Journal of Biomedical Informatics 96 (2019) 103231

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Early detection and risk assessment for chronic disease with irregular longitudinal data analysis

Kai He[a], Shuai Huang[b], Xiaoning Qian[a,*]

[a] Department of Electrical & Computer Engineering, Texas A&M University, United States
[b] Department of Industrial & Systems Engineering, University of Washington, United States

ARTICLE INFO

Keywords:
Early diagnosis
Risk monitoring
Longitudinal measurements
Machine learning
Structured output
Support Vector Machine

ABSTRACT

Early detection and risk assessment of complex chronic disease based on longitudinal clinical data is helpful for doctors to make early diagnosis and monitor the disease progression. Disease diagnosis with computer-aided methods has been extensively studied. However, early detection and contemporaneous risk assessment based on partially labeled irregular longitudinal measurements is relatively unexplored. In this paper, we propose a flexible mixed-kernel framework for training a contemporaneous disease risk detector to predict the onset of disease and monitor the disease progression. Moreover, we address the label insufficiency problem by identifying the pattern of disease-induced progression over time with longitudinal data. Our method is based on a Structured Output Support Vector Machine (SOSVM), extended to longitudinal data analysis. Extensive experiments are conducted on several datasets of varying complexity, including the contemporaneous risk assessment with simulated irregular longitudinal data; the identification of the onset of Type 1 Diabetes (T1D) with irregularly sampled longitudinal RNA-Seq gene expression dataset; as well as the monitoring of the drug long-term effects on patients using longitudinal RNA-Seq dataset containing missing time points, demonstrating that our method enhances the accuracy in both early diagnosis and risk estimation with partially labeled irregular longitudinal clinical data.

1. Introduction

The rapid advancement of sensor and information technologies in recent decades such as the high-throughput next generation sequencing and imaging techniques provide unprecedented opportunities for us to develop methods for early diagnosis and contemporaneous monitoring of the disease. For example, a dynamic biological process of living organisms can be manifested by the changes in the gene expression, whose dysfunction and variation help better understand disease progression. The positron emission tomography (PET) scan imaging technique shows characteristic changes in the brains of patients with Alzheimer's disease (AD), and in prodromal and even presymptomatic states that can help estimate the AD pathophysiological process [25]. Early diagnosis is beneficial for disease prevention and early treatment as it plays an important role to raise cure rates, achieve better care and quality of life, and/or extend survival for chronic diseases which progress over time or have persistent and long-lasting in its effect [38,15]. For example, type 1 diabetes (T1D), a genetic chronic disease, whose disease progression can be subdivided into multiple stages while the symptoms only appear at the last stage as shown in Fig. 1[27]. Early detection can also be applied to the longitudinal study of the clinical responses to drug therapy. Identifying pre-existing and drug-induced signatures is important to predict the clinical response to the drugs [69].

Besides early diagnosis, contemporaneous monitoring of the disease progression is also critical for the care management of the patients with chronic conditions. One of the most important properties of chronic disease is that, as defined by the U.S. National Center for Health Statistics, the disease persists for long time. The speed of progression of the chronic diseases such as Alzheimer's disease and diabetes, varies greatly across patients due to different factors including genetics, physiology, social-economics, gender, and behavior. Contemporaneous monitoring of the disease progression can help patients get more appropriate care and treatments. Furthermore, contemporaneous monitoring of the disease progression can be very helpful in the study of drug response as well. E.g., it's crucial for doctors to have the capability of tracking the drug's longitudinal effects to provide reliable recommendations for the continual usage of medications to treat the disease.

To develop early diagnosis and contemporaneous disease

* Corresponding author.
E-mail address: xqian@ece.tamu.edu (X. Qian).

https://doi.org/10.1016/j.jbi.2019.103231
Received 26 July 2018; Received in revised form 9 May 2019; Accepted 11 June 2019
Available online 13 June 2019
1532-0464/

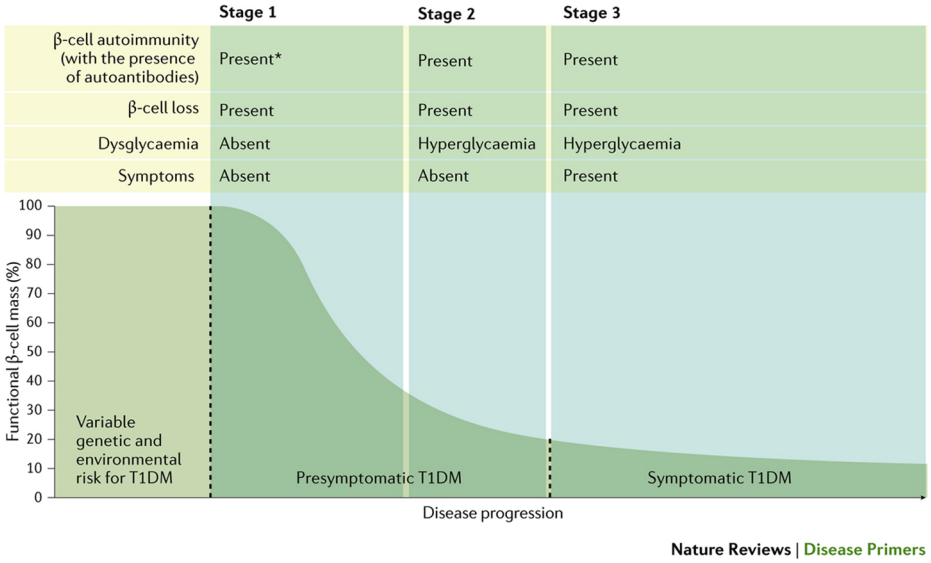


Fig. 1. T1D can be subdivided into three stages: stage 1 is characterized by the presence of autoantibodies and the absence of dysglycaemia; stage 2 is characterized by the presence of both autoantibodies and dysglycaemia; and symptoms only appear at stage 3, which corresponds to symptomatic T1D [27].

monitoring methods, we also need to overcome the following challenges:

- Difficulty in disease detection at the early stage

One property of the chronic disease is that they are slow to develop and may progress over time. This property makes early diagnosis difficult since patients at the early stage of diseases behave similarly as healthy people. E.g., for Alzheimer's disease patients at the early stage, their cognitive functions and living functions usually maintain as normal aging individuals. Fig. 2 provides a simple schematic example with the data containing 2 features ($x1$ and $x2$). In Fig. 2, there are two subjects with repeated observations: the patient (orange points) and the normal control (blue points). The patient can go through multiple stages: mild, moderate and severe. Most points at the early stage of the patient cannot be separated from those of the normal control. Early diagnosis is thus challenging at the early stage of the disease.

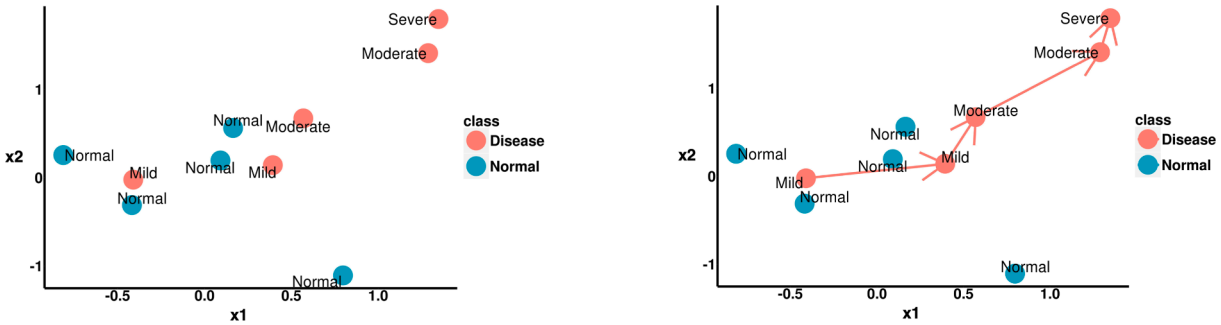
- Lack of information regarding disease progression

Another challenge is the lack of label information to specifically point out the stages of the disease progression. Labeling subjects by the trained medical professionals at each time point, i.e., the information regarding the stage in Fig. 2a, is almost impossible and expensive. In many cases, the only given label information for a subject's longitudinal data is the final diagnosis at the end of a clinical study. Furthermore, subjects' irregular and asynchronous visits as well as the varying disease progression rates make the problem more intractable. For longitudinal dataset with label only on the last time point, it's difficult to apply existing classification methods on the data points observed prior to the last one, since we have no information indicating from which time point the patients start to behave differently from the normal controls.

In contrast to existing methods that need labels of the patients on all the time points, here, we develop an approach that can extract the "change" information from the original data points, and seek to learn the disease progression over time. We have the intuition that although patients at the early stage may not be separable from the normal controls using static measurements if we focus on the magnitude or scale of the measurements, the change patterns over time may separate the two groups, as presented in Fig. 2b.

Fig. 3 demonstrates that such a transformation from the original data to the changes over time enables clear separation between the two classes. Moreover, the changes accumulated over larger time intervals are more separable between the two classes, since they contain more information regarding the disease progression. Meanwhile, since the "change" information is measured based on the different time points within the same subject, the synchronization of the visits across the subjects is not required.

