2013 IEEE 13th International Conference on Data Mining

Extraction of Interpretable Multivariate Patterns for Early Diagnostics

Mohamed F. Ghalwash*§, Vladan Radosavljevic* and Zoran Obradovic*
*Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, USA.
Email: mohamed.ghalwash, vladan, zoran.obradovic@temple.edu
§Mathematics Department, Faculty of Science, Ain Shams University, Cairo, Egypt.

***Abstract*—Leveraging temporal observations to predict a patient's health state at a future period is a very challenging task. Providing such a prediction early and accurately allows for designing a more successful treatment that starts before a disease completely develops. Information for this kind of early diagnosis could be extracted by use of temporal data mining methods for handling complex multivariate time series. However, physicians usually prefer to use interpretable models that can be easily explained, rather than relying on more complex black-box approaches. In this study, a temporal data mining method is proposed for extracting interpretable patterns from multivariate time series data, which can be used to assist in providing interpretable early diagnosis. The problem is formulated as an optimization-based binary classification task addressed in three steps. First, the time series data is transformed into a binary matrix representation suitable for application of classification methods. Second, a novel convex-concave optimization problem is defined to extract multivariate patterns from the constructed binary matrix. Then, a mixed integer discrete optimization formulation is provided to reduce the dimensionality and extract interpretable multivariate patterns. Finally, those interpretable multivariate patterns are used for early classification in challenging clinical applications. In the conducted experiments on two human viral infection datasets and a larger myocardial infarction dataset, the proposed method was more accurate and provided classifications earlier than three alternative state-of-the-art methods.**

Keywords*—*early classification; multivariate time series; early diagnosis; interpretability; pattern extraction;

I. Introduction

Providing correct and timely diagnosis saves lives [1]. The severity of many life-threatening diseases can be greatly reduced by administering treatment before the diseases have fully manifested. In addition, early diagnosis may also result in less-intensive therapy, less time spent recovering, and lower cost of treatment. However, making accurate early diagnoses is a challenging problem. Many diseases (such as primary thyroid lymphoma, Alzheimer disease, acute lymphoblastic leukemia [2]) have heterogeneous presentations where a variety of genetic defects cause the same disease, which conventional diagnostic criteria often fail to recognize. The quality of diagnosis can be significantly improved by using the information from multivariate data collected from the patient over time. However, utilization of those complex high-dimensional temporal data to support clinical decisions is still not fully implemented since physicians lack the tools to extract relevant clinical information in a timely manner. Extracting useful early temporal patterns and building accurate predictive models on such data provides a great challenge for the data mining community.

To successfully implement a data mining method in applications related to medical decisions, it is necessary to consider some specific requirements. In medical applications, one of the most desirable properties of a data mining method is interpretability. Physicians tend to prefer methods where it is clear exactly which factors were used to make a particular prediction [3]. Moreover, physicians prefer to have a concise list of temporal patterns to predict patients' health. For example, as noted by [4], "*the occurrence of hypotension episodes taking place during dialysis treatments when arterial blood pressure decreases, the organism reacts with an increase in the heart rate, which then goes back to normal values as soon as blood pressure increases*". Therefore, interpretable methods are much more convincing than "black-box" methods, especially when only a few key factors are used, and each of them is meaningful.

Another important aspect to consider is the need for predictions to be provided as early as possible. The patient is observed for few time points and then as soon as the match between the patient's data and any of the extracted patterns is found, the decision is provided immediately so that the physician can take an appropriate diagnosis and save the patient's life.

In this work, we develop an optimization-based approach for extracting multivariate Interpretable Patterns for Early Diagnostics (IPED) of time series data that provides the predictions. The contributions of our paper are the following:

- We formulate and solve a novel convex-concave optimization problem to a extract large set of patterns from multivariate time series data. Our method allows the patterns to appear at different time points in different dimensions and to be of different lengths.
- We formulate and solve a novel discrete optimization problem to extract a few key patterns from previously extracted large set of patterns. The dimension of each of these key patterns is less than the dimension of the observed multivariate time series. Therefore, the method is able to identify the relevant dimensions in larger dimensional series.
- We show that easily interpretable key patterns can be used for accurate early classification.

II. Methodology

We start in this section by highlighting the framework of our IPED method and then explaining each part of the method in detail in the following sections.

1550-4786/13 $31.00 © 2013 IEEE
DOI 10.1109/ICDM.2013.19

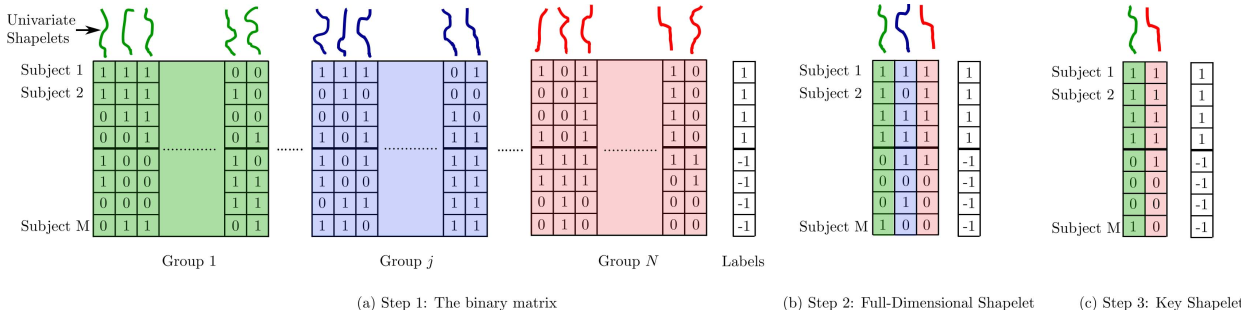


Fig. 1. IPED framework. (a) Transformation of N-variate time series observations on M subjects into a binary matrix using shapelets. The columns are the shapelets and the rows are the subjects. The elements of the matrix indicate the presence of the shapelet in the subject. Each group (different color) contains all shapelets extracted from a single dimension. (b)Extraction of a full-dimensional shapelet from the binary matrix such that only one univariate shapelet from each group is extracted. (c) Extraction of key shapelets (fewer dimensional shapelet) from the full-dimensional shapelet.

A. Method Sketch

First, let us introduce the notations we use throughout the paper. Let $D = \{(\mathbf{T}_i, y_i);\ i = 1 \dots M\}$ be a dataset of M patients where $\mathbf{T}_i$ is a multivariate time series of observations for the i-th patient and y_i is a binary class label describing a status of this patient at the end of an observation period (e.g. 0 - healthy, 1 - sick). For simplicity of representation, we assume that each time series is of length L (this is not necessary in our method). Let N be the number of dimensions of the time series. T_i^j is the j^{th} dimension of the time series $\mathbf{T}_i$ where $j = \{1, \dots, N\}$.

