Classifying Heart Sounds using SAX Motifs, Random Forests and Text Mining techniques

Elsa Ferreira Gomes
GECAD- Knowledge Eng. Dec. Sup. Res. Center
ISEP/IPP-School of Eng., Polytechnic of Porto, Portugal
efg@isep.ipp.pt

Alípio M. Jorge
LIAAD-INESC TEC
DCC-FCUP, Universidade do Porto, Portugal
amjorge@fc.up.pt

Paulo J. Azevedo
HASLab-INESC TEC
Department of Informatics, University of Minho, Portugal
pja@di.uminho.pt

ABSTRACT

In this paper we describe an approach to classifying heart sounds (classes Normal, Murmur and Extra-systole) that is based on the discretization of sound signals using the SAX (Symbolic Aggregate Approximation) representation. The ability of automatically classifying heart sounds or at least support human decision in this task is socially relevant to spread the reach of medical care using simple mobile devices or digital stethoscopes. In our approach, sounds are first pre-processed using signal processing techniques (decimate, low-pass filter, normalize, Shannon envelope). Then the pre-processed symbols are transformed into sequences of discrete SAX symbols. These sequences are subject to a process of motif discovery. Frequent sequences of symbols (motifs) are adopted as features. Each sound is then characterized by the frequent motifs that occur in it and their respective frequency. This is similar to the term frequency (TF) model used in text mining. In this paper we compare the TF model with the application of the TFIDF (Term frequency - Inverse Document Frequency) and the use of bi-grams (frequent size two sequences of motifs). Results show the ability of the motifs based TF approach to separate classes and the relative value of the TFIDF and the bi-grams variants. The separation of the Extra-systole class is overly difficult and much better results are obtained for separating the Murmur class. Empirical validation is conducted using real data collected in noisy environments. We have also assessed the cost-reduction potential of the proposed methods by considering a fixed cost model and using a cost sensitive meta algorithm.

Keywords

Heart sound classification, motif discovery, time series analysis, SAX, Random Forest, Text Mining, cost analysis.

1. INTRODUCTION

IDEAS14 July 07-09 2014, Porto, Portugal

http://dx.doi.org/10.1145/2628194.2628240.

The automatic identification of cardiac pathologies by analyzing the features of heartbeat audio recordings is a challenging problem due to the variability in patient characteristics, existing noise in recordings and subtle differences between heart conditions. The problem becomes particularly difficult when heart sounds are collected using small mobile devices, such as digital stethoscopes or smart phones, in diverse environments such as hospitals or outdoor population screenings. The ability to automatically classify or support the classification of heart sounds can greatly contribute to the spreading of medical care, especially in regions where the access to specialized units is difficult.

This paper contributes with algorithms and methods that are able to perform the first level of screening of cardiac pathologies both in a hospital environment by a doctor (using a digital stethoscope) and at home by the patient (using a mobile device). Such algorithms do what is called Heart Sound Classification. We propose the use of multi-resolution motif discovery to characterize heart sounds. Motifs are discovered using the MrMotif algorithm [3], a SAX-based approach [10]. Discovered motifs are regarded as terms that can be present or absent in heart sounds. In our previous work we have used an approach inspired in text mining, using terms as attributes and term frequency as values [7]. In this paper we analyze the use of other text mining inspired techniques, namely TFIDF (term frequency inverse document frequency) modeling and bi-grams. In our previous work [7] we had already concluded that Random Forest classifiers were the most successful for these data. We use this technique for class separation. We also perform a cost reduction study using the MetaCost algorithm.

Heart sounds were collected from 312 auscultations gathered using the DigiScope Collector system [9] deployed in the Maternal and Fetal Cardiology Unit of the Real Hospital Português (Recife, Brazil). Each auscultation consists of 6 to 10 seconds recorded for each of the four standard cardiac auscultation spots in children. We have three classes: Normal, Murmur and Extrasystole. Normal (N) has a count of 200, Murmur (M) counts 66 cases and Extrasystole (E) 46. The relative class imbalance makes the problem harder.