To articulate this intuition, this paper proposes a flexible mixed-



(a) The patient at the early stage behave similarly as the normal control.

(b) There exists a trajectory of the disease progression for longitudinal data

Fig. 2. How can we train a detector (i.e., for early diagnosis and risk monitoring) with the dataset where most points are not separable?

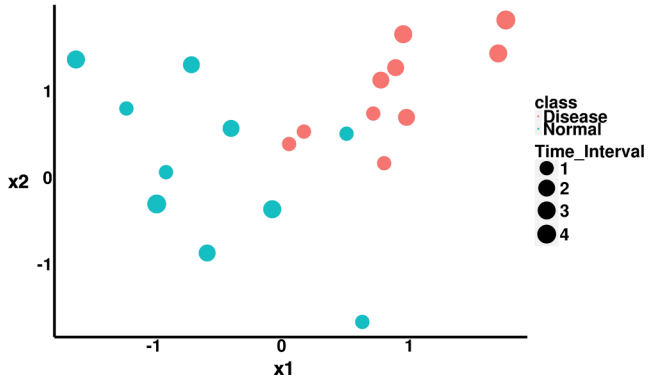


Fig. 3. New data is generated based on the same data in Fig. 2 by transforming the original time points to the change over time: $\widetilde{x}^i_{tt'} \equiv \delta\Phi(x^i_t, x^i_{t'})$, $t > t'$. $\widetilde{x}^i_{tt'}$ is the transformed data point, x^i_t and $x^i_{t'}$ are two original data points from subject i, and $\delta\Phi$ can be any function for measuring the change from t' to t. In this figure, it's simply $\widetilde{x}^i_{tt'} = x^i_t - x^i_{t'}$. The size of the points indicates the length of the time intervals. It can be shown that the change accumulated over large time intervals is more obvious between the two classes.

kernel method, called EDRA (Early Detection and Risk Assessment), which is based on the Structured Output Support Vector Machine (SOSVM) [60] extended to longitudinal data analysis with partial label information. By capturing the pattern of the disease progression over time instead of looking at a single data point, our method is able to achieve better disease diagnosis at the early stage. Another contribution of our method is that it can provide contemporaneous risk assessment of the disease. Meanwhile, EDRA inherits the advantages of SOSVM, including the rescaling of the penalty placed on the misclassification, which enables the smooth and monotonic trajectories for the predicted scores with the proper selections of loss rescaling functions. The properties of smoothness and monotonicity are crucial to reflect the contemporaneous underlying risk over time for slowly progressive diseases such as chronic diseases.

EDRA has the following advantages. First, it achieves early diagnosis with high accuracy. Second, it addresses the disease label/information inefficiency problem for the chronic disease with longitudinal data. Third, it enables contemporaneous risk assessment for tracking the disease/drug-induced progression. Last but not least, it provides a flexible mixed-kernel framework which constructs the kernel as a linear combinations of weighted "sub-kernels" each containing one feature or a subset of features, to take advantage of the prior knowledge about the features. Experiments of varying complexities were conducted to analyze our method performance, including (1) early detection and contemporaneous risk assessment using the simulated irregular and partially-labeled longitudinal data with features that are equally/differently predictive; (2) early detection and contemporaneous risk estimation with irregular longitudinal T1D RNA-Seq gene expression data; (3) monitoring of drug's long-term effect on patients based on longitudinal RNA-Seq gene expression data with missing time points. Our paper is organized as follows. In Section 2, we will review and discuss the related works in literature. In Section 3, the proposed mix-kernel framework for training contemporaneous disease risk detector with longitudinal data will be presented and the corresponding algorithm will be derived. Specifically, in Section 3.4, we will analyze and discuss the properties of the trained risk detector. In Section 4, the performance of our method will be demonstrated and validated on two synthetic datasets and two real-world applications as described above. Finally (Section 5), we will conclude our work and introduce the directions of our future study.

2. Related works

Our method is related to the topics in literature of computer-aided diagnosis methods, longitudinal clinical data analysis, and structured-output learning. Different from these methods, our method can handle irregular longitudinal data with partial label information, focusing on not only early diagnosis, but also contemporaneous monitoring of the disease progression.

2.1. Computer-aided diagnosis methods

Classification methods are widely used in computer-aided diagnosis. Many classification methods care about finding optimal hyperplanes to best separate data from different groups, whereas other methods such as Bayesian methods achieve classification based on probabilistic models. Classic classification methods are frequently applied in DNA micro-array and RNA-seq gene expression data analysis, among which there are statistical methods such as Linear Discriminant Analysis (LDA) [49,1], Quadratic Discriminant Analysis (QDA) [3] and Optimal Scoring (OS) [14,20], which seek to find another space where the between-class covariance is maximized while the within-class covariance is minimized; Bayesian methods including Naive Bayes and Bayesian Networks classifier apply Bayes rule for the inference of classes [56,50,36]; Machine learning methods like Support Vector Machine (SVM) [28,1,46,42] and ensemble learning methods such as Random Forest (RF) [7,47,35], representing modern techniques, are commonly applied in computer-aided disease diagnosis because of their robust performance. More complicated models are considered to address diverse range of challenges and specific complexities in some applications. For example, Zhou et al. formulate the prediction problem as a multi-task regression problem to predict the longitudinal outcomes for Alzheimer's disease based on the static baseline MRI features [70]. Multi-model frameworks are proposed to combine data of different types, e.g., Chen et al. propose a convolutional neural network (CNN)-based multimodel disease prediction algorithm using structured and unstructured data [12]. In [68], Zhang et al. propose a multimodel classifier combining three modalities of biomarkers to classify Alzheimer's disease (AD) or its prodromal stage (i.e., mild cognitive impairment (MCI)) from the healthy controls. Various recurrent neural networks (RNN)-based approaches have been developed for temporal data analysis. GRU-D, that is based on Gated Recurrent Unit (GRU), proposed by Che et al. to address the missing values problem in time series data by utilizing the missing patterns to achieve better prediction results [9]. Choi et al. propose Doctor AI, a temporal model using recurrent neural networks (RNN) that was applied to longitudinal time stamped electronic health record (EHR) data to leverage large historical data to make multilabel predictions (one label for each diagnosis or medication category) for patients' subsequent visits [13].

Nevertheless, most of the methods discussed above are supervised learning, it's difficult to directly apply these methods on partially labeled data. Moreover, comparing to these methods, our objectives are different, since we not only aim at discriminating classes, but also contemporaneously estimating the underlying risk scores with the irregular longitudinal data.

2.2. Longitudinal clinical data analysis

Longitudinal study is widely used in diagnosis, prediction and monitoring of the disease, that involves repeated observations of same variables over short/long period of time. There exist many time series models applied in longitudinal clinical data analysis. State-spaced models focusing on latent states inference, such as HMM and Linear Dynamic Systems (LDS) with its variants including Kalman filter, have been proved to be useful for the prediction of the disease progression [34,41,40]. Among this line of efforts, HMM-based methods are widely used in clinical data analysis. For instance, Wang et al. propose a continuous-time HMM-based model that learns a continuous-time progression model from discrete-time observations with non-equal intervals to address the problems like irregularity and the incompleteness of

the observation [64]. Jackson et al. develop a multistage Hidden Markov Model and apply it to an aneurysm screening study [23]. Sukkar et al. apply Hidden Markov Model to Alzheimer's disease [58]. Trajectory studies including Fixed/Random/Mixed-effect models, Latent Growth Mixture Modeling (LGMM), Latent Class Growth Modeling (LCGM) have been increasingly recognized for their usefulness for identifying homogeneous subpopulations within the larger heterogeneous population [44,51,26,57,19,54]. However, most of these methods aim at either prediction or discrimination, which is not enough to cover our objectives, nor are they feasible for the cases where the clinical data is of high dimension. Ke et al. exploit the low-rank property of a spatial-temporal matrix via the bilinear formalism and further use the matrix completion technique to fill the missing data for predicting the time to SSI onset by using dynamic data [29]. A least-square loss function as well as a squared hinge loss function are contained in their proposed bilinear formulation to obtain an unbiased learning formulation with complete and censored samples. Although their problem shows some relevance to ours, they put more focus on prediction. Meanwhile, the continuous measurements are required for constructing the spatial-temporal matrix, whereas our method focuses more on contemporaneous risk assessment and can deal with data points with irregular time intervals.

In the field of temporal predictive pattern learning, there have been efforts to extend supervised learning to time series data analysis to summarize and represent this complex time-series data in order to make them amenable to statistical analysis and modeling. Temporal predictive pattern mining techniques are developed to improve the classification of time series data and can be applied on the identification of the onset of disease. They usually aim to mining the predictive temporal patterns or extracting the time series shapelets to be the alternatives of the original features. These methods are usually applied as a preprocessing step prior to classification or regression, or sometimes can be directly used as the detectors [69,4,67,59].

Most of these methods, however, hold the assumptions that the data points are sampled on regular time points. Thus they are not suitable for the data with irregular time intervals, asynchronous visits and varying disease progression rates like our case. What's more, the underlying risk of disease we seek to monitor is not directly observed, so that it cannot be easily captured by temporal predictive pattern learning techniques. More importantly, high dimensionality of the time series data poses great challenge to this line of methods since mining high dimensional time series data directly is very expensive in terms of both processing and storage cost.