A recent study notes "*transforming the data is the simplest way of achieving improvement in problems where the discriminating features are based on similarity in change and similarity in shape*" [5]. Following this principle, we extract all time series subsequences of different lengths from each dimension of the multivariate time series. These subsequences are called univariate shapelets (or simply shapelets). The shapelet is used for discriminating between groups of time series. Therefore, we compute a distance threshold for each shapelet to maximize the information gain [6]. Using the shapelets, we construct a binary matrix (Figure (1a)) where each element of the matrix indicates the presence of the shapelet (column) in the subject time series (row) based on the shapelet's distance threshold. The shapelets are organized into N groups such that each group contains all shapelets extracted from one dimension.

Next, for each class we extract a multivariate shapelet from the binary matrix such that only one shapelet from each group is extracted (Figure (1b)). We call the extracted multivariate shapelet a *full-dimensional shapelet* since it contains a representative from each dimension of the observed time series. The problem of extracting a full-dimensional shapelet that maximizes the accuracy is formalized as an optimization problem, which is solved using convex-concave procedure [7].

As humans are able to estimate relatedness of only a few variables at a time [8], and some variables of the time series might be irrelevant for early classification, we extract *key shapelets* from the full-dimensional shapelet (Figure (1c)). We formalize this problem as a mixed integer optimization problem. Clearly, the resulting dimension of the key shapelet is less than or equal to the dimension of the observed time series. Then, we repeat that process to extract several full-dimensional shapelets and key shapelets for each class.

Finally, we use all the extracted key shapelets, as a representative for the class, for early classification. In early classification context, the objective is to observe the patient for a few time points and then compare the observed time series of the patient with the extracted key shapelets. If there is a match between the shapelet and the observed time series, the prediction is done. Otherwise, we observe the patients for more time points until a match is found.

We call our method, an optimization-based approach for Interpretable Pattern extraction for Early Diagnostics (IPED). We address the problem in three steps. In following Section II-B we transform this problem into binary matrix presentation. We use such representation in Section II-C to develop a convex-concave optimization method that extracts discriminative multidimensional patterns. From multidimensional patterns we extract a few key patterns by novel discrete optimization method presented in Section II-D. Finally, in Section II-E we show how key patterns can be used for early classification.

B. Problem Transformation to a Binary Matrix Representation

Let $S_{ikl}^j = T_i^j[k, \dots, k+l-1]$ be a contiguous subsequence of the time series T_i^j of length l that starts at time point k. These subsequences are patterns of interest called univariate shapelets (or simply shapelets) [6]. An effective method to extract all shapelets is described in [9], [10]. This method iterates over all dimensions j of all M time series, and for each subject univariate time series T_i^j it extracts all subsequences S_{ikl}^j (shapelets) of length l that start at the k^{th} time point. In other words, the method extracts all univariate shapelets from the time series dataset. Each shapelet will be referred to as S_{ikl}^j which means the subsequence of length l that is extracted from j^{th} dimension of the time series T_i from the start position k.

Our objective is to extract a few key shapelets and to use them for discriminating between classes. The discriminative shapelets are those that are present in the time series of one class but not in time series of another. Therefore, we transform our problem into a problem of selecting patterns from binary matrix which indicate presence of shapelets in time series. The transformation process is summarized in Algorithm 1.

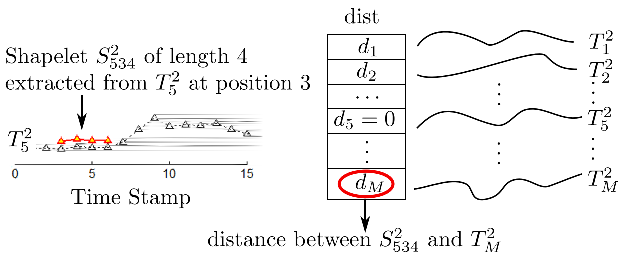


Fig. 2. In this example, the shapelet S^2_{534} (short red line) of length 4, is extracted from the 2^{nd} dimension of the time series $\mathbf{T}_5$, of length 15, from the starting position 3. $dist$ is an array of distances d_m between the shapelet S^2_{534} and the 2^{nd} dimension of each time series $\mathbf{T}_m : m = \{1, 2, \ldots, M\}$. Note that the distance between S^2_{534} and T^2_5 is zero because the shapelet is extracted from T^2_5.

For each extracted shapelet S^j_{ikl} we compute the distance between the shapelet and j^{th}-dimension of each of M multidimensional time series in the dataset using the function $ComputeDist$. The distance between S^j_{ikl} and time series T^j is defined as the minimum distance between S^j_{ikl} and each subsequence of length l from T^j. The function $ComputeDist$ returns an array $dist$ of length M where the m^{th}-element represents the distance between the shapelet S^j_{ikl} and the time series T^j_m (Figure (2)). Next, we need to determine threshold d^j_{ikl} on distances such that if the distance is less than threshold we say that the shapelet is present in the time series and vice versa. For each shapelet we compute threshold so that it maximizes the separation of the dataset into two classes by maximizing information gain [6] (an ideally discriminative shapelet is the one that is present in all time series of one class but not in any time series of another class).

Note that for a shapelet S^j_{ikl} that is extracted from j-th dimension we only determine its presence in j-th dimension of time series of all patients. We say that a shapelet S^j_{ikl} is present in (*covers*) the time series T^j if distance between S^j_{ikl} and T^j is less than a threshold d^j_{ikl}.

Algorithm 1 Transformation of N-dimensional Time Series

Input: A training dataset D of M N-dimensional time series; $minL$; $maxL$ (user parametersfor minimum and maximum shapelet's length)
Output: A binary matrix F where rows represent subjects and columns represent shapelets
for $j = 1$ **to** N **do** {each dimension}
  for $i = 1$ **to** M **do** {each subject time series}
    for $l = minL$ **to** $maxL$ **do** {each shapelet's length}
      for $k = 1$ **to** $L - l + 1$ **do** {each start position}
        $dist = ComputeDist(S^j_{ikl}, D)$
        Compute distance threshold d^j_{ikl}
        Construct column feature F^j_{ikl}
      end for
    end for
  end for
end for

Consequently, we construct a binary vector F^j_{ikl} which depicts the presence of the shapelet S^j_{ikl} in all training subjects (e.g. a column in Figure (1)). The m^{th} element in the vector F^j_{ikl} equals 1 if the distance between the shapelet S^j_{ikl} and the time series T^j_m is less than the shapelet's distance threshold d^j_{ikl}. Otherwise, it is 0. We call binary vector F^j_{ikl} the profile of the shapelet S^j_{ikl} [11]. For example, each column in Figure (1a) represents a univariate shapelet profile.

Algorithm 1 returns a binary matrix F of size $M \times R$ where M is the number of subjects and R is the number of shapelets, where $R = NM \sum_{k=minL}^{maxL} k(L - k + 1)$. The columns of the matrix F are partitioned into N groups such that each group j contains all extracted shapelets from the dimension j as in Figure (1a).

