Project Proposal: Heart sound Analysis

Mutasem Aldmour

Domain Background

Doctors are trained to listen to heart sounds through stethoscopes and look for abnormalities in heart sound, off of which they might ask the patient to do more screening. This simple diagnosis technique is an obvious candidate for machine learning, as an initial diagnosis tool. People can use cheap electronic stethoscopes at home with their smartphones to record their own heart sound and get a recommendation to see a doctor or not worry about it. It could also remove the necessity for doctors to train in this area, especially since they get worse at it over time without enough practice in this critical skill.

There are a few datasets with heart sound recordings. Most of them are aimed at training medical students to learn heart sounds, and as such they are generally too small in size. One of the most interesting attempts at this was through the PASCAL challenge[2] database which resulted in three different solutions[3] to this problem. In this project, I want to work with the biggest dataset that attempted to tackle this problem, the Physionet 2016 challenge dataset.

In fact, my college graduation project in Biomedical Engineering was to prototype a 3d-printed stethoscope plug-in that can be used to turn any normal stethoscope into an electronic recording stethoscope. The goal was to deploy the device in rural India where rural nurses would record and share heart sounds with doctors in the city to make a diagnosis. As an alternative, a machine learning model can give real time recommendation to seek a doctor. One of the challenges I had when doing the analysis on the recorded heart sounds was to find the components S1 and S2 to calculate Heart Rate Variability. Rather than the feature-based approach I used to find them which was very susceptible to noise, machine learning offers a dynamic solution to working with this data.

[1] Liu et al. An open access database for the evaluation of heart sound algorithms. Physiol Meas. 2016 Nov 21;37(12):2181-2213

[2] Bentley P, Nordehn G, Coimbra M, Mannor S and Getz R 2011 The PASCAL classifying heart sounds challenge 2011 (CHSC2011)

[3] PASCAL Challenge Paper 1, 2 and 3

Problem Statement

The main question that I want to answer: can we train a learning model to predict heart problems using recordings of heart sounds as input? Can we classify a heart sound recording as either normal or abnormal? An abnormal classification could mean that a user should follow up with a doctor or to let the doctor know that further investigation should be done. This has the benefit of enabling anyone with a smartphone and the right tool to record their heart sound with to be able to get a quick diagnosis that can let them know to seek further diagnosis. I also want to decrease the number of false negatives that a doctor or a nurse will make over their careers listening to heart sounds.

Dataset and inputs

In this project, I will be using the Physionet heart sound classification dataset[1] used for the 2016 challenge. The Physionet dataset totals to 3126 heart sound recordings, the largest database of heart sound recordings which range from 5 to 120 seconds. These heart rates were sourced from both clinical and non-clinical settings, 8 different databases, and 764 different patients. No two recording was recorded from the same patient in the training set. The database is divided into 5 different sets and each recording is labeled as either normal, abnormal, or unsure (too noisy to tell).

The testing dataset has yet to be released, and as such cannot be used to validate the learning model. As such, I will be using one of the 5 datasets as a testing dataset. I will choose dataset F as the testing dataset, as it represents 3.62% (114) of all the available recordings and 5% of the number of heart beats in all of the recordings. That leaves us with a total of 3039 recordings as part of the training dataset. The dataset is unbalanced with more normal recordings than abnormal. The training set is 73% normal to 18% abnormal, which matches the training set which is 68% normal to 27% abnormal.

I might also use PASCAL challenge dataset[2] which has segmented normal heart recordings into S1 and S2, the two major components of heart sounds, as well as the R-R intervals between heart beats. It also has labeled data into multiple categories of normal and different diseases. Unfortunately, the dataset size is small and has a low-pass filter applied to the recordings, which removes some distinguishing components of heart sounds. However, it could be flattened to be used as more training or testing data.

Segmentation will allow us to derive the features that facilitate classification. Some of these features include median and standard deviation of R-R intervals, median and standard deviation of S1-S1 intervals, median and standard deviation of S2-S2 intervals, ratio of amplitudes between S1 and S2, and ratio of the systolic to diastolic intervals.