Background noise of different types can frequently be found in these data gathered in real-world situations. The differences between heart sounds corresponding to different heart symptoms can also be extremely subtle and challenging to separate. Success in classifying this form of data requires extremely robust classifiers. Despite its medical significance, this is still a relatively unexplored application for machine

learning [1]. The audio files in the data set are of varying lengths, between 1 second and 30 seconds (some have been clipped to reduce excessive noise and provide the salient fragment of the sound).

2. APPROACHES

Heart sound signals of a normal heart have two main components: the first heart sound, $S1$ (or lub), corresponding to the systolic period, and the second heart sound, $S2$ (or dub), the diastolic period [1].

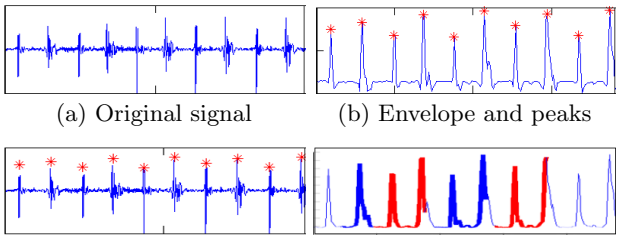
In all approaches we start to preprocess the recorded signals. The original signal was decimated with factor 5. Then, a band-pass filter was applied. Considering the frequency components of $S1$ and $S2$ heart sounds, the chosen filter was a fifth order Chebyshev type I low pass filter. Then, the signals were normalized to the absolute maximum of the signal and the Shannon Envelope of the normalized signal was calculated [8].

In our previous work we tested two very different approaches. The first one was based on peak detection [6]. The process has two steps: segmentation, where $S1$ and $S2$ sound segments are located within audio data and, and classification. In the second approach, we have used SAX-based Multiresolution Motif Discovery for Heart Sound Classification [7]. This technique had never been used in this problem.

Whereas the peak based approach relies on the identification of prominent local maxima in the time series, motif discovery allows the surveying of frequent local patterns (motifs) in the time series, not necessarily peaks. In the case of heartbeat sounds in a cardiac audio, different kinds of sounds should correspond to different motifs. Both approaches lead to the construction of features that characterize sounds and enable the application of powerful general purpose classification algorithms. In our previous work [7] we have concluded that Random Forests [2] gave best results with motifs in terms of average accuracy.

Nevertheless, the segmentation of a heart sound signal is quite difficult since the number of sound components may be different and the existence of anomalies like arrhythmia or murmurs are unpredictable. The periods of heart beat cycles are inconsistent in a phonocardiographic (PCG) signal.

In [6] we presented a methodology for locating S1 and S2 sounds within audio data (Figure 1(a)), segmenting the Normal audio files in the dataset. Our algorithm identified the peaks (S1 and S2) on the envelope calculated using the normalized average Shannon energy [8] (Figures 1(b) and 1(c)).



(a) Original signal (b) Envelope and peaks

(c) Original signal and peaks (d) Motifs over the original

Figure 1: Peaks and one exemplar motif (thicker line) are shown overimposed on the original time-series (x-Time/y-Amplitude).

2.1 Motifs

The extraction of frequent patterns (motifs) from a time series database is an important data mining task. These patterns provide useful insights to the domain expert about the problem at hand [5] and help summarize the time series database. Motif discovery has been used in different areas namely in health and medicine. In particular in EEG time series a motif may be a pattern that usually precedes a seizure [10].

A motif in a time series is a frequently repeated subsequence (frequent pattern). Figure 1(d) shows an example of a motif. There are different approaches for this task. We have followed the Multiresolution Motif Discovery (in Time Series) algorithm [3] to detect the common (and relevant) patterns. This algorithm uses the SAX methodology to discretize the continuous signals and looks for patterns in the resulting discrete sequences. In particular, we have used the MrMotif algorithm as implemented by its authors [7].

2.2 Text Mining

The characterization of sounds using motifs is done by counting the number of occurrences of each motif in each sound. In table 1 we can see an example of a small data set. Each attribute M_i is a top-5 motif of resolution 4. Values are frequencies of motifs. The last column is class.