To address the challenges brought by the high dimensionality of the time series data, various works are presented in literature on developing representation techniques that can reduce the dimensionality of time series, while still preserving the fundamental characteristics of a particular data set [65]. High-level representations such as Discrete Fourier Transformation (DFT) [17], Singular Value Decomposition (SVD), Discrete Wavelet Transformation (DWT) [8], Piecewise Aggregate Approximation (PAA) [31], Adaptive Piecewise Constant Approximation (APCA) [32] were considered previously. In conjunction of these techniques, different similarity-based approaches represent a promising direction of time series analysis. For instance, Dynamic Time Warping (DTW), introduced by Berndt and Clifford [30], and its variants such as Weighted DTW (WDTW) that adopts a weighting scheme [24] and Derivative DTW (DDTW) that uses the difference between consecutive time values [33], are classical speech recognition tools allowing a time series to be "streched" or "compressed", that are considered to be strong for many time series data problems [6]. Another group of similarity measures for time series such as LCSS (Longest Common SubSequence) [63], EDR (Edit Distance on Real sequence) [11] and ERP (Edit Distance with Real Penalty) [10] have been developed based on the concept of the edit distance for strings [65]. More recent works for similarity measurement adopt tree-based methods to increase the robustness and the parameters tuning problems. TCK (time series cluster kernel) proposed by Mikalsen et al., leverages the missing data handling properties of Gaussian mixture models (GMM) augmented with informative prior distributions, and uses an ensemble learning approach to ensure robustness to parameters by combining the clustering results of many GMM to form the final kernel [43]. Baydogan et al. propose a method to model the dependency structure in time series that generalizes the concept of autoregression to local autopatterns, which generates a pattern-based representation along with a similarity measure called learned pattern similarity (LPS). Moreover, it adopts a tree-based ensemble-learning strategy that is fast and insensitive to parameter settings [5].

Another category of methods for longitudinal data analysis builds upon principal component analysis [41,45,53]. The general idea is to use a factor-analytic, or principal-component type analysis to first reduce the dimensionality of the response vector, and then, use standard longitudinal models for the analysis of the latent variables [62]. The drawback of these methods is that when applied to longitudinal data, bias will be introduced in principal factors by the within-individual effects, since the estimated covariance is the sum of the covariance of interest caused by disease progression and the unwanted covariance of the within individual effects due to the repeated measurements.

2.3. Structured-output learning

Since our method is based on SOSVM, we review the previous works on structured-output learning methods and their applications in this section.

Let us first review the basic idea of Support Vector Machine (SVM). SVM is a popular supervised learning method for classification by looking for optimal hyperplanes so that the projected data from different groups could have the largest separations. SVM allows misclassification by including "slack variables" for each training sample, and aims at minimizing the sum of the slack variables in the objective function. Given a training dataset (x_1, y_1), ...,(x_n, y_n), where y_i is the label indicating the class that the data point x_i belongs to, and x_i is a p dimensional vector. Let $F(x_i, y_i;w, b)$ be the score of x_i, where w denotes the parameter vector, and b is the intercept, the objective function of a soft-margin SVM can be written as:

$$\begin{aligned} &\min_{w,\xi_i} \quad \frac{1}{2}\left\|w\right\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i \\ &\text{s.t.} \quad F(x_i, y_i;w, b) \geqslant 1-\xi_i, \quad \xi_i \geqslant 0 \\ &\qquad \forall\, i = 1, ..., n \end{aligned} \tag{1}$$

Here we may ask two questions: (1) Instead of assigning labels, can we also give a confidence level about the classification results? (2) Given partial information about the labels, how can we apply it in semi-supervised scenarios where the label information is not available for each time point?

Tsochantarid et al. propose SOSVM [60], a general framework which extends SVM to the scenario where there exists some structure of the output classes. SOSVM's approach is to rescale the slack variables according to the loss incurred in each of the linear constraints:

$$\forall\, i, \forall\, y \in \mathcal{Y}\backslash y_i \quad f\left(x_i, y_i;w\right) - f\left(x_i, y;w\right) \geqslant 1 - \frac{\xi_i}{\Delta(y_i, y)} \tag{2}$$

where $f(x_i, y_i;w)$ is the same as the score function $F(x_i, y_i;w, b)$ with the intercept parameter b excluded and $\Delta(y_i, y)$ is the slack variable rescaling function, which measures the loss incurred by the misclassification of the true label y_i by $y \in \mathcal{Y}\backslash y_i$.

Methods related to learning using privileged information are extensively studied recently (e.g. [37,48,61]). Although these methods take output structures into account, not only the labels but also the privileged information such as the rankings of the labels are required for the training, and most of them are not specifically designed in

longitudinal data analysis. Hoai et al. [21] adopt the idea from SOSVM and apply it on computer vision for early detection of temporal events. However, it's difficult to directly apply their method since the detailed label information about the target events for training is needed. Huang et al. [22] consider longitudinal data with partial labels, but they apply same weights for all slack variables and don't take the advantage of rescaling loss functions as the SOSVM-based methods mentioned above to model the irregular time intervals.

3. Early disease detector and risk estimator

As we described above, most existing methods are not designed for early diagnosis and risk assessment of disease with partially labeled longitudinal data. In this section, we propose a learning formulation to address this problem.

3.1. Notations

Let $(X^1, y^1), \ldots, (X^i, y^i), \ldots (X^n, y^n)$ be the set of longitudinal data with the diagnosis result made on the last time point, where $y^i \in [1, -1]$ is the final diagnosis result for the ith patient and X^i is a matrix drawn from the input domain $\mathcal{X} \in \mathcal{R}^{T_i \times p}$, which includes the measurements for subject i with T_i visits in total. X^i can thus be represented as $X^i = \begin{bmatrix} x^i_{t_1} \\ \vdots \\ x^i_{t_{T_i}} \end{bmatrix}$, in which $x^i_{t_l} \in \mathcal{R}^{1 \times p}$ denotes the p measurements of the lth visit at time t_l for patient i.

There's a record of the visiting times for each subject: $T = \{T[1], \ldots, T[n]\}$, where $T[i]$ records the visiting time of patient i, i.e., $T[i] = [t_0, t_1, t_2, \ldots, t_{T_i}]$. For instance, T can be the number of months for each follow up after the initialization of the drug therapy; it can also be the number of months prior to the diagnosis. Please note that the visiting times of a patient can be irregular and asynchronous.

3.2. Feature representation in mixed kernel space

In order to provide a flexible framework for taking advantage of the prior knowledge about the rankings of features' discriminating power, apart from directly applying "kernel trick" on the original data to project it to the kernel space, we constructed a kernel as the linear combinations of "sub-kernels" each containing only one feature: $K(x, x') = \sum_{d=1}^{p} \beta_d K_d(x, x') = \sum_{d=1}^{p} \beta_d \langle \Phi_d(x), \Phi_d(x') \rangle$, where $\Phi_d(x) = \Phi(x_d)$, only works on the dth feature of x. β is a vector of dimension p for the feature weights, and it satisfies $\sum_{d=1}^{p} \beta_d = 1$.

To measure the augmented information till time t_l, we check both the cumulative moving average and the running total in our experiments to obtain the information augmented until time t_l, which have been applied in the implementation of MMED (Max-Margin Early Event Detectors) [21]. However, we decide to use the cumulative moving average to obtain the augmented information in our method for the following reasons: (1) we would like to smooth out the short-term fluctuations; and (2) different from MMED that aims at localizing the interval for an event, we care more about the risk at a time point given the cumulative information prior to that. The representation can be written as:

$$X^i_{t_l} = \overline{X^i_{[1:l]}} = \frac{1}{l} \sum_{s=1}^{l} x^i_{t_s}$$

Let $\Phi(X^i_{t_l})$ denotes the projection of $X^i_{t_l}$ in the kernel space:

$$\Phi(X^i_{t_l}) = diag\left(\sqrt{\beta_1}, \ldots, \sqrt{\beta_p} \right) \begin{bmatrix} \Phi_1(X^i_{t_l}) \\ \vdots \\ \Phi_p(X^i_{t_l}) \end{bmatrix}$$

With this representation, the similarity assessment of the information between the two subjects i and j at the time points l and l', respectively, can be represented by $K(X^i_{t_l}, X^j_{t_{l'}}) = \sum_{d=1}^{p} \beta_d K_d(X^i_{t_l}, X^j_{t_{l'}}) = \sum_{d=1}^{p} \beta_d \Phi_d(X^i_{t_l})^T \Phi_d(X^j_{t_{l'}})$.

3.3. Learning with longitudinal data

Recall that instead of learning individual data points, we identify the signatures of the disease/drug-induced changes to address the problem of inseparability and label insufficiency.