C. Extracting Full-dimensional Shapelets

The binary matrix F contains profiles for all shapelets from all dimensions of the time series. Having in mind the need for interpretability, for each class, our objective is to retain exactly one "class representative" univariate shapelet from each dimension of multivariate time series such that classification accuracy is maximized. The naïve approach to extract such representative shapelets is to look at each dimension separately and extract the shapelet which maximizes the accuracy (or minimizes classification loss) within the observed group. However, with this approach the possible interactions among shapelets from different groups are not taken into account.

We propose a novel method that simultaneously finds exactly one representative shapelet in each group. In order to mathematically define the problem that we are solving, let us assume that we have N groups of shapelets, where R^j is the number of shapelets in group j, $j = \{1, \ldots, N\}$. We then assign weight w^j_i to each shapelet $i = \{1, \ldots, R^j\}$ in group j which measures the importance of the shapelet in the classification. If we do not impose any restrictions on weights $W = \{w^j_i\}$, the optimal weights can be found by minimization of logistic loss over M subjects in the training data. The objective becomes minimization of logistic loss $\mathcal{L}_1$ with respect to the weights W.

$$\underset{W}{\text{minimize}} \quad \underbrace{\sum_{m=1}^{M} \log(1 + \overbrace{e^{-y_m \cdot \sum_{j=1}^{N} \sum_{i=1}^{R^j} (w^j_i \cdot f^j_{im})}}^{\mathcal{I}_m})}_{\mathcal{L}_1} \qquad (1)$$

where f^j_i represents the profile of the shapelet i from the group j, i.e. one column in the binary matrix F in Figure (1). f^j_{im} represents the m^{th} component of f^j_i and y_m is the label for the m^{th} subject.

To be able to identify exactly one shapelet within each group we solve the following optimization problem with constraints on the weights W

$$\begin{aligned}
&\underset{W}{\text{minimize}} && \mathcal{L}_1 && && \text{(2a)}\\
&\text{subject to} && 0 \le w_i^j \le 1, && \forall i,j, && \text{(2b)}\\
& && \sum_{i=1}^{R^j} w_i^j = 1, && \forall j, && \text{(2c)}\\
& && \max_{i=1\ldots R^j} w_i^j = 1, && \forall j. && \text{(2d)}
\end{aligned}$$

Equation (2b) indicates that all weights are bounded. The lower bound has to be 0 as we are interested only to extract representative shapelets for the positive class. The upper bound can be any positive real number. For simplicity we set the upper bound to 1. We need to impose this upper bound to be able to assure that exactly one weight within a group is not 0 and all other weights in the same group are equal to 0 by (2c) and (2d). Equation (2c) restricts weights to be normalized making sum of all weights within a group to be equal to 1. When (2c) is used together with (2d), which says that maximum weight within a group has to be equal to 1, we achieve that exactly one w_i^j within a group j is 1 while all other weights within a group j have to be 0.

The constrained optimization problem (2) is hard to solve since the max function is not differentiable. To be able to apply standard convex apparatus, we propose a transformation of (2) into a new optimization problem in which we: 1) relax hard equality constraints by introducing penalized terms in the objective function and 2) approximate the max function with convex differentiable $log\text{-}sum\text{-}exp$ function. The objective function with relaxed hard equality constraints becomes:

$$\begin{aligned}
&\underset{W}{\text{minimize}} && \mathcal{L}_1 + C_1 \cdot \overbrace{\sum_{j=1}^{N}(\sum_{i=1}^{R^j} w_i^j - 1)^2}^{\mathcal{L}_2}\\
& && + C_2 \cdot \sum_{j=1}^{N}(1 - \max_{i=1\ldots R^j} w_i^j) && \text{(3a)}\\
&\text{subject to} && 0 \le w_i^j \le 1, \quad \forall i,j. && \text{(3b)}
\end{aligned}$$

where we set the penalization parameters C_1 and C_2 to some values ($C_1 = C_2 = 0.1$ in our experiments). The intuition behind these penalization terms is that we would like to penalize the difference between sum of weights and 1, as well as maximum weight in each group and 1. We use quadratic penalty in the first term, as the sum of weights can be both greater or less than 1. We do not need quadratic penalty in the second term, as $\max_{i=1\ldots R^j} w_i^j$ is always less than or equal to 1 because of constraint (3b). For that reason, we need to penalize the difference between 1 and max of weights without the need for squaring the term.

To get a differentiable optimization function we need to approximate max with a smooth function. We start with the following lower bound for the max function [12]

$$\max_{i=1..R^j} w_i^j \ge \log(\sum_{i=1}^{R^j} e^{w_i^j}) - \log R^j. \qquad (4)$$

Then we have

$$1 - \max_{i=1..R^j} w_i^j \le \underbrace{1 - \log(\sum_{i=1}^{R^j} e^{w_i^j}) + \log R^j}_{\mathcal{M}^j}. \qquad (5)$$

This means that the second penalization term is upper bounded with smooth $log\text{-}sum\text{-}exp$ function. Penalizing the right hand side of (5) assures that the maximum is close to 1. If we combine (5) and (3) we get an optimization problem:

$$\begin{aligned}
&\underset{W}{\text{minimize}} && \mathcal{L}_1 + C_1 \cdot \mathcal{L}_2 + C_2 \cdot \mathcal{L}_3 && \text{(6a)}\\
&\text{subject to} && 0 \le w_i^j \le 1, \quad \forall i,j. && \text{(6b)}
\end{aligned}$$

where $\mathcal{L}_3 = \sum_{j=1}^{N} \mathcal{M}^j$. The objective function (6a) is convex-concave. $\mathcal{L}_1$ and $\mathcal{L}_2$ are convex while $\mathcal{L}_3$ is concave as it is equal to negative convex $log\text{-}sum\text{-}exp$ function. Therefore, we can apply the convex-concave procedure (CCCP) [7], [13] to find the global optimal solution. The application of CCCP to find optimal solution is shown in Algorithm 2. The term

Algorithm 2 Extract Full-Dimensional Shapelet

Initialize W^0
repeat
$$W^{t+1} = \underset{W}{\text{argmin}} \overbrace{\mathcal{L}_1 + C_1 \cdot \mathcal{L}_2 + C_2 \cdot W \cdot \left(\frac{d\mathcal{L}_3}{dW}\right)_{W=W^t}}^{\mathcal{J}}$$
until Convergence of W $\{W^{t+1} - W^t \le 0.01\}$

$(d\mathcal{L}_3/dW)_{W=W^t}$ is the derivative of $\mathcal{L}_3$ at the point W^t. The advantage of CCCP is that no additional hyper-parameters are needed. Furthermore, each update is a convex minimization problem and can be solved using classical and efficient convex apparatus.