M1	M2	M3	M4	M5	Class
1	0	5	2	0	N
7	6	4	4	5	M
2	0	2	2	0	E

Table 1: Sample of a resulting dataset of resolution 4, Top-5.

Using an analogy with text mining, each motif can be regarded as a term, and each sound as a document. What we have done so far is to model the sounds using term frequency (TF). Extending the text mining analogy, we now use the TFIDF model (term frequency, inverse document frequency). Given a motif m and a sound s from a collection $Sounds$, $TFIDF(m, s)$ is calculated by multiplying relative term frequency (the frequency of m in s divided by the number of sounds N) by $IDF(m, s)$. This latter part penalizes motifs that are too frequent and hence not useful as discriminants.

$$TFIDF(m,s) = TF(m,s).log\frac{\#\{s \in Sounds\}}{\#\{s \in Sounds : m \in s\}} \quad (1)$$

Another idea brought from text mining is to exploit sequential occurrences of terms in documents. A sequence of size n is an n-gram. A special case are bi-grams for $n = 2$. Frequent n-grams correspond to potentially relevant term compositions. For example, "New York" is a relevant bi-gram that can be automatically identified from documents where it occurs frequently.

In this paper we exploit the ability of these text mining inspired approaches to provide information for the task of separating classes and for reducing screening costs. In our previous work [7] we have shown that the motif based approach yields best results on this heart sounds data set than the approach based on peak location and characterization.

Here we compare three different approaches based on motifs. The basic one is TF, where each sound is characterized by the frequency of relevant motifs found. The two alternatives are TFIDF, where we use the "term frequency - inverse document frequency" model, and TF with bi-grams, where we find common pairs of motifs occurring in sequence (bi-grams). In this latter case, each motif or bi-gram based feature is characterized by plain frequency.

3. EXPERIMENTS

We have used the Weka API and the Weka explorer to run the experiments from our motif-based feature generator program. Here we only show best results of classifier per data set. In the first set of experiments we assess the ability of TF, TFIDF and bi-gram representation for class separation. Then we show an exemplar study of cost reduction using a cost sensitive meta algorithm. In all cases we use Random Forest as base classifier since this technique has consistently given best results on these data with the motif based approach [7].

3.1 Separating classes

The first series of experiments measures how each approach is able to separate classes. We did not consider here class size or importance. These experiments explore the ability of the motif based techniques to separate the Normal class from Murmur, Normal from non-Normal and all three classes at once. The separation of Normal from Extrasystole yield no positive results, due to the large similarity between these two classes. We have performed 10 times 10 fold cross validation on each data set and each classifier set up.

3.1.1 *Normal vs Murmur*

Murmur is the largest non-normal class. Here we consider only the cases from the data set belonging to the classes Normal (200 cases - 75.2%) and Murmur (66 cases - 24.8%). In table 2 we show best results for each approach for the accuracy performance measure.

Data set	Method	Acc.
TF	RF(200,6)	84.77
TFIDF	RF(100,3)	84.81
TFWBG	RF(200,6)	84.89

Table 2: Best accuracy results Normal vs Murmur.

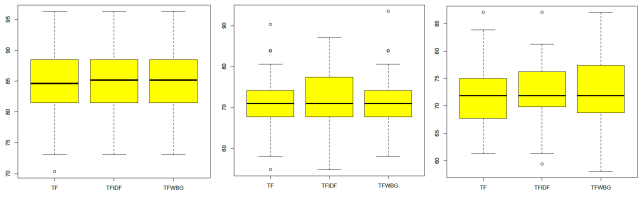


Figure 2: Distribution of accuracy values for 10×10 cross validation. Left: Normal vs Murmur; Center: Normal vs not Normal; Right: Normal vs Murmur vs Extrasystole

The parameters of the random forest algorithm were chosen by systematic grid search in the number of trees and number of attributes. We can observe an increase in the accuracy with TFIDF and bi-grams. In the left image of Figure 2 we can see the distribution of accuracy values over the 100 cross validation runs. Although the accuracy mean increases with the use of TFIDF and bi-grams, we observe no significant differences (Wilcoxon signed rank test).