First consider a linear function $g(\delta\Phi_i(t_l, t_{l'}); w) = \langle w, \delta\Phi_i(t_l, t_{l'}) \rangle$, where $\delta\Phi_i(t_l, t_{l'})$ is the shorthand defined as $\delta\Phi_i(t_l, t_{l'}) \equiv \Phi(X^i_{t_l}) - \Phi(X^i_{t_{l'}})$, which is for measuring the changes of the ith subject from time $t_{l'}$ to t_l in the mixed-kernel space. The function $g(\delta\Phi_i(t_l, t_{l'}); w)$ is expected to have the following properties:

$$\forall\, i, \quad \forall \begin{bmatrix} l', l \end{bmatrix} \in L_i, \begin{cases} g(\delta\Phi_i(t_l, t_{l'}); w) \geqslant 0, y^i = 1 \\ g(\delta\Phi_i(t_l, t_{l'}); w) \leqslant 0, y^i = -1 \end{cases}$$

where $L_i = \{[1, 2], [1, 3], \ldots, [T_i - 1, T_i]\} \cup \{[0, T_i]\}$ contains all the pairwise combinations of the visit index for subject i, for i in $1, \ldots, n$.

In the framework of SOSVM, the loss of misclassifying x^i to a class $y \in \mathcal{Y} \backslash y^i$ is rescaled by a non-negative weight function $\Delta(y, y_i)$, i.e., $\Delta: \mathcal{Y} \times \mathcal{Y} \to \mathcal{R}$, and it quantifies the loss associated with a prediction y, if the true output value is y_i[60]. It's saying that with the prior knowledge about the structure of the output y and y_i, we put greater penalty for the misclassification if $\Delta(y, y_i)$ is large when training the classifier.

In our early detection case on longitudinal data, $\Delta(y, y_i)$ here can be a function with respect to the time interval between two time points: $\mu(t_l, t_{l'})$. More strict classification rules should be applied for larger time intervals, so that the penalty $\mu(t_l, t_{l'})$ placed on the misclassification should be greater when the two time points are far from each other. The design of function μ will be discussed in detail in the later context.

The desired constraints then become:

$$\forall\, i, \quad \forall \begin{bmatrix} l', l \end{bmatrix} \in L_i, \quad y^i g\left(\delta\Phi_i\left(t_l, t_{l'} \right); w \right) \geqslant 1 - \frac{\xi_i}{\mu(t_l, t_{l'})} \tag{3}$$

Together with the goal of max-margin hyperplane, we obtain the following objective function:

$$\begin{aligned} &\min_{w, \xi_i, b} \frac{1}{2} \left\| w \right\|^2 + \frac{C}{n} \sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad &y^i \left\langle w, \delta\Phi_i\left(t_l, t_{l'} \right) \right\rangle \geqslant 1 - \frac{\xi_i}{\mu(t_l, t_{l'})}, \quad g\left(\Phi(X^i_{t_0}); w \right) = -b, \quad \xi_i \geqslant 0 \\ &\forall\, i, \quad \forall\, [l', l] \in L_i \end{aligned} \tag{4}$$

The constraints containing b are active only for cases $[l', l] = [0, T_i]$, where $l' = 0$ is a virtual time point, so that the constraints for cases $[l', l] = [0, T_i]$ shrink to the constraints of a standard soft-margin SVM.

3.4. Properties of EDRA

In this section, let us analyze several properties of the scores assigned by the risk detector learned with the above objective function.

- Monotonicity

To develop EDRA, we focus on the early detection and contemporaneous risk estimation for the disease/drug-induced progression prior to the diagnosis, for which we utilize the monotonic progression characteristic (either towards disease or recovery) as the model assumption to learn EDRA. For instance, as shown in Fig. 1, functional beta-cell mass declines as T1D progresses. For the degenerative disease

conditions such as Alzheimer's disease, the underlying disease degradation process is also monotonic. This generative nature leads to the monotonic assumption of EDRA.

Here we may ask such question: After we obtain the reliable detection of the changes from the time intervals of a subject, how can the risk scores reflect the progressive property of the underlying disease progression for each time point?

Based on the linear property of function g, the constraints (4) can be rewritten as:

$$\forall\, i, \quad \forall \left[l', l\right] \in L_i, \quad y^i\left\{g\left(\Phi(X^i_{t_l});w\right) - g\left(\Phi(X^i_{t_{l'}});w\right)\right\} \geqslant 1 - \frac{\xi_i}{\mu(t_l, t_{l'})} \tag{5}$$

The learning formulation actually naturally enforces monotonicity of the detector function. Moreover, the function μ is desired to have the following properties: (1) $\mu(t_l, t_{l'}) \in (0, 1)$, and (2) $\mu(t_l, t_{l'}) \propto |t_l - t_{l'}|$, to serve as a rescaling function to adjust the penalty for the misclassification based on the distance between two time points. In our study, we set $\mu\left(t_l, t_{l'}\right) = 1 - e^{-\left(\frac{t_l - t_{l'}}{\sigma}\right)^2}$, where σ is a tuning parameter. The proposed learning formulation achieves the monotonicity with respect to the information contained within the time intervals accounting for the disease/drug-induced progression. Such learning formulation provides a flexible framework that is able to deal with irregular time intervals, and enables not only the property of monotonicity, but also the property of smoothness for the trajectories of the predicted scores, which will be discussed in the following context. Both of these properties reflect the progressive property of the chronic disease and drug response.

- Smoothness

A smooth trajectory of the risk scores assigned to one subject over time is desired, since usually in the real case, the disease progresses gradually, so that the difference between the risk scores of two close neighbor time points should be relatively small. The smoothness of the trajectory can be controlled by the design of the slack variable rescaling function μ, which is used to adjust the penalty of the misclassification in (3) and (4). Since $\mu\left(t_l, t_{l'}\right) = 1 - e^{-\left(\frac{t_l - t_{l'}}{\sigma}\right)^2}$, when two time points are very close, the penalty of the misclassification is close to zero, i.e., $\mu(t_l, t_{l'}) \to 0$, when $t_{l'} \to t_l$. This enables the smoothness of the risk score trajectories for the subjects, since the disease/drug-induced progression contained in a very small time interval is very limited, so that the difference between the predicted scores of two very close time points should be relatively small compared to the ones of the large intervals.

With the linear property of function g, $g(\delta\Phi_i(t_l, t_{l'});w) = g(\Phi(X^i_{t_l});w) - g(\Phi(X^i_{t_{l'}});w)$, so that we have:

$$R(X^i_{t_l}) - R(X^i_{t_{l'}}) = g(\delta\Phi_i(t_l, t_{l'});w) \to 0 \tag{6}$$

for the cases when $(t_l - t_{l'}) \to 0$

- Separation

The risk scores can be wrongly estimated if we only care about the difference of the scores between two time points since either one of them can start from or end up in a random place. It's important to "fix" at least one point of the whole trial so that the predicted score of which can separate the two classes. In our study, the detector should be trained to be able to classify the last single time point, since the only label we have is the diagnosis at the end of the clinical trial.

In contrast to the smoothness property with the rescaling penalty function $\mu(t', t) \to 0$ when $t' \to t$, the penalty placed on the misclassification is scaled to be the highest for the greatest time interval of the ith subject, i.e., $[l', l] = [0, T_i]$, since the information augmented from the initial time point till the last one reaches the maximum.

Recall that we have the constraint $g(\Phi(X^i_{t_0});w) = -b$, so that the constraints regarding $[l', l] = [0, T_i]$ in (4) turn out to be the constraints of a soft-margin SVM:

$$\begin{aligned} y^i g\left(\delta\Phi_i\left(t_{T_i}, t_0\right);w\right) &= y^i\left\langle w, \Phi\left(X^i_{t_{T_i}}\right) - \Phi(X^i_{t_0})\right\rangle \\ &= y^i\left(\left\langle w, \Phi\left(X^i_{t_{T_i}}\right)\right\rangle + b\right) \geqslant 1 - \frac{\xi_i}{1 - e^{-\left(\frac{t_{T_i}}{\sigma}\right)^2}} \end{aligned} \tag{7}$$

The problem thus shrinks to a standard SVM classifier training problem. This constraint is to model the real case where the diagnosis is only available at the end of the study. With constraint (7) the trajectories of the two groups are enforced to depart from each other as the disease progresses.

