Predicting Heart Beats using Co-occurring Constrained Sequential Patterns

Shameek Ghosh[1], Mengling Feng[2,3], Hung Nguyen[1], Jinyan Li[1]

[1]University of Technology Sydney Australia
[2]Massachussets Institute of Technology, Cambridge, USA
[3]Institute for Infocomm Research, Singapore

Abstract

The aim of this study is to develop and evaluate a robust method for heart beat detection using a sequential pattern mining framework, based on the multi-modal Physionet 2014 challenge dataset. Each multi-modal patient time series was initially transformed to a symbolic sequence using Symbolic Aggregation Approximation (SAX). A training set was created, by randomly selecting 70% of the data and the rest 30% was used as the test set. Later, all segments of length 100 were extracted, for annotated beat occurrences. Subsequently, an algorithm was used to extract repetitive frequent subsequences, where consecutive symbols are separated by a pre-defined gap range. The patterns for ECG and BP were then ranked based on length and frequency support. For tests, the highest ranked patterns were used to mark beat segments. True beat occurrences were only considered when patterns co-occurred for both ECG and BP within a width of 150 time points. Our results comprise two parts viz. extracted top ranked sequences and gross test statistics. An interpretive highest ranked sequential pattern for ECG looks like [7,7,7,5,5,5,5,5,4,3,10,10,10,2,2,3,3,4,3,4,5,5,5,6,7], for 10 discrete symbols which identify regional signal activity, with a gap range of [2,4] between contiguous elements. As per our test results, the method gives us a sensitivity of 51.66% and a positive predictivity (PPV) of 67.15%. The novelty of mining gap constrained co-occurring frequent sequential patterns lies in its ability to capture approximate co-occurring long clinical episodes across multiple variables, even if the quality of one signal suffers for a certain period of time. A higher PPV indicates that our method did not have a lot of false positives (detecting non-beats). The method is still being improved and will be further tested in the next stages of the Physionet Challenge 2014.

1. Introduction

Heart beat patterns in an ECG have traditionally been identified using the popular P-QRS-T waveform cycle. In this context, early detection of heartbeats is an important problem, owing to its applicability in the identification of irregular heartbeats or while differentiating normal from abnormal beats. Yet, a QRS cycle just by itself is not enough to always detect a beat signature, if the ECG signal is noisy and scrambled. Towards this aspect, the 14th Physionet/Computing in Cardiology 2014 challenge was instituted to detect heartbeats based on a multi-modal dataset, for a set of patients. The objective of the challenge was aimed at the exploration of robust techniques for detecting heartbeats from multi-modal data. Each clinical record consisted of 10 minute excerpts of four to eight signals, which among others included data from ECG, blood pressure (BP) and EEG. Moreover, the records in the multi-modal training dataset were also annotated for occurrences of heart beats based on expert opinion.

In recent literature, numerous studies have been reported on the detection of heartbeats using ECG features like morphological signatures, frequency and interval features, neural networks and support vector machines [1-3]. Moreover, there have been some studies related to the mining of sequential time series motifs in ECG signals, which were predictive of cardiovascular diseases [5-6, 12]. Although a number of methods have been proposed using statistical features, we intended to explore the heart beat detection problem using a sequential pattern mining framework. In this context, the proposed methodology consisted of a number of stages involving pre-processing of data, mining of sequential patterns and the employment of these patterns for the final prediction of heart beat segments. In the following sections, we describe our methodology for mining sequential patterns, and our test statistic results based on a test sets, created from the existing training set.

2. Methodology

The sequential pattern mining methodology consists of mainly three stages. In the first stage, continuous time series signals are transformed into a symbolic representation. In the second stage, the collection of known heart beat segments is used for mining frequently occurring sub-sequences. In the third stage, the set of all patterns are ranked to obtain relevant sequential patterns based on increased symbol activity in a given pattern.

2.1. Data preprocessing

For the purposes of finding frequent sub-sequences, the ECG and BP time series records for each patient were initially pre-processed based on a symbolic discretization method known as symbolic aggregate approximation (SAX) [4]. SAX performs symbolization of continuous time series by transforming the data into a piecewise aggregate representation and later converting the same into a symbolic string. As a result, numerous data mining algorithms which operate on discrete representations can be applied to SAX based symbolic sequences to extract interesting sequential patterns. A SAX representation would thus be a transformed sequence of symbols as shown in Figure 1. X and Y are cut-points separating regions which are denoted by P, Q and R. The given signal is thus represented as QQRQPQR.