3.1.2 *Normal vs non-normal*

Here we consider all the cases but join the two non-normal classes as one. Class Normal has 200 cases (64.1%) and Non-Normal has 112 cases (35.9%). In table 3 we show best results for each approach for the accuracy performance measure.

Data set	Method	Acc.
TF	RF(100,3)	70.81
TFIDF	RF(100,2)	71.65
TFWBG	RF(100,2)	71.84

Table 3: Best accuracy results Normal vs non-Normal.

As with Normal vs. Murmur results improve slightly but steadily from TF to TFIDF and then with bi-grams. In the center of Figure 2 it is visible that TFIDF folds' results spread to a region of higher accuracy values. Performing a Wilcoxon signed rank test we have statistical significance for a confidence level of $\alpha = 0.1$. The p-values are 0.066 (TF vs. TFIDF) and 0.005 (TF vs. bi-grams).

3.1.3 *All vs all*

In the situation where all classes are presented as separate we have again a progression from TF to TFIDF and then to bi-grams (table 4).

Data set	Method	Acc.
TF	RF(200,3)	72.37
TFIDF	RF(200,3)	72.54
TFWBG	RF(200,6)	72.79

Table 4: Best accuracy results with all the classes.

Figure 2 (right) shows that distributions of accuracy values are visibly different in the three situations. Kolmogorov-Smirnov test to compare the distributions indicates significant differences ($\alpha = 0.1$) in TF vs TFIDF and TF vs bi-grams. The Wilcoxon signed-rank test, however, shows no significant differences between the means.

3.2 Minding costs

Now that we have assessed the ability of the different motif based approaches to separate classes, we investigate their potential impact in cost reduction, given that different misclassification errors may have different consequences. In the case of heart disease screening, the cost of classifying a normal case as non-normal implies in further more expensive exams. The cost of classifying as normal what is not implies in patient suffering or even human lives. In any case, a late diagnosis will have higher costs with a multiplication of more expensive exams and medical procedures. The quantification of such costs is a complex matter that depends highly on the local characteristics of the screening operation and usually requires specific investigation [11]. In this section we hypothesize a plausible cost model that assumes a

higher cost of missing non normal cases and compare the three motif/text mining based approaches.

In our experiments we will use one cost model that assumes that the cost of misclassifying a non-normal case is 10 times higher than the opposite. To estimate the overall costs for each model we use 10 fold cross validation. To deal with different class costs we employ the MetaCost meta classifier [4] with Random Forest as base classifier. Here we consider only the task of separating Normal and Murmur.

Data set	Method	FN	FM	Cost
TF	RF(200,6)	36	5	365
TFIDF	RF(100,3)	37	4	374
TFwBG	RF(200,6)	35	5	365
TF	MC+RF(200,2)	10	93	193
TFIDF	MC+RF(300,2)	10	91	191
TFwBG	MC+RF(300,3)	10	92	192

Table 5: Cost model analysis. Key: FN: False Normal, FM: False Murmur. $Cost = 10 \times FN + FM$. MC: Meta cost.

The best models identified above for class separation are not concerned with the cost of different classes. As we can see in table 5 costs are quite high for accuracy optimizing models. Using Meta cost to find the model that minimizes cost leads to a cost reduction of about 50% as we can see in the last 3 rows of the table. Note however that the optimal Random Forest parameters are also different from the three first rows. The differences between the three motif approaches are residual. Note also that the cost for the trivial solution of classifying everything as Murmur would be 200. We have therefore an advantage of 9 $(200 - 191)$ cost units per 266 cases (Normal + Murmur). For a practical example, if each cost unit is 100€(the cost of a more expensive screen of the detection of Murmur) and we screen a population of 5000 people, we have an estimated cost reduction of around 17000€.

For the case of the other variants of the dataset, we have not obtained more than marginal cost reductions. Therefore we focused on the Normal vs. Murmur case. Note also the Murmur is the largest non-normal class.