Since we now have a smooth, differentiable objective function $\mathcal{J}$ with only inequality constraints, we can use the trust-region-reflective algorithm for solving the problem [14]. In order to quickly solve the problem we compute first derivatives of the objective function with respect to the weights W, and approximate Hessian with diagonal matrix [15] as follows

$$\begin{aligned}
\frac{\partial \mathcal{J}}{\partial w_i^j} &= \sum_{m=1}^{M} \frac{\mathcal{I}_m \cdot (-y_m \cdot f_{im}^j)}{1+\mathcal{I}_m} + 2 \cdot C_1 \cdot (\sum_{i=1}^{R^j} w_i^j - 1)\\
&\quad - C_2 \cdot \left(\frac{\exp^{w_i^j}}{\sum_{i=1}^{R^j} \exp^{w_i^j}}\right)_{W=W^t}, && \text{(7)}\\
\frac{\partial^2 \mathcal{J}}{\partial w_i^j \partial w_i^j} &= \sum_{m=1}^{M} \frac{\mathcal{I}_m}{(1+\mathcal{I}_m)^2} \cdot (y_m \cdot f_{im}^j)^2 + 2 \cdot C_1. && \text{(8)}
\end{aligned}$$

D. Extracting Key Shapelets

As humans are able to perceive only a few variables at a time [8], using the full-dimensional shapelets for early classification has several drawbacks, evidenced by our experiments reported in Section III-D. First, if the number of the dimensions of the time series is high, then it would be implausible for all components of the shapelet to cover all dimensions of the time series. This will affect the earliness of the method and the decision of the classification would come late, if it comes at all. Second, if some of the dimensions are irrelevant to the target, then these irrelevant dimensions will be subsequently inherited to the full-dimensional shapelet and would affect the overall accuracy of the method. Therefore, we propose a method to extract automatically, for each class, discriminating representative key shapelets by extracting all relevant dimensions from the full-dimensional shapelet.

Let's simplify the idea of the key shapelet with the following example. Assume that we have 8 subjects where the class labels of the subjects are defined in y. Suppose we have extracted a large dimensional (3-dimension) shapelet S with its profile $\hat{S}$ as representative for the positive class (as in Figure (1b)). We would like S to have maximum coverage for the positive class and minimum coverage for the negative class (*that is present in all time series of the positive class but not in the time series of the negative class*).

$$y = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ -1 \\ -1 \\ -1 \\ -1 \end{bmatrix} \quad \hat{S} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad \hat{S}' = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}$$

The full-dimensional shapelet S covers a time series if each component of S covers the corresponding dimension of the time series. Since the first row of $\hat{S}$ has three ones, it means that each component of the shapelet S covers the corresponding dimension of the time series for subject 1. Therefore, the shapelet S covers the first subject. The same for the 3^{rd} and 4^{th} subjects. The shapelet S does not cover the 2^{nd} subject because the 2^{nd} component does not cover the corresponding dimension of the time series of the 2^{nd} subject. Then, S has positive coverage 75% (covers 3/4 of the subjects in the positive class). It is clear from the second column in $\hat{S}$ that the 2^{nd} component in S is noisy because it is represented three times in the positive class and two times in the negative class (1^{st}, 3^{rd}, 4^{th}, 5^{th} and 7^{th} subjects).

Obviously, we can extract a key shapelet S' with only two components from S. S' has better positive coverage than S because it covers all positive subjects and does not cover any negative subject. Sometimes we can extract multiple key shapelets from one full-dimensional shapelet.

A recent paper proposed a novel mixed integer optimization approach for extracting interpretable compact rules for the classification task [16], [17]. Thus, we adapt the approach to selecting several key shapelets from a full-dimensional shapelet. Here we review the method along with our modifications.

It is worth mentioning that we could minimize the logistic loss with L_1 and L_2 regularization terms (elastic net) instead of using the integer discrete optimization problem because the penalized logistic regression is much faster. However, properties of penalized logistic regression are not suitable for early classification which we will explain later in this section why elastic net does not work for our application. However, the results when applying penalized logistic regression to our datasets were worse than using a novel mixed integer optimization approach (Section III-F).

Let $b \in \{0,1\}^N$ be a binary vector that encodes the presence of the N components in the key shapelet. For example, S' is encoded by the binary vector $b = [101]$ which means that only the first and third components are in the key shapelet S'. Let x_i be defined as

$$x_i = \begin{cases} 1 & \text{if } S' \text{ covers the subject } i, \\ 0 & \text{otherwise.} \end{cases}$$

The problem of selecting a key shapelet from the full-dimensional shapelet S can then be formulated as the following optimization problem

$$\begin{aligned} \underset{b,x}{\text{maximize}} \quad & \sum_{i_+} x_i - R_1 \sum_{i_-} x_i - R_2 \sum_{j=1}^{N} b_j & \text{(9a)} \\ \text{subject to} \quad & x_i \leq 1 + (S_{ij} - 1) b_j, \forall i, j, & \text{(9b)} \\ & x_i \geq 1 + (S_i - \mathbf{1_N})^T b, \forall i, & \text{(9c)} \\ & \sum_{j=1}^{N} b_j \geq \mathcal{B}, & \text{(9d)} \\ & b_j \in \{0,1\}, & \text{(9e)} \\ & 0 \leq x_i \leq 1. & \text{(9f)} \end{aligned}$$

where $\mathbf{1_N}$ is an N-dimensional column vector of all ones and $i = \{1 \ldots M\}, j = \{1 \ldots N\}$ unless otherwise stated. i_+ are indices for all positive subjects and i_- are indices for all negative subjects. R_1, R_2, and $\mathcal{B}$ are parameters that will be discussed later.

The first term of the objective function (9a) corresponds to maximizing the coverage of the shapelet for the positive class while the second term ensures that the shapelet is not represented in the negative class. The last term controls the sparsity of the multivariate shapelet. So, if there are several optimal solutions that maximize the coverage for the positive class and minimize the coverage in the negative class, we choose the one that has less components. R_1 and R_2 control the weights for those terms and have been set as $R_1 = 10$ to emphasize the importance of non-coverage for the negative class, and $R_2 = \frac{0.1}{MN}$ as was set previously [16].

For the constraint (9b), we have two cases for b_j. If $b_j = 0$ then the constraint is just $x_i \leq 1$ which has no effect. If $b_j = 1$ then the constraint becomes $x_i \leq S_{ij}$ which means that $x_i = 0$ if the subject i is not covered by the component j. The constraint (9c) explains the case when the subject i is covered by the component j. Therefore, the constraints (9b) and (9c) combined ensure that $x_i = 1$ if and only if all the components of the shapelet cover all the corresponding dimensions of the subject i (*this is the main reason why this approach outperformed the elastic net approach in early classification. Elastic net allows some components of the shapelets to be zero. Although it seems to be an advantage for elastic net, it did not work in early classification context*).