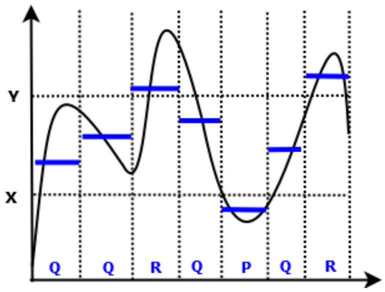

Figure 1. SAX approximation of a continuous time series signal

Following the construction of symbolic sequences for each patient record in the training dataset, reference heart beat annotations are used to extract symbolic segments of a given length, say *k*. This process is carried out by selecting each reference heart beat annotation and extracting the corresponding segment at the concerned position, such that the given annotation is in the middle of the extracted segment. This is repeated for both the ECG and the BP symbolic representations for all the patient records. Subsequently, the extracted segments are collected together to build a new set. Finally, we obtain a set of known heart- beat specific symbolic segments for both ECG and BP, respectively.

2.2. Mining gap constrained sequential patterns

In the past, there have been a several studies detailing various methods of generating frequent sequential patterns from a set of sequences [8, 10, 11]. In this context, sequential pattern mining aims to extract a set of significant sub-sequences based on the concept of frequency support in different types of data such as time series, transactions and sequences. Relevant definitions are provided next.

For a set of symbols (also known as an alphabet) given as $\boldsymbol{I} = \{i_1,i_2,...i_n\}$, a *sequence* is an ordered list of itemsets like $s=\{s_1,s_2,....,s_n\}$ where s_i belongs to $\boldsymbol{I}$. A sequence $S_a = (a_1,a_2,...,a_n)$ is said to be a *sub-sequence* of $S_b = (b_1,b_2,...b_m)$ if there exists integers $1\leq i_1\leq i_2 \leq ... \leq i_n \leq m$, such that $a_1 \subset b_{i1}, a_2 \subset b_{i2},...,a_n \subset b_{in}$. Thus for a given a set of symbols as $P= \{a, b, c\}$, *ab* is a subsequence of *acbc*, but not *ba*. Now, let us consider $\boldsymbol{D} = \{d_1, d_2,...., d_n\}$ to be a set of sequences in a database. Given D, if there exists a subsequence *P*, such that *P* is found in '*k*' number of entries in **D**, then '*k*' is defined as the frequency support of *P*. Moreover, it is not necessary for all symbols in *P* to occur consecutively. Instead a maximum gap constraint denoted as '*g*', allows the algorithm to search for a sequence, such that consecutive elements in P can be distant from each other in the matched sequence entry, up to a maximum value of *g*. Thus for example, if *g*=2, then *'ab'* is a subsequence of *'acb'* but not *'acccb'*.

For the purpose of mining frequent gap-constrained sub-sequences, we employed the ConSGapMiner algorithm, which can be used to extract sub-sequences with user-defined gap constraints [8]. It involves growing a set of candidate sub-sequences in the form of a prefix-based lexicographic sequence tree [8], while recording the frequency support of each candidate sequence. Moreover, the generation of a candidate subsequence is based on an important condition. If there exists a candidate sequence which does not satisfy the minimum user-defined frequency support threshold condition, then the concerned sequence need not be extended. This is because the descendants of the concerned sequence (symbolic extension of the sequence when considered as a prefix) are also expected to be infrequent [8]. In addition, the gap constraint in a sequence is verified by the bitmap checking procedure [7]. Further, details on satisfying gap-constraint based on bitmaps can be found in [7].

2.3. Ranking patterns

Pattern mining algorithms typically generate a significant number of patterns as part of reporting the list of patterns which satisfy gap constraints and frequency support. In this context, there have been numerous interestingness measures, which have been used for ranking of patterns [9]. To minimize the number and identify a suitable set of patterns, we used the following measure to rank the patterns.

$$\rho = \frac{Number\ of\ Distinct\ Symbols}{Pattern\ Length}$$

The rationale behind using the given measure was to identify patterns with more number of symbols indicating increased activity. In addition, before applying ρ to rank the list of patterns, we initially extracted patterns with a length of at least 10 as part of the sequential pattern mining algorithm.

