 and 2 demonstrate that the presented approach is indeed feasible to classify heart sound recordings based on their spectral information. In the final configuration, a sensitivity of 0.74 at a specificity of 0.83 was achieved on the complete hidden test data, resulting in an overall score of 0.78. For this, expected distributions of $p(\mathrm{n}) = 14\%$, $p(\mathrm{u}) = 3\%$ and $p(\mathrm{a}) = 83\%$ that almost mirror the actual distribution in the training data had to be chosen. It is interesting to note that these skewed priors had no negative effect on overall score when cross-validating the training data but had an overall positive effect on the balanced hidden test set.

In addition, one can see that the overall score on the hidden test set (0.78) is low compared to the mean score of 0.88 when performing 10-fold cross-validation of the training data. Table 3 delivers a potential explanation as it shows a very low mean score of 0.46 when performing leave-one-subset-out cross-validation. This indicates that the presented approach does not generalize well to data that was recorded using a different scenario. In phase I of the challenge, a modified version of this algorithm achieved a top score of 0.87 ($Se = 0.90$, $Sp = 0.84$), which is close to the cross-validation result. Thus, we suspect that the data added in phase II is (from the point of view of the algorithm) significantly different from training sets a to f. It is also worth noting that the specificity is almost identical in the cross-validation and in the hidden test set results. Thus, the lower score originates from a lower sensitivity. This might indicate that the algorithm misses certain pathologies.

Several measures to improve the algorithm can be taken. First, the segmentation process can be improved. It was observed in the development of this algorithm that even a randomly initialized NMF resulted in temporal excitation vectors W that were correlated with different phases of the cardiac cycle. We thus plan to further examine the potential of the NMF for PCG segmentation. In particular, instead of initializing W, the matrix representing the spectral shapes, H, could be initialized. This way, the occurrence of typical pathological or benign patterns could be detected from analyzing W.

As an alternative, the segmentation algorithm as presented in [6] could be augmented by robust beat-to-beat heart rate estimation [9]. Although originally developed for cardiac vibration signals, the algorithm has since been proven useful on a variety of cardiac signals [10,11]. Since

Fold		test data from subset training-						test data distribution			test data results		
iterate	OOBE	a	b	c	d	e	f	normal	unsure	abnormal	Se	Sp	Overall
1	0.11	46	47	1	6	203	12	233	26	56	0.94	0.83	0.88
2	0.11	35	44	5	6	216	10	234	35	47	0.87	0.82	0.84
3	0.12	34	63	0	5	206	7	231	30	54	0.98	0.84	0.91
4	0.11	34	56	2	8	205	11	237	23	56	0.94	0.83	0.89
5	0.11	44	47	5	8	196	16	240	19	57	0.86	0.81	0.83
6	0.12	46	49	6	2	201	11	221	29	65	0.93	0.84	0.89
7	0.12	39	38	2	6	225	5	233	27	55	0.94	0.91	0.92
8	0.12	43	52	3	5	201	11	221	35	59	0.94	0.83	0.89
9	0.13	43	44	3	6	200	19	226	23	66	0.97	0.82	0.90
10	0.12	45	50	4	3	201	12	226	32	57	0.85	0.80	0.82
Mean	0.12	40.9	49.0	3.1	5.5	205.4	11.4	230.2	27.9	57.2	0.92	0.83	0.88
SD	0.01	5.0	6.9	1.9	1.9	8.7	4.0	6.5	5.3	5.4	0.05	0.03	0.03

Table 2. Results for 10-fold cross-validation on the complete, unbalanced training data.

Excluded		test data results		
database	OOBE	Se	Sp	Overall
training-a	0.18	0.50	0.59	0.54
training-b	0.07	0.22	0.43	0.32
training-c	0.11	0.54	0.29	0.41
training-d	0.11	0.57	0.37	0.47
training-e	0.16	0.91	0.09	0.50
training-f	0.11	0.94	0.04	0.49
Mean	0.12	0.61	0.30	0.46
SD	0.04	0.27	0.21	0.08

Table 3. Leave-one-out cross-validation of the training data.

the duration of the four states of the PCG are depending on the beat-to-beat interval, an improved estimation could ultimately enhance segmentation results.

It should be noted that the classification is developed purely data-driven. Thus, the inclusion of physiological information, such as expected frequency bands in the respective states of the cardiac cycle associated with certain pathological situations, could help to reduce features and improve the generalization capability of the algorithm. Purely data-driven methods for feature reduction such as principal component analysis did not improve classification results.

References

[1] Liu C, Springer D, Li Q, Moody B, Juan RA, Chorro FJ, Castells F, Roig JM, Silva I, Johnson AE, Syed Z, Schmidt SE, Papadaniil CD, Hadjileontiadis L, Naseri H, Moukadem A, Dieterlen A, Brandt C, Tang H, Samieinasab M, Samieinasab MR, Sameni R, Mark RG, Clifford GD. An open access database for the evaluation of heart sound algorithms. Physiological Measurement 2016;37(11):in press.

[2] Clifford GD, Liu CY, Springer D, Moody B, Li Q, Juan RA, Millet J, Silva I, Johnson A, Mark RG. Classification of normal/abnormal heart sound recordings: the PhysioNet/Computing in Cardiology Challenge 2016. Comput Cardiol 2016;43.

[3] Bhatikar SR, DeGroff C, Mahajan RL. A classifier based on the artificial neural network approach for cardiologic auscultation in pediatrics. Artificial intelligence in medicine 2005;33(3):251–260.

[4] Sepehri AA, Hancq J, Dutoit T, Gharehbaghi A, Kocharian A, Kiani A. Computerized screening of children congenital heart diseases. Computer methods and programs in biomedicine 2008;92(2):186–192.

[5] de Vos JP, Blanckenberg MM. Automated pediatric cardiac auscultation. IEEE Transactions on Biomedical Engineering 2007;54(2):244–252.

[6] Springer DB, Tarassenko L, Clifford GD. Logistic regression-hsmm-based heart sound segmentation. IEEE Transactions on Biomedical Engineering 2015;in press.

[7] Lee DD, Seung HS. Learning the parts of objects by nonnegative matrix factorization. Nature 1999;401(6755):788–791.

[8] Breiman L. Random forests. Machine learning 2001; 45(1):5–32.

[9] Brüser C, Winter S, Leonhardt S. Robust inter-beat interval estimation in cardiac vibration signals. Physiol Meas 2013; 34(2):123–138.

[10] Hoog Antink C, Brüser C, Leonhardt S. Detection of heart beats in multimodal data: a robust beat-to-beat interval estimation approach. Physiol Meas 2015;36(8):1679–1690.

[11] Hoog Antink C, Gao H, Brüser C, Leonhardt S. Beat-to-beat heart rate estimation fusing multimodal video and sensor data. Biomed Opt Express Aug 2015;6(8):2895–2907.

Address for correspondence:

Christoph Hoog Antink
Chair for Medical Information Technology
Helmholtz-Institute, RWTH Aachen
Pauwelsstr. 20 / D-52074 Aachen / Germany
hoog.antink@hia.rwth-aachen.de