The constraint (9d) determines the minimum number of components of the multivariate shapelet. Complex shapelets (high values of $\mathcal{B}$) increase the accuracy performance of the method because $\mathcal{B}$ dimensions of the time series would be covered. However, this might affect the earliness. Therefore,

Algorithm 3 Extract Key Shapelets

Input: full-dimensional shapelet profile $\hat{S}$
Output: List of key shapelets
while *PosCoverage*$(S') \geq CovThreshold$ **do**
  S' = Solve problem (9)
  Add S' to $Solutions$
  Add new constraint (10)
end while
Rank $Solutions$

Algorithm 4 IPED

Input: A time series dataset D.
Transform D into binary matrix (Algorithm 1)
for $c = 1$ **to** C **do**
  Consider c as the positive class
  for $i = 1$ **to** I **do**
    Extract full-dimensional shapelet (Algorithm 2)
    Extract key shapelets (Algorithm 3)
  end for
end for

the multivariate shapelet can not be neither too simple nor too complex. In Section III-C, we study the sensitivity of the method on the model complexity parameter $\mathcal{B}$.

By solving the problem (9), we obtain one key shapelet. Sometimes, there are several optimal key shapelets that could be extracted from the same full-dimensional shapelet. Therefore, we need to resolve the problem (9) to obtain more key shapelets. However, we need to make sure that the second time we solve the problem, we will not get one of the previous solutions. So, each time we solve (9) we add the following constraint

$$1 - \sum_{j:b_j^*=0} b_j - \sum_{j:b_j^*=1} (1 - b_j) \leq 0, \qquad (10)$$

where b^* is the previous solution. The constraint (10) ensures that the new solution does not equal to the previous solution. So, the final key shapelets would have different components.

Therefore, we initially solve problem (9) and get an optimal solution. Then we add a new constraint (10) and resolve the problem (9) to get a new optimal key shapelet. We repeat this process by solving (9) until the solution has positive coverage (number of positive subjects covered by the shapelet using function *PosCoverage*) less than a threshold or we exceed a certain number of iterations. After we extract all key shapelets, we rank them such that the first ranked shapelet has maximum positive coverage. The process is shown in Algorithm 3.

Up to this point we have extracted several key shapelets that cover subjects in the positive class. We can iterate this procedure (extracting a full-dimensional shapelet followed by extracting key shapelet) for additional I iterations. Having in mind the need for interpretability and selecting only key shapelets as well as the need to avoid overfitting with too complex model [18], we consider $I = 2$ iterations. We then flip the labels of the subjects and do the same process again by iteratively extracting a full-dimensional shapelet and then key shapelets for the other class. The whole IPED procedure is summarized in Algorithm 4.

E. Key Shapelets for Early Classification

After running Algorithm 4, we end up with several key shapelets for each class. For a time series with unknown label, the IPED method initially reads $minL$ (minimum length of any shapelet which is user parameter) time stamps from the test time series. First, the highest-ranked key shapelet is considered. If any of the components of the key shapelet covers the corresponding dimension of the current stream of the test multivariate time series, we mark the component of the shapelet that covers the time series. *Recall that the component of the shapelet covers a time series if the distance between the time series and the component of the shapelet is less than the distance threshold of the component.* If all components of the shapelet are marked, then the time series is classified as the class of the shapelet and the process stops. Otherwise, next key shapelet from the ranked list is considered and the process is repeated. If none of the shapelets cover the current stream of the test time series, the method reads one more time stamp and continues classifying the time series. If the method reaches the end of the time series and none of the shapelets cover it, the method marks the time series as an unclassified subject.

We note that the components of the multivariate shapelets may classify the time series at different time points. So, one component of the multivariate shapelet may cover the time series at time point t_1 and the other component may cover the time series at t_2 where $t_2 > t_1$. However, the final decision is made only when all components cover the time series, which is the last time point t_2. Therefore, the test time series could be classified after reading a number of time points greater than the shapelet's length.

III. Experimental Results

The IPED method is applicable in any context where interpretable early classification is desired. In order to compare IPED to the state-of-the-art method for interpretable early classification [19], [20] and for time series classification [18], we show the accuracy and earliness performance of IPED on three challenging medical datasets. A brief description of the data is shown in Table I. A description for the methods we compare to is provided in Section IV.

We used datasets for blood gene expression from human viral studies with influenza A (H3N2) and live rhinovirus (HRV) to distinguish individuals with symptomatic acute respiratory infections from uninfected individuals [21]. Since the dataset has relatively small number of patients, we apply leave one out cross validations.

PTB [22] is an ECG database available at the Physionet website[1] [23]. In this application, we are interested in distinguishing between ECG signals of individuals with myocardial infarction (368 records) and those of healthy controls (80 records). The dataset consists of records of the conventional 12 leads ECGs together with the 3 Frank lead ECGs. Since the dataset is imbalanced, we report the accuracy as the average between sensitivity and specificity. 20 patients (10

[1] http://www.physionet.org/physiobank/database/ptbdb

TABLE I. DATASETS DESCRIPTIONS. THE NUMBER BETWEEB PARANTHESE ARE THE NUMBER OF SUBJECTS IN EACH CLASS.

	H3N2	HRV	PTB
Number of Subjects	17(9/8)	20(10/10)	448(368/80)
Time Series Length	16	14	3200
Number of dimensions	23	26	15

from each class) are used as training data (to simulate the real-life scenario where a small number of temporal observation are provided) and 428 patients are used for tests.

A. Experimental Setup and Evaluation Measures

We compare our method to 3 alternative methods. Logical Shapelet (LS) [18], Early Distinctive Shapelet for Classification (EDSC) [20], and Multivariate Shapelet Detection (MSD) [19] (see Section IV for descriptions of these methods). LS is designed for classification of univariate time series and EDSC for early classification of univariate time series. Therefore, we applied both methods on two settings: 1) on each variable of the dataset separately. 2) on a new variable that is constructed by concatenating all the variables in the dataset. We report the one that gives the best result. However, LS is not designed for early classification, therefore, it utilizes the full time series of the test data. The MSD method is designed for early classification of multivariate time series so it is applicable directly to our datasets.

The coverage threshold used in Algorithm 3 to retrieve more key shapelets is set to be 50% of the training class subjects, to ensure that more training subjects are covered and to allow for variability among subjects. The $maxIter$ is set to 5 to limit the computational time. For the MSD method, the parameters are optimized using internal cross validations.

We report the following measures:

1) *Coverage* which is the percentage of subjects who were classified. LS has always 100% coverage because it utilizes the full time series.
2) *Accuracy* which is computed as the average between sensitivity and specificity. For early classification methods (IPED, MSD and EDSC) the accuracy is computed with respect to the coverage.
3) *Earliness* which is the fraction of the time points used in the test. LS has always 100% earliness because it utilizes the full time series.
4) *Full Accuracy* since we report the coverage and the accuracy, it might not be easy to directly compare the methods. For example, if one method has better accuracy and less coverage and the second has better coverage and less accuracy, it is not clear which one is better. Therefore, for those examples not covered by the method, the method classifies them using the majority of examples in the training data or randomly in case of a tie.