An illustration of the different stages along with the pattern mining framework is given in Figure 3.

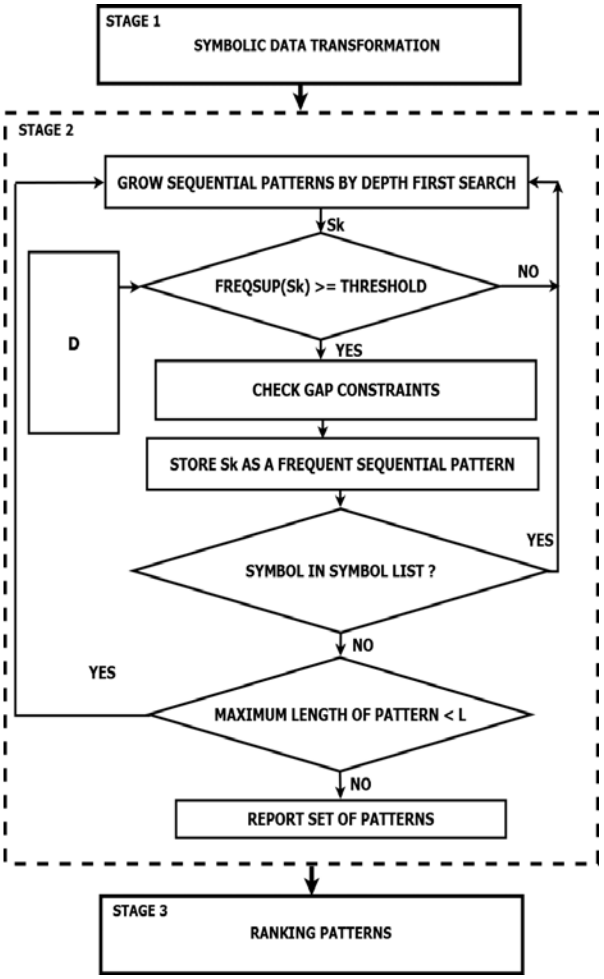

Figure 2. Stages of extracting sequential gap-constrained patterns briefly explained in section 2.

3. Results and discussion

The above described methodology was utilized to extract sets of frequent gap-constrained sequential patterns, for both the ECG and BP symbolic segments. Towards this purpose, we employed 70% of the Physionet 2014 Challenge dataset for training and 30% as the testing set. A sliding window of length 100 was used to shift through the entire length of a test record, while marking potential beats (for both ECG and BP sequences in a test record).Thus, if a BP or ECG pattern was found in the test segment, then the segment mid-point would be marked as 1 at the corresponding position in the original record for ECG or BP. Next, the window would move forward by a value of 30. The difference of 30 was considered based on the maximum possible length of a sequential pattern. The ECG and BP pattern matching step would be repeated for the sliding window segments. Finally, after both BP and ECG records were marked, a true beat was considered only when both ECG and BP sequences were marked as '1' within a maximum difference of 150 time points. Table 1 shows the parameters used for extracting sequential patterns from the training set.

Table 1. Algorithmic Parameters for Sequential Patterns

Parameters	Value
Extracted Beat Segment Length	100
Number of SAX Symbols	10
Gap Range	[2,4]

Table 2 provides the interpretations provided to the definitions of TP, FN and FP in the given challenge. The measures employed as per the challenge were sensitivity (Se) and positive predictivity (+P). These are defined as below.

$$Se = 100 \cdot TP / (TP + FN)$$
$$+P = 100 \cdot TP / (TP + FP)$$

Table 2. Definitions of TP, FN and FP

Term	Meaning
TP	Number of correctly detected beats
FN	Beats missed
FP	Detection of non-beats

In Table 3, we report the gross and average test statistics as required by the challenge, for our test dataset.

Table 3. Test Results

Test Statistics	
Average Sensitivity	51.66%
Gross Sensitivity	65.3%
Average positive predictivity	67.15%
Gross positive predictivity	72.1%

Although the set of generated patterns was extremely large, some of the top ranked patterns were significant to be ECG and BP signatures, which were used to annotate heart beats. The highest ranked ECG and BP sequential patterns are listed in Table 4. In this context, earlier studies have emphasized the importance of longer subsequences in building better models [12]. ECG and BP sequential patterns can thus help interpret detect of instances of transitions that are positively associated with various clinical events.