3.5. Optimization: dual problem and algorithm

To solve the primal problem (4), first we move the constraints to the objective function to obtain the Lagrangian form:

$$\begin{aligned} &\max_{\alpha,\zeta}\min_{w,b,\xi} L\left(w, b, \xi, \alpha, \zeta\right) \\ &= \frac{1}{2}\left\|w\right\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\sum_{[l',l]\in L_i}\alpha^i_{l,l'}\left[1 - \frac{\xi_i}{\mu(t_l, t_{l'})} - y^i\left\langle w, \delta\Phi_i\left(t_l, t_{l'}\right)\right\rangle\right] \\ &\quad - \sum_{i=1}^{n}\zeta_i\xi_i \\ &\qquad \text{s.t.} \quad \forall\, i, \quad \forall\, [l', l] \in L_i \quad \xi_i \geqslant 0, \quad \zeta_i \geqslant 0, \quad \alpha^i_{l,l'} \geqslant 0 \end{aligned} \tag{8}$$

The third part which is related to α is the sum of the terms regarding the changes detection and the last time point classification:

$$\begin{aligned} &\sum_{i=1}^{n}\sum_{[l',l]\in L_i\setminus[0,T_i]}\alpha^i_{l,l'}\left[1 - \frac{\xi_i}{\mu(t_l, t_{l'})} - y^i\left\langle w, \delta\Phi_i\left(t_l, t_{l'}\right)\right\rangle\right] \\ &+ \sum_{i=1}^{n}\alpha^i_{T_i,0}\left[1 - \frac{\xi_i}{\mu(t_{T_i}, 0)} - y^i\left(\left\langle w, \Phi(X^i_{T_i})\right\rangle + b\right)\right] \end{aligned} \tag{9}$$

To derive the dual problem, we need to minimize the Lagrangian form with respect to w, b and ξ to get:

$$\begin{aligned} &\max_{\alpha^i_{l,l'}} \sum_{i,[l',l]\in L_i}\alpha^i_{l,l'} \\ &- \frac{1}{2}\sum_{i,[l',l]\in L_i}\sum_{j,[\tilde{l}',\tilde{l}]\in L_j} y^i y^j \alpha^i_{l,l'}\alpha^j_{\tilde{l},\tilde{l}'}\left\langle \delta\Phi_i\left(t_l, t_{l'}\right), \delta\Phi_j\left(t_{\tilde{l}}, t_{\tilde{l}'}\right)\right\rangle \\ &\qquad \text{s.t.} \quad \forall\, i \quad 0 \leqslant \sum_{[l',l]\in L_i}\frac{\alpha^i_{l,l'}}{\mu(t_l, t_{l'})} \leqslant \frac{C}{n}, \quad \sum_{i=1}^{n} y^i\alpha^i_{T_i,0} = 0 \end{aligned} \tag{10}$$

The inner product of $\delta\Phi_i(t_l, t_{l'})$ and $\delta\Phi_j(t_{\tilde{l}}, t_{\tilde{l}'})$ can be expanded as:

$$\begin{aligned} &\left\langle \delta\Phi_i(t_l, t_{l'}), \delta\Phi_j(t_{\tilde{l}}, t_{\tilde{l}'})\right\rangle \\ &= K\left(X^i_{t_l}, X^j_{t_{\tilde{l}}}\right) - K\left(X^i_{t_l}, X^j_{t_{\tilde{l}'}}\right) - K\left(X^i_{t_{l'}}, X^j_{t_{\tilde{l}}}\right) + K\left(X^i_{t_{l'}}, X^j_{t_{\tilde{l}'}}\right) \end{aligned} \tag{11}$$

Specifically, all terms $K(X^i_{t_{l'}}, \cdot)$ with $l' = 0$ are set to be zero, since $l' = 0$ is the virtual time point. When $[l', l] = [0, T_i]$ and $[\tilde{l}', \tilde{l}] = [0, T_j]$, we have $\left\langle \delta\Phi_i\left(t_{T_i}, t_0\right), \delta\Phi_j\left(t_{T_j}, t_0\right)\right\rangle = \left\langle \Phi\left(X^i_{t_{T_i}}\right), \Phi\left(X^j_{t_{T_j}}\right)\right\rangle$, which is of the same form as a standard SVM problem.

One challenge of the above dual problem is that the number of constraints is very large and thus the computation complexity of the optimization is high. To relieve this problem and speed up the

algorithm, we use constraint generation (cutting plane algorithm) [29] to handle the large set of constraints in the original problem (4). The outline of the algorithm is described as Algorithm 1.

4. Experiments

This section describes our experiments on two synthetic datasets and two real-world datasets of varying complexity: (1) Simulated longitudinal data considering irregularity in observation time with features of equal/different predictive power; (2) Irregularly sampled T1D longitudinal RNA-Seq gene expression dataset from TrialNet; (3) Longitudinal RNA-Seq gene expression dataset with missing time points for IFNβ drug response. In this paper, both the real-world longitudinal datasets used in our experiments to evaluate the performance are RNA-Seq data, but our method can also be applied to other clinical data where the longitudinal data analysis is involved. The performance of our method is evaluated regarding how early the detection of the disease can be made and how well the risk scores reflect the actual disease progression.

4.1. Evaluation

In our experiments, we evaluate the performance of our method based on two criteria: (1) The earliness of detection, (2) The correlation between the risk scores with the disease progression. We use the area under the ROC curve (AUC) over the normalized time points for benchmarking the earliness of detection when comparing our method with other algorithms, and we plot the risk scores over time for evaluating the performance of our method as a contemporaneous risk monitoring tool for the disease progression.

Algorithm 1. Algorithm for solving the dual problem (10) of EDRA

Input: $(X^1, y^1), \ldots, (X^n, y^n)$, β, T, L, C, ϵ.
Output: α, b.
1: *Initialize*: α, $\xi \leftarrow 0$ and $S \leftarrow \varnothing$
2: **repeat**
3: $V \leftarrow \varnothing$
4: **for** $i = 1$ to n **do**
5: Compute the loss for all $[l', l] \in L_i$

$H(l', l) \equiv (1 - y_i\langle w, \delta\Phi_i(t_l, t_{l'})\rangle)\mu(t_l, t_{l'})$

6: where $w = \sum_{i=1}^{n} \sum_{[l',l]\in L_i} \alpha_{l,l'}^i y^i \delta\Phi_i\left(t_l, t_{l'}\right)$
7: Find the most violated constraint:

$$\left[\hat{l}', \hat{l}\right] = \max_{[l',l]\in L_i} H\left(l', l\right)$$

8: $\xi_{i_{new}} = \max\{0, H(\hat{l}', \hat{l})\}$
9: **if** $(\xi_{i_{new}} \geqslant \xi_i + \epsilon)$ **then**
10: $\xi_i \leftarrow \xi_{i_{new}}$
11: $V \leftarrow V \cup \{[\hat{l}', \hat{l}]_i\}$
12: $\alpha \leftarrow$ optimize dual problem (10) over $S = S \cup V$
13: **until** $V = \varnothing$

4.2. Time normalization

Since the visiting times of subjects can be irregular and asynchronous in many cases, the visiting time is normalized as the fraction of the whole trial to get better evaluation. For instance, the normalized time for the lth visit of the subject i can be represented as: $t = 1 - \frac{t_{T_i} - t_l}{L}$, where L is the length of the whole trial, i.e., the maximum length of all the subjects, so that the normalized time $t \in [0, 1]$. Since often in clinical settings the delta time to an event is more useful as it provides how early the event of interest can be estimated, we also consider to evaluate the performance based on the normalized delta time to an event (such as diagnosis/recovery) in our experiments, which can be represented as: $\Delta t = \frac{t_{T_i} - t_l}{L}$. When the subject reaches the last time point and receives the diagnosis ($t_l = t_{T_i}$), the normalized time $t = 1(\Delta t = 0)$. At the initiation of the trial, $t = 0(\Delta t = 1)$ for the subjects whose length of study equal L. This set up is for the cases where some of the subjects start to take the test early while some of them start late. For the subjects with late starting time or skipped visits, they may not be available on some certain normalized time points according to their actual skipped visits. Fig. 4 illustrates the time normalization for three subjects of different cases.

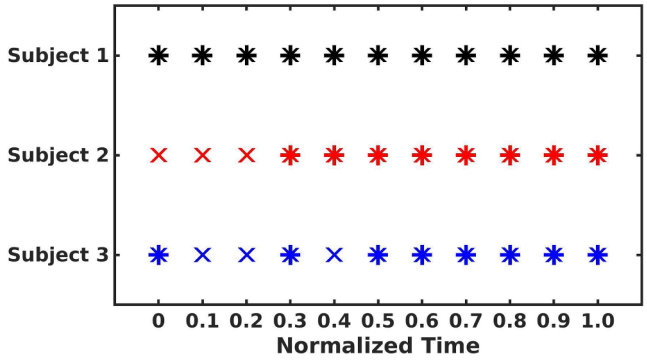


Fig. 4. Time Normalization: Asterisk symbol "$*$" denotes the available visits; Cross symbol "x" denotes the unavailable visits. Subject 1: early starting time without skipped visits; Subject 2: late starting time; Subject 3: skipped visits.

4.3. Simulation

We first validate the performance of our method on the synthetic longitudinal data. The synthetic longitudinal data is generated for 100 subjects in total, and each subject has different number of time points ranging from 12 to 14. The prior for the class of disease equals the prior for the class of the normal controls, which is 0.5. The disease progression is modeled by 4 different stages: Stage 0 (Normal), Stage 1 (Mild), Stage 2 (Moderate) and Stage 3 (Severe). For normal controls, they only stay in Stage 0 and will never proceed to the other three stages.

For patients, however, the disease progression is modeled by a Markov Chain model starting from either Stage 0, Stage 1, or Stage 2 and can proceed to more severe stages as disease develops, or it can start from one stage and skip the adjacent stage to directly jump to any one of the more severe stages (e.g., jump from Stage 0 to Stage 2/Stage 3). Specifically, to evaluate the robustness of the proposed approach on irregular longitudinal data, we randomly skipped the time points within one subject to model the irregularity in the observations, as shown in Fig. 6.