All the evaluation measures have the property "*higher is better*" except the earliness. In order to preserve this property and to simplify the presentation of the results, we report 100-*Earliness*.

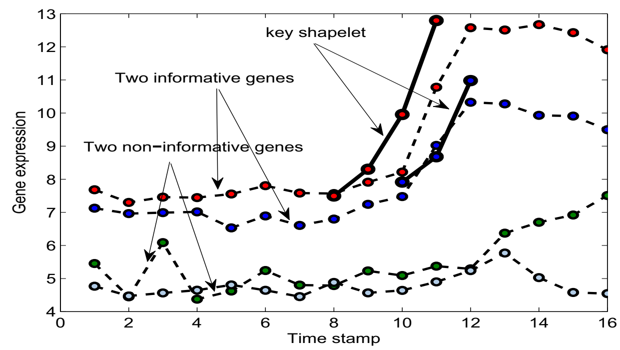


Fig. 3. The effectiveness of the IPED method is illustrated on a single patient from H3N2. Four genes for a symptomatic test subject were observed over 16 time points. Two genes are informative for early classification (red and blue) while the other two gene are non-informative. The key shapelet covers the subject at 12 the time point. We add displacement in the gene expression to make the visualization easier.

B. Interpretability of IPED

To show the benefit of extracting a key shapelet and to explain some features of the IPED method, we show a real case from the H3N2 dataset. We used two-dimensional shapelets just for simplicity of the presentations. The experiments in Section III-C show that a 3-dimensional shapelet is more reasonable for our application.

In Figure (3), we show four genes observed over time for the same symptomatic subject. The red and blue genes are informative because we have jump in the gene expression, while the other two genes are non-informative.

The IPED method is interpretable because the key shapelet is interpreted as "*symptomatic subject is identified when we observe high increase in the red gene accompanied with high increase in the blue gene*".

We use this example to explain how IPED overcomes the issues presented in the MSD method. First, the IPED method has extracted a key shapelet which contains only the two informative genes. This shows that the key shapelet is more accurate than the full-dimensional shapelet, as in the case of the MSD method, because it contains only the relevant dimensions. Second, each component of the key shapelet is of different length and the onset of each component is different. For instance, the 1^{st} component of the shapelet covers the corresponding dimension of the subject time series at the 11^{th} time point while the 2^{nd} component covers the corresponding dimension at the 12^{th} time point when the decision is made.

C. Model Complexity

As mentioned in Section II-D, the parameter $\mathcal{B}$ determines the complexity of the multivariate shapelet. We assessed the sensitivity of IPED to the parameter $\mathcal{B}$. For each value of $\mathcal{B}$, we compute the evaluation measures. The results are shown in Figure (4).

As expected, when increasing the number of components of the shapelet, the accuracy of the model increases but the the coverage decreases and the classification decisions are

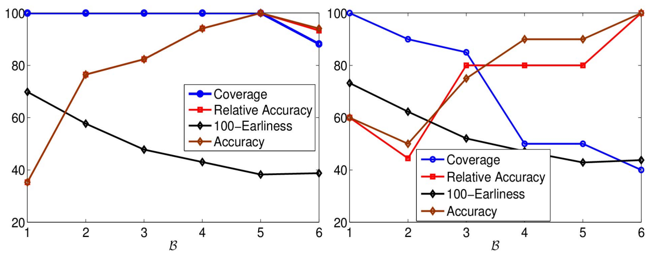


Fig. 4. The sensitivity of the model complexity $\mathcal{B}$ on H3N2 (left) and HRV (right) datasets.

TABLE II. EVALUATION OF DIFFERENT OPTIMIZATIONS TO EXTRACT THE KEY SHAPELETS. *EN* IS THE ELASTIC NET APPROACH. *None* IS THE METHOD WITHOUT THE SECOND OPTIMIZATION. COV IS THE COVERAGE, ACC IS THE ACCURACY, AND EAR IS THE EARLINESS.

		Cov	Acc	100-Ear
H3N2	IPED	100	82.35	47.8
	EN	21.2	80.0	5.4
	None	17.6	60	4.5
HRV	IPED	85.0	80.6	47.2
	EN	19.0	54.3	9.5
	None	1	10	0.4
PTB	IPED	100	89.78	47.43
	EN	25	70.2	60.2
	None	1.17	90.2	4.2

provided later. Since our objective is to provide the classification as early as possible and to maintain comparable accuracy performance, we use $\mathcal{B} = 3$ in the remaining experiments for all datasets to guard against overfitting and not to use different values of $\mathcal{B}$ for different datasets.

D. Evaluation of IPED

In a recent article [19], experiments on viral infection datasets (Table 1) provided evidence that MSD was more accurate than 1-nearest neighbor (1NN) on the full time series. Therefore, here we compare IPED to MSD, LS and EDSC on the 3 datasets described earlier using the four aforementioned evaluation measures.

As shown in Figure (5), our IPED method has better performance than the alternative MSD method on the PTB dataset. In addition, IPED outperforms all other classifiers and provides the classification much earlier, using approximately half of the time series.

For the viral infection datasets, IPED has comparable results with the MSD method. For univariate classifiers, LS and EDSC, the results obtained by concatenating all variables are worse than the results obtained on one variable. It shows that LS and EDSC are not appropriate for multivariate time series. LS is the most accurate of the classifiers. However, LS provides the classification at the end of the time series while IPED provides comparable results even earlier.

Since the early classification is important in the medical domain, we have shown that on a longer time series datasets such as PTB, the classification was provided using only half of the time series, which could have significant impact by providing treatment to the patient at an earlier time. In addition, the classification accuracy of IPED was comparable to or even better than those classifiers that use the full time series.

E. Runtime Complexity Analysis

The IPED method has three steps to extract key shapelets from the time series data. The first step is the shapelets extraction. This step is performed using exhaustive search to extract all univariate shapelets from all dimensions of the time series data. In our experiments on the PTB dataset (the largest dataset in our experiments), the shapelet extraction step took less than a minute. However, there are some recent works to speed up the process of that step [9], [10]. We did not use them because it is not the main point of our paper.

The second step of the IPED method is extraction of a full-dimensional shapelet by solving a convex-concave problem. This is the most time-consuming part of the method because it requires solving a convex optimization on a dataset that is constructed using all extracted univariate shapelets. On the PTB dataaset, it took about 15 minutes to solve the optimization problem.

The third step is the extraction of the key shapelets from the full dimensional shapelet. This part is very fast (2-3 seconds) because the dimension of the input data to the problem is the same as the number of dimensions of the original time series (which in our case was 15-26). Therefore, the most time is consumed to apply the second step of the method. However, this is only for training the model, which is done offline.