Solution Statement

We want to classify heart sound records as either normal or abnormal, indicating that the patient should follow up with a doctor. For this, I will be experimenting with supervised learning techniques, especially Logistic Regression and Ensembles. Logistic Regression, for example, was used as part of the benchmark model used in the Physionet challenge1. As such, I will be measuring accuracy of labels as the key metric.

Furthermore, as part of trying to classify recordings of heart sound, we need to be able to identify the key features of a heart sound wave. A heart sound is comprised of two major components: S1 (lub sound) and S2 (dub sound). This segmentation requires its own learning model that could process each recording and output the key features needed to classify the heart sounds as normal or abnormal. For this, I will be experimenting with unsupervised learning techniques like K-means clustering, which has been used successfully to achieve 92.1% sensitivity and 88.4% positive predictions on 27 recordings from healthy subjects6. Segmentation will also help tell apart noisy recordings, to let the user know that they should try to record again, labeled as unsure.

Benchmark Model

As a benchmark model, I will be using the same benchmark model outlined in the Physionet challenge. This model uses Springer's[5] segmentation code to locate S1 and S2, generating 20 different features for each recording. Moreover, a learning model based on Logistic Regression was used to classify the recordings as normal or abnormal. The open-source code for this model was written in Matlab, so I will have to rewrite it in Python, and apply the same evaluation metrics that will be applied to the final model.

Evaluation Metrics

As evaluation metrics, I will be looking at the accuracy and f1 score of the models on the labeled testing dataset. I will also be calculating Sensitivity (Se) and Specificity (Sp) as outlined by the Physionet paper[1].

Here's a table detailing the classification rules to calculate the score:

	Diagnosis	Signal quality	Percentages of recordings	Challenger report result		
				Abnormal	Unsure	Normal
Reference label	Abnormal (1)	Good (1)	wa_1	Aa_1	Aq_1	An_1
		Poor (0)	wa_2	Aa_2	Aq_2	An_2
	Normal (−1)	Good (1)	wn_1	Na_1	Nq_1	Nn_1
		Poor (0)	wn_2	Na_2	Nq_2	Nn_2

These equations will be used to calculate Sensitivity (Se) and Specificity (Sp):

$$\mathrm{Se} = \frac{wa1 \times Aa1}{Aa1 + Aq1 + An1} + \frac{wa2 \times (Aa2 + Aq2)}{Aa2 + Aq2 + An2}$$

$$\mathrm{Sp} = \frac{wn1 \times Nn1}{Na1 + Nq1 + Nn1} + \frac{wn2 \times (Nn2 + Nq2)}{Na2 + Nq2 + Nn2}$$

Using the following equation, I will score the different learning models, as well as the benchmark model.

$$\mathrm{MAcc} = (\mathrm{Se} + \mathrm{Sp})/2$$

Project Design

The first part of the project is developing the benchmark model in Python and evaluating its performance according to the metrics outlined above. This process will give me insight into working with the dataset. The original benchmark model was only trained on one part of the training dataset and not optimized. I plan to do that, show the results, and then attempt to optimize it and train it on the full training dataset.

The second part will be developing the segmentation strategy. I will use the results of the Springer's segmentation method in addition to some hand labeling to validate the model. The Pascal Challenge dataset could be very helpful with this, as it has identified the S1 and S2 for normal recordings in the dataset. I will be building an unsupervised training model starts with K-means clustering. The goal is to develop a strong set of features that can help classify the heart sound recordings as normal or abnormal, as well as tell apart the noisy recordings.

The third part of the project will be building the learning model using supervised learning techniques. I will first experiment with Logistic Regression and then Ensembles with gradient boosting or other techniques. I will attempt to narrow down the number of features to the ones that weigh the most on the prediction.

References

[4] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23):e215-e220 [Circulation Electronic Pages]; 2000 (June 13).

[5] Clifford GD, Liu CY, Moody B, Springer D, Silva I, Li Q and Mark RG 2016 Classification of normal/abnormal heart sound recordings: the PhysioNet/Computing in Cardiology Challenge 2016 Computing in Cardiology (Vancouver: IEEE) pp 609–12

[6] Chen T, Kuan K, Celi L and Clifford G D 2009 Intelligent heartsound diagnostics on a cellphone using a hands-free kit AAAI Spring Symp. on Articial Intelligence for Development (Stanford University) pp 26–31