The lth visit of the ith subject's can be represented as:

$$x_l^i = \mu_s + \varepsilon^i + \varepsilon_l^i \tag{12}$$

where μ_s is a vector of mean values for the measurements including 4 features for the stage corresponding to the lth visit of the ith patient. The design of μ follows the structure of the stages. Further, linear and nonlinear co-existing predictive relationships are considered for generating the synthetic data. Fig. 5 illustrates the design of μ_s in our experiments.

The individual effects and the technical noise are modeled by ε^i and ε_l^i, respectively. For longitudinal dataset, it's necessary to model the "baseline" information ε^i for each subject, that won't change over repeated measurements, and is shared by all the time points of the subject i. The technical noise is modeled by ε_l^i, which varies among all the data points. Both ε^i and ε_l^i are drawn from multivariate normal distribution.

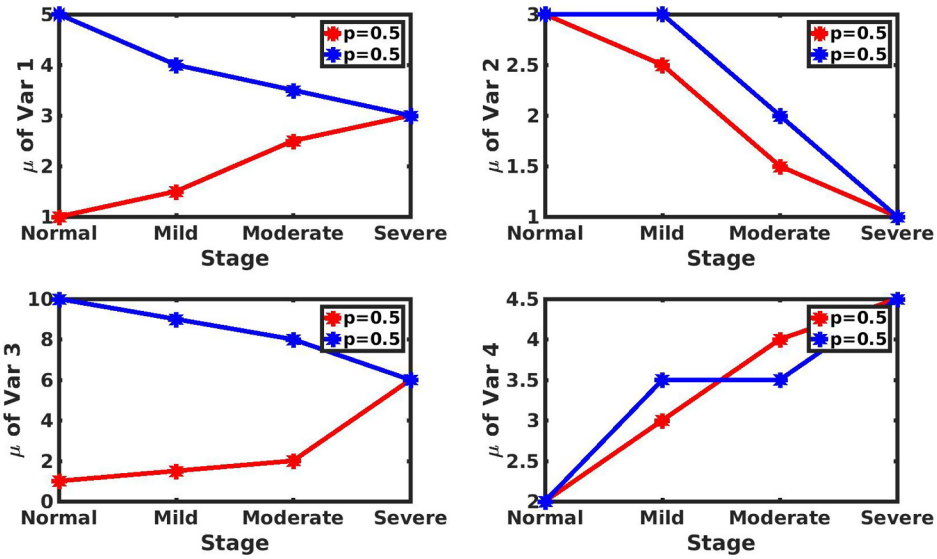


Fig. 5. Generation of synthetic data: design of μ for the 4 features as disease progresses over time. μ of variable 1 and variable 3 are designed to model the nonlinear predictive relationship, while variable 2 and variable 4 follow linear predictive relationship with different progression rates. The probability of choosing the pattern of the blue line is same as the red line, which equals 0.5. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We randomly divide the synthetic data into training and testing dataset. 80 percent of the generated synthetic data is contained in the training dataset, and the rest 20 percent is used as testing data for evaluating the performance.

In the first experiment, we evaluate the performance using the synthetic data with all features contributing to the discrimination of the two classes. The feature weights β_k are set to be same for the 4 features in the experiments of this synthetic dataset: $\beta_k = 0.25,\ k = 1,\ ...,4$.

We first investigate the performance of the risk assessment. Since we know the ground truth about the stages, the stage information is illustrated by different colors for better illustration. However, please note that the information regarding the stage is only used for demonstration, and it's not available when we train the models.

Fig. 6 provides two subjects from the testing dataset to illustrate how the trained detector monitors disease progression in longitudinal study for (1) a normal control stays at the "Normal" stage and (2) a patient goes through different stages over time. The curve of the risk scores over time is relatively flat for the normal control, and the predicted scores throughout the trial are less then zero. Nevertheless, for the patient with increasingly severe situation, the risk score increases as the disease progresses, and turns out to be positive since the third normalized time point of the trial.

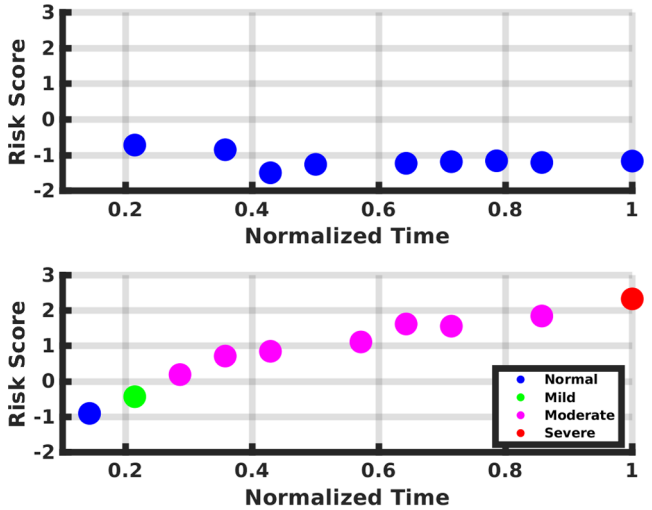


Fig. 6. Synthetic data experiments: Risk scores over time for the two subjects. Top: a normal control without disease; Bottom: a patient with 4 different stages.

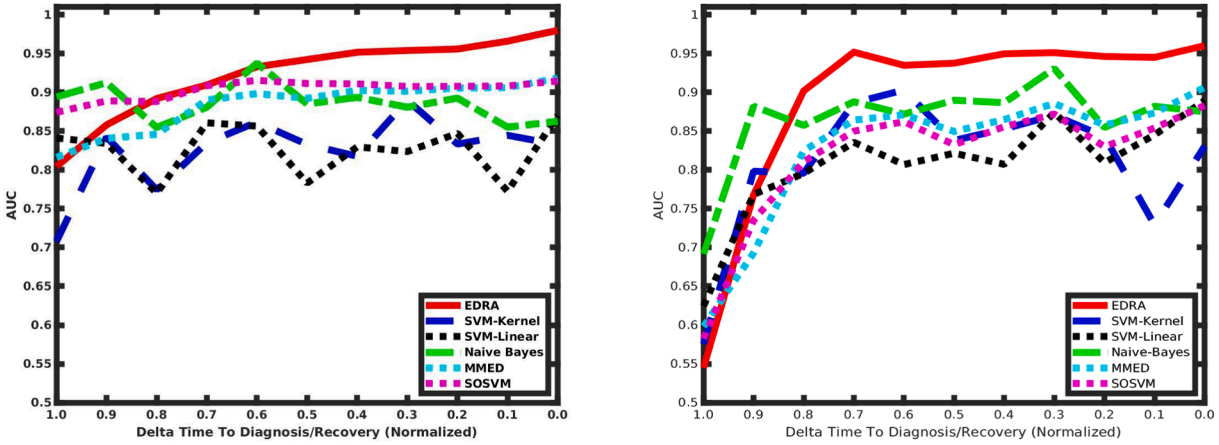
To further evaluate the effect of the mixed-kernel framework with consideration of the prior knowledge about the feature discriminating power, another synthetic dataset containing features with different predictive power is discussed in the following experiments. This synthetic dataset is simulated with two additional inactive features whose mean values for the measurements stay the same over different stages, to the original feature set. Therefore the feature weights for the kernel construction are: $\beta_k = 0.25$, for $k = 1, ...,4$ and $\beta_k = 0$, for $k = 5, 6$. Table 1 provides the detailed information about how β_k is determined for this experiment.

We analyze the earliness and accuracy of the detection by EDRA. We repeat our experiments 50 times and record the average performance. We randomly divide the synthetic data for training and testing each time as described above. To obtain better evaluation of the performance, we compare our method with three other popular classifiers: Linear SVM, Naive Bayes (NB) and Kernel SVM (RBF). When we train Linear SVM, Kernel SVM and Naive Bayesian classifier, since the only information about the label for the longitudinal data we have is the final diagnosis, we apply the final diagnosis result to the time points prior to the last one, i.e., given the time points of the subject i: $x_1^i, x_2^i, ..., x_{T_i}^i$, we assign the label $y_{T_i}^i$ for the last time point $x_{T_i}^i$ as the label to the other time points prior to that. Specially, since linear SVM, Naive Bayes (NB) and Kernel SVM are not designed for longitudinal data, we treat the data points independently, without considering the temporal structure within them.

In addition to the methods mentioned above, since our method is inspired by SOSVM, we compare our methods to SOSVM and another SOSVM-based method Max-Margin Early Event Detectors (MMED), which are more state-of-the-art approaches specifically designed for the early detection of temporal data analysis. We train and evaluate MMED

Table 1
Synthetic data experiment: Feature information

ID	Var1	Var2	Var3	Var4	Var5	Var6
Active	T	T	T	T	F	F
β	0.25	0.25	0.25	0.25	0	0



(a) Same predictive power (b) Different predictive power

Fig. 7. Synthetic data experiments: AUC over the normalized delta time to the diagnosis.

and SOSVM the same way the authors of MMED did in their experiments [21]. During the training of MMED and SOSVM, since both methods require the starting and ending time of an event to train the model for localizing the event of interest, so that we set the first time point as the starting point and the last one as the ending point of the event for an subject with disease; for healthy controls, we set the time interval for the event of interest to be empty. We follow MMED's implementation to perform the detection with MMED and SOSVM: given a data point at time t, we calculate the scores for all the data points prior to t, and use the highest score as the risk score indicating if an event has been happening until time t.