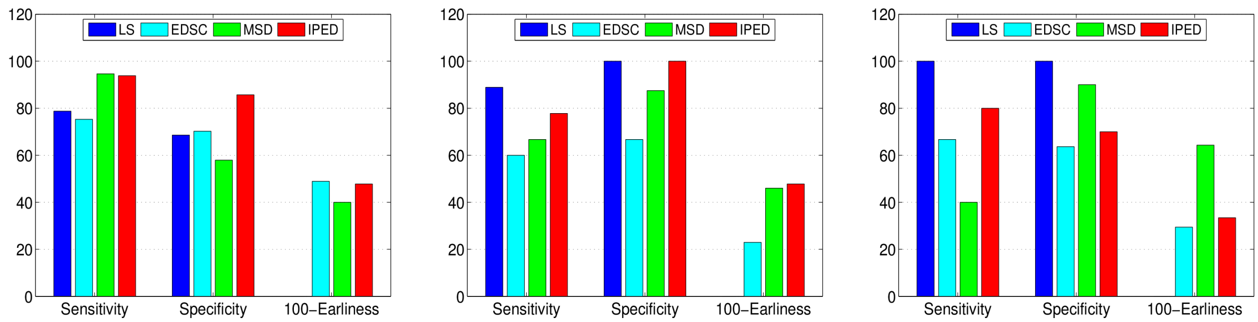
Using the extracted key shapelets for early classification (which is done online), is a very fast operation, because the number of extracted shapelets is small. In our case, the early classification of the test time series for each patient takes less than a second to classify the patient.

F. Evaluation of the Second Optimization

To evaluate the importance of our mixed integer optimization formalization, we conducted two experiments. First, we removed the second optimization completely and kept only the first optimization, which we call "*None*". In the second experiment, we replaced the mixed integer optimization with the penalized (elastic net) logistic loss as the objective function, which we call "*EN*". The results are shown in Table II. Using the mixed integer programming as the second optimization in the IPED method outperformed all other alternatives because *EN* and *None* have very low coverage which affects the overall accuracy.

IV. Related Work

Multivariate time series classification has been widely analyzed in a context of learning a hypothesis to discriminate between groups of time series [24]–[26]. Time series shapelets, introduced by [6], were initially proposed for univariate time series where the objective is to extract time series subsequences which are in some sense maximally representative of a class. The time series shapelet concept has gained a lot of attention in the last few years because of its simplicity, interpretability and accuracy. However, the shapelet representation has limited expressiveness for the time series. This problem has been



(a) PTB (b) H3N2 (c) HRV

Fig. 5. Comparison between IPED and the alternative methods on three datasets. LS and EDSC methods are applied to each variables separately and to a new variable that is constructed by concatenating all the variables, and the best result is reported. Sensitivity, specificity and the 100-earliness are reported for each method (the higher is the better). LS is the Logical Shapelet method. EDSC is the Early Ddistinctive Shapelet for early Classification of univarite time series. MSD is the Multivariate Shapelet Detectiom method. IPED is our proposed method. LS utilizes the full time series so that the 100-earliness for the LS method is zero.

addressed by considering combination (conjunction and disjunction) of shapelets which are called logical shapelet [18]. As noted by [18], *for the sake of simplicity and to guard against over fitting with too complex a model, they only consider two cases, only AND, and only OR combinations*. Although the method has been shown to be more accurate than other time series classification techniques, it is not proposed for early classification.

Early classification methods have been developed recently [27]. A method called ECTS (Early Classification on Time Series) has been proposed to tackle the problem of early prediction on time series data. ECST uses a novel concept of MPL (Minimum Prediction Length) and effective 1-nearest neighbor (1NN) classification method which makes prediction early and at the same time retains an accuracy comparable to that of a 1NN classifier using the full-length time series. Although ECTS has made good progress, it only provides classification results without extracting and summarizing patterns from training data. The classification results may not be satisfactorily interpretable and thus end users may not be able to gain deep insights from and be convinced by the classification results. This problem has been addressed by extracting shapelets for early classification [20]. The method, which is called EDSC (Early Distinctive Shapelet Classification), extracts local shapelets, which distinctly manifest the target class locally and effectively use them for interpretable early classification. Although very effective for univariate time series, EDSC is not applicable for multivariate time series.

A method called hybrid HMM/SVM was proposed for early classification of multivariate time series [28], [29]. The method uses HMMs to examine all segments from the original time series and generate likelihood of the membership of the pattern which is then passed to SVM to decided the probability of the class membership of the time series. Although the method was shown to be an accurate method, it does not provide interpretable results which limits the application of the method by the clinical practitioners. More recently, a method called Multivariate Shapelet Detection (MSD) was proposed for early classification of multivariate time series, which appears to be more accurate than several alternative methods when evaluated on various benchmark clinical multivariate

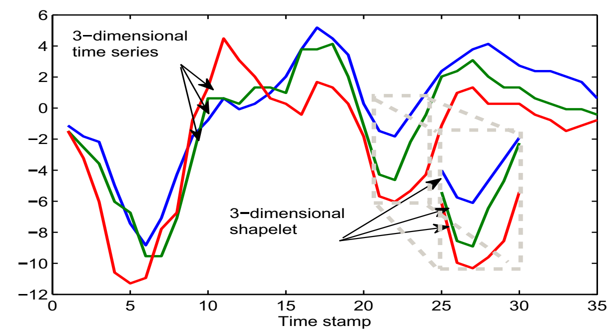


Fig. 6. Example of a 3-dimensional shapelet extracted from a 3-dimensional time series of length 35. The components of the 3-dimensional shapelets have the same start position at the 20^{th} time point and the same length 5.

time series [19]. The method extracts all multivariate shapelets from the training data and effectively use them for early classification. However, the multivariate shapelet is extracted such that it has one component from each variable in the time series. For instance, the extracted multivariate shapelet from a 3-dimensional time series has 3 components (univariate shapelets) as in Figure (6). Each component is extracted from one variable. To control computational complexity in MSD, the components of the multivariate shapelet are extracted such that they have the same starting position. Therefore, all components have to appear at the same time which we believe affects the efficiency of MSD. So, the method would not be able to capture a multivariate shapelet such that one component appears at time point t_1 and the other component appears at time point t_2.

In our approach we relaxed that condition so that our method is able to extract multivariate shapelets such that the components of the shapelets may appear at different time points and can be of different length. In addition, our IPED method extracts key shapelets, which have much fewer dimensions than the original time series. That contributes more to the interpretability of the IPED method.

V. Conclusion

We proposed an optimization-based method for building predictive models on multivariate time series data and mining relevant temporal interpretable patterns for early classification (IPED). The IPED method starts with transforming the multivariate time series data into a binary matrix representation over the span of all extracted shapelets from all the dimensions of the time series. Then, the matrix is used to build predictive models via solving the proposed optimization formulations. The IPED method extracts a full-dimensional shapelet for each class from the binary matrix by solving a convex-concave optimization problem. Then, IPED extracts key shapelets for each class by solving a mixed integer optimization problem. The extracted key shapelets are low-dimensional shapelets. These key shapelets are then used for early interpretable classification. We are able to make our results interpretable by using only a few patterns from the observed time series data.