4. CONCLUSION AND FUTURE WORK

In this paper we have presented results of the application of frequent motif based algorithms and methods that are able to perform the first level of screening of cardiac pathologies both in a hospital environment by a doctor and at home by the patient, using a relatively inexpensive mobile device. We proposed the use of multi-resolution motif discovery to characterize heart sounds. Heart sounds are previously preprocessed with standard signal processing procedures. Motifs are discovered in sounds using the MrMotif algorithm [3], a SAX-based approach [10]. Discovered motifs are regarded as terms that and their expression in heart sounds can be quantified by motif frequency (term frequency). In this paper we extended the term frequency (TF) study to other text mining inspired techniques, namely TFIDF (term frequency inverse document frequency) and bi-grams. We have used one real data set and performed several experiments. For class separation we have used 10 times 10 fold cross validation and statistical validation. We have also performed cost analysis using 10 fold cross validation and a proposed cost model. We have observed some significant advantage for class separation by using either TFIDF or bi-grams instead of the simple TF approach. Cost reduction can be achieved for the main non-normal class.

We will continue the exploration of motif based heart sound characterization and optimize the combined effect of signal pre-processing techniques with motif discovery. The discovery of relevant motifs highly depends on the noise filtering performed on the signal. Signal over-smoothing may hide important traces of heart sounds and make them untraceable by motif characterization. This would explain the difficulty with the Extra-systole class, which is very hard to detect by any technique applied on this data set.

Acknowledgements: Portuguese Funds through the FCT - Fundação para a Ciência e a Tecnologia (proj. FCOMP-01-0124-FEDER-037281 and FCOMP-01-0124-FEDER- PEst-OE/EEI/UI0760/2014).

5. REFERENCES

[1] P. Bentley, G. Nordehn, M. Coimbra, and S. Mannor. Classifying Heart Sounds Challenge. www.peterjbentley.com/heartchallenge, 2011.

[2] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[3] N. Castro and P. J. Azevedo. Multiresolution Motif Discovery in Time Series. In *SDM*, pages 665–676, 2010.

[4] P. Domingos. Metacost: A general method for making classifiers cost-sensitive. In U. M. Fayyad, S. Chaudhuri, and D. Madigan, editors, *KDD*, pages 155–164. ACM, 1999.

[5] P. G. Ferreira, P. J. Azevedo, C. G. Silva, and R. M. M. Brito. Mining approximate motifs in time series. In *Discovery Science*, pages 89–101, 2006.

[6] E. F. Gomes, P. J. Bentley, E. Pereira, M. Coimbra, and Y. Deng. Classifying heart sounds - approaches to the pascal challenge. In D. Stacey, J. Solé-Casals, A. L. N. Fred, and H. Gamboa, editors, *HEALTHINF*, pages 337–340. SciTePress, 2013.

[7] E. F. Gomes, A. M. Jorge, and P. J. Azevedo. Classifying heart sounds using multiresolution time series motifs: an exploratory study. In B. C. Desai, A. M. de Almeida, and S. P. Mudur, editors, *C3S2E*, pages 23–30. ACM, 2013.

[8] H. Liang, S. Lukkarinen, and I. Hartimo. Heart sound segmentation algorithm based on heart sound envelogram. In *Computers in Cardiology 1997*, pages 105–108, Sep 1997.

[9] D. Pereira, F. Hedayioglu, R. Correia, T. Silva, I. Dutra, F. Almeida, S. Mattos, and M. Coimbra. Digiscope - unobtrusive collection and annotating of auscultations in real hospital environments. In *Eng. Med. Bio. Soc., EMBC, 2011 An.Int.Conf.of the IEEE*, pages 1193 –1196, 30 2011-sept. 3 2011.

[10] D. Yankov, E. J. Keogh, J. Medina, B. Y. chi Chiu, and V. B. Zordan. Detecting time series motifs under uniform scaling. In P. Berkhin, R. Caruana, and X. Wu, editors, *KDD*, pages 844–853. ACM, 2007.

[11] M. S. Yi, T. R. Kimball, J. Tsevat, J. M. Mrus, and U. R. Kotagal. Evaluation of heart murmurs in children: cost-effectiveness and practical implications. *Pediatrics*, 141:504–522, Oct 2002.