When applying the trained classifiers to the testing dataset, AUCs are calculated on each normalized delta time point. The curves of AUC over the normalized delta time points are depicted in Fig. 7. Fig. 7a demonstrates the AUC trajectories over the normalized delta time to the diagnosis based on the synthetic dataset with 4 active features. Fig. 7b provides the AUC trajectories over the normalized delta time to the diagnosis based on the synthetic dataset with features of different predictive power.

In Fig. 7a, at the beginning of the trial, EDRA performs similarly with Naive Bayes method but better than the other SVM-based methods. However, EDRA outperforms all the other methods at the last three time points of the trial, with AUC reaching 0.96 ± 0.05 at the end of the trial, while the AUCs of the other methods are 0.88 ± 0.08, 0.88 ± 0.07, 0.87 ± 0.09, 0.85 ± 0.11, 0.90 ± 0.07 for MMED, SOSVM, Linear SVM, Kernel SVM (RBF) and Naive Bayes, respectively. What's more, it can be seen that the SOSVM-based methods considering temporal structure, such as EDRA, MMED and SOSVM, successfully capture the disease progression with the smoothly increasing trajectories of AUCs over time, while the other methods failed in this point.

In Fig. 7b, EDRA outperforms the other methods by a large margin after the third normalized time point, which is much earlier than Fig. 7b. The AUC of EDRA keeps increasing till the last time point and ends up at 0.96 ± 0.04, while the AUCs for MMED, SOSVM, Linear SVM, Kernel SVM and Naive Bayes are: 0.91 ± 0.06, 0.88 ± 0.06, 0.89 ± 0.06, 0.83 ± 0.09 and 0.87 ± 0.06, respectively.

Comparing the earliness and the accuracy of the detection, EDRA outperforms the other methods. Regarding the contemporaneous risk assessment, it can be shown that the models considering the structure within the temporal data such as EDRA, MMED and SOSVM, capture the risk progression better with the smoother and increasing AUC trajectories over time, compared to the relatively fluctuating AUC trajectories by Linear SVM, Kernel SVM (RBF) and Naive Bayes. With the synthetic data incorporating different rates of irregularity in observations, the experiments show that the proposed method is robust to irregularly-sampled longitudinal data. The experiments also demonstrate that the mixed-kernel framework incorporating the prior knowledge about the features' discriminative power improves the performance compared to the methods without such consideration. For the experiments on this dataset, we perform 5-fold cross validation for determining the hyperparameter C for the SVM-based methods and the tuning parameter σ for the kernel construction.

4.4. Longitudinal T1D RNA-Seq data from TrialNet

This section describes our experiments on RNA-Seq gene expression dataset from TrialNet, which includes 42 subjects with the final diagnosis of T1D, and 37 subjects as normal controls. For each subject diagnosed to have diabetes, there are at least 3 time points and at most 11 time points. There's only one time point for each normal control. The pattern of the visiting time of the patients with multiple time points is irregular and asynchronous, and the time stamps are recorded by the months prior to the diagnosis, as illustrated in Fig. 8.

Since there are 16618 genes contained in the original dataset, we first perform differential expression test by edgeR [55] to identify 50 differential expressed genes (DEGs) that show differences in expression level between conditions for our experiments. The importance of each DEG is measured by the absolute value of the fold change (FC). The weight of the kth DEG is calculated based on the absolute value of the dth DEG's log_2FC and is normalized by the sum of the absolute values of log_2FC of all DEGs, i.e., $\beta_k = \frac{abs(log_2FC_k)}{\sum_{d=1}^{50} abs(log_2FC_d)}$.

In this experiment, since the only information regarding the disease situation is the medical diagnosis made at the last time point for each subject, how early our method can detect the disease prior to that time point is of great interest. To investigate the performance of our method

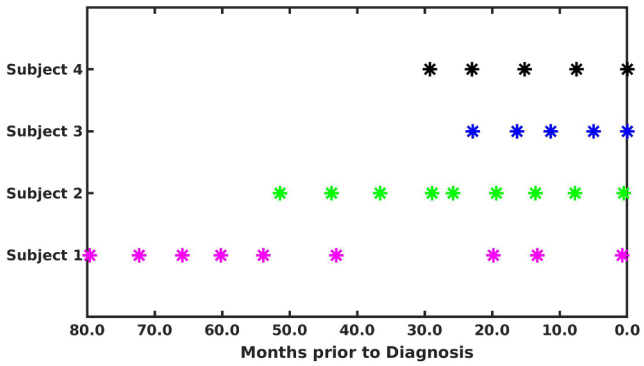
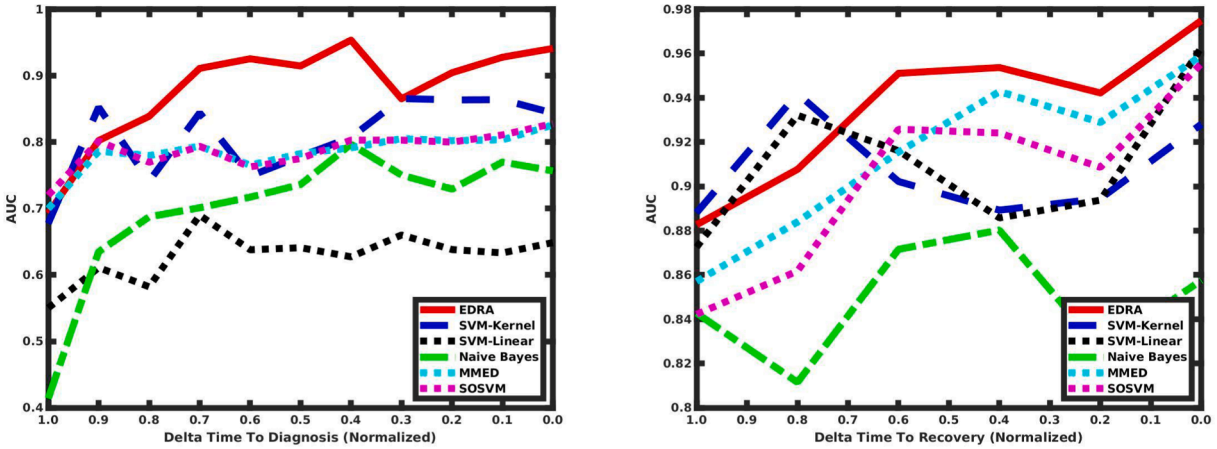


Fig. 8. Visiting time points (Months prior to diagnosis): "$*$" denotes the available visits.



(a) TrialNet: AUC over the normalized delta time to diagnosis

(b) IFNβ Drug Response: AUC over the normalized delta time to recovery

Fig. 9. Real-world data experiments: AUC over the normalized delta time to the diagnosis/recovery.

for early disease detection, we plot the curves of AUC over the normalized delta time points for benchmarking the earliness of the detection.

Similar to the simulation, we randomly divide the dataset for training and testing. The training dataset contains 80 percent of the whole data, and the testing dataset contains the rest 20 percent. We repeat our experiments 50 times and record the average performance.

During the testing, since each normal control only has one time point, we use the predicted scores for the single time point of the subjects without disease as the baseline scores, and compare the scores assigned for the patients at each normalized delta time point against the baseline scores. AUCs thus can be calculated for each normalized delta time point. The curves of AUC over the normalized delta time are depicted as Fig. 9a.

In this dataset, it's difficult to classify patients from the normal controls at the beginning due to the slow progression property of the chronic disease. However, at the second normalized time point (the eighth normalized delta time point), EDRA is able to detect the disease with the AUC of 0.84 ± 0.15, while the AUCs for MMED, SOSVM, SVM-kernel, SVM-linear and Naive Bayes at that time point are: 0.78 ± 0.17, 0.77 ± 0.16 and 0.74 ± 0.17, 0.58 ± 0.23 and 0.69 ± 0.19 respectively. In the end, EDRA still performs best, whose AUC is 0.94 ± 0.07, while MMED, SOSVM and kernel SVM perform slightly worse with AUCs as 0.83 ± 0.11, 0.83 ± 0.11 and 0.84 ± 0.10. Regarding the earliness and accuracy of the detection, EDRA outperforms all the other classifiers for most extent of the trial. For the experiments on this dataset, the hyperparameter C for the SVM-based methods and the tuning parameter σ are selected based on 5-fold cross validation.

4.5. Longitudinal RNA-Seq data from IFNβ Drug Response Study

This section describes our experiment on the longitudinal RNA-Seq data from a drug therapy called Recombinant human interferon beta (rIFNβ), which is routinely used to control exacerbations in multiple sclerosis patients with only partial success, mainly because of adverse effects and a relatively large proportion of non-responders [3]. Therefore, early prediction and contemporaneous monitoring of the drug responses based on gene expression is important for doctors or researchers who would like to identify the suitable recipients of the specific drug therapy as well as to learn the long-term drug-induced effects.