Our IPED method addresses three issues in the state-of-the-art MSD method. First, the components of the multivariate shapelet do not have the constraints of the same starting time point and are not required to be of the same length. Therefore, the onset of each pattern could be different, which simulates a real life scenario. Second, the extracted multivariate shapelet contains only the relevant dimensions of the observed multivariate time series data. On several medical datasets, we have shown that IPED outperformed the MSD method and other alternative methods for univariate time series data. One of the shortcomings of IPED is that it assumes the time series are sampled at regular time points. We can mitigate this issue by including the time span of each shapelet in the computation.

Acknowledgment

This work was funded, in part, by DARPA grant [DARPAN66001-11-1-4183] negotiated by SSC Pacific grant.

References

[1] M. F. Ghalwash, V. Radosavljevic, and Z. Obradovic, "Early diagnosis and its benefits in sepsis blood purification treatment," in *International Workshop on Data Mining for Healthcare*, Philadelphia, PA, Sep 2013.

[2] I. Tzeng and W. Lee, "Detecting differentially expressed genes in heterogeneous diseases using control-only analysis of variance," *Annals of Epidemiology*, vol. 22, no. 8, pp. 598 – 602, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1047279712001111

[3] D. Nauck and R. Kruse, "Obtaining interpretable fuzzy classification rules from medical data," *Artif Intell Med*, vol. 16, no. 2, pp. 149–169, 1999.

[4] R. Bellazzi, L. Sacchi, and S. Concaro, "Methods and tools for mining multivariate temporal data in clinical and biomedical applications," in *31st Annual International Conference of the IEEE EMBS*, 2009.

[5] A. Bagnall, L. Davis, J. Hills, and J. Lines, "Transformation based ensembles for time series classification," in *Proceedings of the 12th SIAM International Conference on Data Mining*, 2012, pp. 307–319.

[6] L. Ye and E. Keogh, "Time series shapelets: A new primitive for data mining," in *Proceesings of the 15th ACM SIGKDD Conference on Knolwedge Discovery and Data Mining*, 2009, pp. 947–956.

[7] A. Yuille and A. Rangarajan, "The concave-convex procedure (CCCP)," in *Neural Computation*, vol. 15, 2003, pp. 915–936.

[8] D. L. Jennings, T. M. Amabile, and L. Ross, *Judgment Under Uncertainty: Heuristics and Biases*, D. Kahneman, P. Slovic, and A. Tversky, Eds. Cambridge Press, 1982.

[9] K.-W. Chang, B. Deka, W.-M. W. Hwu, and D. Roth, "Efficient pattern-based time series classification on gpu," in *IEEE International Conference on Data Mining*, 2012.

[10] T. Rakthanmanon and E. Keogh, "Fast-shapelets: A fast algorithm for discovering robust time series shapelets," in *Proceedings of 11th SIAM International Conference on Data Mining*, 2011.

[11] J. Lines, L. M. Davis, J. Hills, and A. Bagnall, "A shapelet transform for time series classification," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 289–297.

[12] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge university press, 2004.

[13] R. Collobert, F. Sinz, J. Weston, and L. Bottou, "Trading convexity for scalability," in *International Conference of Machine Learning*, 2006.

[14] T. F. Coleman and Y. Li, "An interior trust region approach for nonlinear minimization subject to bounds," *SIAM Journal on Optimization*, vol. 6, pp. 418–455, 1996.

[15] N. Djuric, M. Grbovic, and S. Vucetic, "Convex kernelized sorting," in *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, 2012, pp. 893–899.

[16] A. Chang, D. Bertsimas, and C. Rudin, "Ordered rules for classification: A discrete optimization approach to associative classification," Poster in Proceesings of Neural Information Processing Systems Foundations (NIPS), 2012.

[17] D. Bertsimas, A. Chang, and C. Rudin, "Ordered rules for classification: A discrete optimization approach to associative classification," in *Proceedings of ICML Workshop on Optimization in Machine Learning*, July 2012.

[18] A. Mueen, E. Keogh, and N. Young, "Logical-shapelets: An expressive primitive for time series classification," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 1154–1162.

[19] M. F. Ghalwash and Z. Obradovic, "Early classification of multivariate temporal observations by extraction of interpretable shapelets," *BMC Bioinformatics*, vol. 13, Aug. 2012.

[20] Z. Xing, J. Pei, P. S. Yu, and K. Wang, "Extracting interpretable features for early classification on time series," in *Proceedings of 11th SIAM International Conference on Data Mining*, 2011, pp. 439–451.

[21] A. K. Zaas, M. Chen, J. Varkey, T. Veldman, A. O. H. III, J. Lucas, Y. Huang, R. Turner, A. Gilbert, R. Lambkin-Williams, N. C. ien, B. Nicholson, S. Kingsmore, L. Carin, C. W. Woods, and G. S. Ginsburg, "Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans," *Cell Host and Microbe*, vol. 6, pp. 207–217, 2009.

[22] B. R, K. D, and S. A, "Nutzung der ekg-signaldatenbank cardiodat der ptb ber das internet," *Biomedizinische Technik*, p. 317, 1995.

[23] G. AL, A. LAN, G. L, H. JM, I. PCh, M. RG, M. JE, M. GB, P. C-K, and S. HE, "Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, pp. e215–e220, 2000.

[24] T. Fu, "A review on time series data mining," *Engineering Applications of Artificial Intelligence*, vol. 24, pp. 164–181, 2011.

[25] J. Lin, S. Williamson, K. Borne, and D. DeBarr, "Pattern recognition in time series," in *Advances in Machine Learning and Data Mining for Astronomy*, M. Way, J. Scargle, K. Ali, and A. Srivastava, Eds. New York, NY: Chapman and Hall, 2012, ch. 28, to appear.

[26] Z. Xing, J. Pei, and E. Keogh, "A brief survey on sequence classification," in *ACM SIGKDD Explorations Newsletter*, vol. 12, 2010, pp. 40–48.

[27] Z. Xing, J. Pei, and P. S. Yu, "Early prediction on time series: A nearest neighbor approach," in *Proceedings of the 21st international joint conference on Artifical intelligence*, 2009, pp. 1297–1302.

[28] M. F. Ghalwash, D. Ramljak, and Z. Obradovic, "Early classification of multivariate time series using a hybrid HMM/SVM model," in *Proceesings of IEEE International Conference on Bioinformatics and Biomedicine*, Philadelphia, PA, Oct 2012.

[29] M. F. Ghalwash, D. Ramljak, and Z. Obradovic, "Patient-specific early classification of multivariate observations," *International Journal of Data mining and Bioinformatics*, 2013, in press.