Table 4. Sequential Pattern examples for ECG and BP

Physiologic Variables	Sequential Pattern
ECG Pattern	[7,7,7,5,5,5,5,5,4,3,10,10, 10,2,2,3,3,4,3,4,5,5,5,6,7]
BP Pattern	[6,5,4,4,3,3,3,3,2,2,3,4,5,6,7,8]

Figure 3 provides a visual interpretation of the ECG pattern, indicating the interpretive capability of a sequence of clinical events. In this context, visual depictions of sequential patterns could be extremely useful to a clinician, for identifying differences in the types of heart beats.

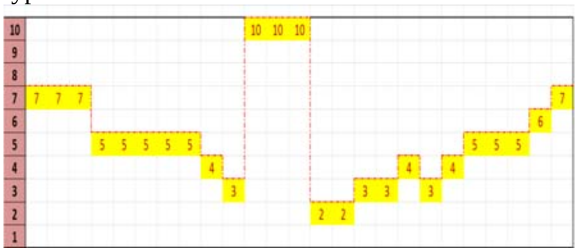

Figure 3. Example of a Top-Ranked Long Range ECG Sequential Pattern

4. Conclusion

In this study, we presented the application of a gap-constrained sequential pattern mining methodology to obtain frequent sub-sequences for annotating heart beat segments, using both ECG and BP. Towards this aspect, we employed the SAX discretization technique to discretize the continuous ECG and BP series into a symbolic form. Later, sequential patterns with an increased density of symbols were considered as more relevant and ranked for predicting heart-beats. As future work, more physiological variables may be used in a clinical record, for applying the sequential pattern mining framework, apart from ECG and BP. Moreover, finding relevant patterns from a given list, turns out to be an important problem in a clinical context and more suitable interestingness measures could be applied.

References

[1] Ye C, Kumar BV, Coimbra MT. Heartbeat classification using morphological and dynamic features of ECG signals. IEEE Transactions on Biomedical Engineering 2012; 59: 2930-2941.

[2] Osowski S, Hoai LT, Markiewicz T. Support vector machine-based expert system for reliable heartbeat recognition. IEEE Transactions on Biomedical Engineering 2004; 51: 582-589.

[3] Osowski S, Markiewicz T, Hoai LT. Recognition and classification system of arrhythmia using ensemble of neural networks. Measurement 2008; 41: 610-617.

[4] Lin J, Keogh E, Lonardi S, Chiu B. A symbolic representation of time series, with implications for streaming algorithms. In Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery 2003; 2-11.

[5] Syed Z, Guttag J, Stultz C. Clustering and symbolic analysis of cardiovascular signals: discovery and visualization of medically relevant patterns in long-term data using limited prior knowledge. EURASIP Journal on Applied Signal Processing 2007; 1: 97-97.

[6] Syed Z, Stultz C, Kellis M, Indyk P, Guttag J. Motif discovery in physiological datasets: a methodology for inferring predictive elements. ACM Transactions on Knowledge Discovery from Data (TKDD) 2010; 4:2.

[7] Ayres J, Flannick J, Gehrke J, Yiu T. Sequential pattern mining using a bitmap representation. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining 2002; 429-435.

[8] Ji X, Bailey J, Dong G. Mining minimal distinguishing subsequence patterns with gap constraints. Knowledge and Information Systems 2007; 11: 259-286.

[9] Ohsaki M, Abe H, Tsumoto S, Yokoi H, Yamaguchi T. Evaluation of rule interestingness measures in medical knowledge discovery in databases. Artificial Intelligence in Medicine 2007; 41: 177-196.

[10] Li C, Wang J. Efficiently Mining closed subsequences with gap constraints. In SDM 2008: 313-322.

[11] Li C, Yang Q, Wang J, Li M. Efficient mining of gap-constrained subsequences and its various applications. ACM Transactions on Knowledge Discovery from Data (TKDD) 2012; 6: 2.

[12] McMillan S, Chia C-C, Esbroeck AV, Rubinfeld I, Syed Z. ICU Mortality prediction using time series motifs. Computing in Cardiology 2012; 39:265-268.

Address for correspondence:

Jinyan Li
Advanced Analytics Institute, University of Technology Sydney (UTS), NSW, Australia,
Jinyan.Li@uts.edu.au