The IFNβ drug response dataset is a longitudinal 70-gene expression dataset that contains the longitudinal gene expression data of 53 subjects. Patients with relapsing-remitting multiple sclerosis (MS) were followed for at least 2y after the initiation of therapy with IFNβ. Patients were classified as either good (33) or poor (20) responders at the end of therapy based on strict criteria [3]. Blood sample was obtained during each clinical follow-up every 3 months after the initialization of the therapy with IFNβ in the 1_{st} year, and every 6 months in the 2_{nd} year. In the previous research, there are 23 genes identified as predictive [3,18]. For the detailed information about the genes identified as being predictive, readers can refer to the supplementary document of the work [18]. The weights for the features for constructing the kernels for EDRA are therefore determined based on the prior knowledge about the predictive power of the features, i.e., the genes identified as not being predictive in literature are viewed as inactive features for the kernel construction. The AUC curves over the normalized delta time points prior to the recovery of EDRA and the other methods are depicted in Fig. 9b.

The experiments of IFNβ drug response dataset differ from the above experiments in the sense that there are pre-existing signatures that are able to separate good and poor responders before the initiation of the drug therapy, so that all the methods perform similarly well at the beginning. However, EDRA captures the long-term drug-induced progression via the increasing performance for classifying the good responders from the poor ones over time, while the other methods are not able to reflect the progression by the increased classification ability. This experiment demonstrates EDRA's contemporaneous risk evaluation performance. The hyperparameter C for the SVM-based methods and the tuning parameter σ for the kernel construction are selected based on 5-fold cross validation for this experiment.

5. Conclusions

This paper addresses problems of early detection and contemporaneous risk assessment for the diseases with irregular longitudinal data. We propose EDRA, a contemporaneous risk detector that is trained with the aim of identifying signatures of disease/drug-induced progression instead of individual data points. Our method is particularly suitable for the chronic disease with slow progression, which is hard to detect at the early stage. Experiments of varying situations from synthetic data to gene expression data of T1D study and gene expression data of drug response study are adopted to evaluate the performance of the proposed methods. Specifically, to evaluate the robustness of the proposed method on irregular longitudinal data, we consider irregularity and label insufficiency problems for synthesizing the data. The results obtained from the experiments demonstrate that EDRA enables early detection and contemporaneous risk assessment on

irregular and partially labeled longitudinal data. It is not only able to detect the onset of disease earlier with higher accuracy compared with the other methods, but also monitor the disease progression contemporaneously in difficult classification situation, such as in the early stage of the disease. What's more, the experiments also demonstrate the advantage of the methods that consider the temporal structure within data for capturing the disease/drug-induced progression over the other methods. Furthermore, we propose a flexible mixed-kernel framework, which incorporates the prior knowledge about features' discriminating power for the kernel construction. In future, we plan to improve our method to consider dynamic kernel selection by learning the correlation between the prediction outcomes with the individual sub-kernels, to simultaneously perform prediction as well as dynamic feature selection. Although there exist approaches enabling simultaneous kernel selection and prediction, such as MKL (Multiple Kernel Learning), HKL (Hierarchical Kernel Learning), Sparse Additive Models (SpAM) and HSIC (Hilbert-Schmidt independence criterion) Lasso, they are not specifically designed for longitudinal data analysis and thus they usually assume that the importance of the features to the classification is fixed for all the data points over time [39,2,52,66]. However, dynamic kernel/feature selection is needed in many applications. For instance, in the pathogenesis of chronic diseases, the features could play distinct roles between different stages [16]. In this paper, we illustrate the benefits of EDRA in the context of disease early detection and risk assessment, but our method can also be applied in many other domains where the longitudinal data analysis is involved.

Declaration of Competing Interest

The authors (Kai He, Shuai Huang, Xiaoning Qian) declare no competing interests.

Acknowledgment

This work was partially supported by the National Science Foundation (NSF)–Division of Communication and Computing Foundations (CCF) awards #1718513, #1715027, #1714136 and the JDRF award #2-SRA-2018-513-S-B.

References

[1] Duygu aliir, Esin Doantekin, An automatic diabetes diagnosis system based on LDA-wavelet support vector machine classifier, Expert Syst. Appl. 38 (2011) 8311–8315.
[2] Francis Bach, High-dimensional non-linear variable selection through hierarchical kernel learning, 2009. arXiv preprint arXiv: 0909.0844.
[3] Sergio E. Baranzini, Parvin Mousavi, Jordi Rio, Stacy J Caillier, Althea Stillman, Pablo Villoslada, Matthew M Wyatt, Manuel Comabella, Larry D. Greller, Roland Somogyi, et al., Transcription-based prediction of response to IFNβ using supervised computational methods, PLoS Biol. 3 (1) (2004) e2.
[4] Iyad Batal, Hamed Valizadegan, Gregory F. Cooper, Milos Hauskrecht, A pattern mining approach for classifying multivariate temporal data, Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on, IEEE, 2011, pp. 358–365.
[5] Mustafa Gokce Baydogan, George Runger, Time series representation and similarity based on local autopatterns, Data Min. Knowl. Disc. 30 (2) (2016) 476–509.
[6] Donald J. Berndt, James Clifford, Using dynamic time warping to find patterns in time series, KDD Workshop, Seattle, WA, vol. 10, 1994, pp. 359–370.
[7] Leo Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[8] Kin-Pong Chan, Wai-Chee Fu, Efficient time series matching by wavelets, ICDE, IEEE, 1999, p. 126.
[9] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, Yan Liu, Recurrent neural networks for multivariate time series with missing values, Sci. Rep. 8 (1) (2018) 6085.
[10] Lei Chen, Raymond Ng, On the marriage of lp-norms and edit distance, Proceedings of the Thirtieth international conference on Very Large Data Bases, vol. 30, VLDB Endowment, 2004, pp. 792–803.
[11] M. Lei Chen, Tamer Özsu, Vincent Oria, Robust and fast similarity search for moving object trajectories, Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, ACM, 2005, pp. 491–502.
[12] Min Chen, Hao Yixue, Kai Hwang, Lu Wang, Lin Wang, Disease prediction by machine learning over big data from healthcare communities, IEEE Access PP (2017) 1.
[13] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, Jimeng Sun, Doctor ai: predicting clinical events via recurrent neural networks, Machine Learning for Healthcare Conference, 2016, pp. 301–318.
[14] Line Clemmensen, Trevor Hastie, Daniela Witten, Bjarne Ersbøll, Sparse discriminant analysis, Technometrics 53 (4) (2011) 406–413.
[15] Wikipedia contributors, Chronic condition, Wikipedia, The Free Encyclopedia, 1 Apr. 2018, Web, 12 Apr. 2018.
[16] Bruno Dubois, Howard H. Feldman, Claudia Jacova, Harald Hampel, José Luis Molinuevo, Kaj Blennow, Steven T. DeKosky, Serge Gauthier, Dennis Selkoe, Randall Bateman, et al., Advancing research diagnostic criteria for alzheimer's disease: the iwg-2 criteria, Lancet Neurol. 13 (6) (2014) 614–629.
[17] Christos Faloutsos, Mudumbai Ranganathan, Yannis Manolopoulos, Fast Subsequence Matching in Time-series Databases vol. 23, ACM, 1994.
[18] vMohamed Ghalwash, Zoran Obradovic, Early classification of multivariate temporal observations by extraction of interpretable shapelets, BMC Bioinform. 13 (2012) 195.
[19] Fiona Imlach Gunasekara, Ken Richardson, Kristie Carter, Tony Blakely, Fixed effects analysis of repeated measures data, Int. J. Epidemiol. 43 (1) (2013) 264–269.
[20] Trevor Hastie, Robert Tibshirani, Andreas Buja, Flexible discriminant analysis by optimal scoring, J. Am. Stat. Assoc. 89 (428) (1994) 1255–1270.
[21] Minh Hoai, Fernando De la Torre, Max-margin early event detectors, Int. J. Comput. Vision 107 (2) (2014) 191–202.
[22] Yijun Huang, Qiang Meng, Heather Evans, Yu. William Lober, Xiaoning Qian Cheng, Ji Liu, Shuai Huang, Chi: A contemporaneous health index for degenerative disease monitoring using longitudinal measurements, J. Biomed. Informat. 73 (2017) 115–124.
[23] Christopher H. Jackson, Linda D. Sharples, Simon G. Thompson, Stephen W. Duffy, Elisabeth Couto, Multistate markov models for disease progression with classification error, J. Roy. Stat. Soc. Ser. D (Statist.) 52 (2) (2003) 193–209.
[24] Young-Seon Jeong, Myong K Jeong, Olufemi A. Omitaomu, Weighted dynamic time warping for time series classification, Pattern Recogn. 44 (9) (2011) 2231–2240.
[25] Keith A. Johnson, Nick C. Fox, Reisa A. Sperling, William E. Klunk, Brain imaging in alzheimer disease, Cold Spring Harbor Perspect. Med. (2012) a006213.
[26] Tony Jung, Kandauda Wickrama, An introduction to latent class growth analysis and growth mixture modeling, Soc. Pers. Psychol. Compass 2 (2008) 302–317.
[27] Anastasia Katsarou, Soffia Gudbjörnsdottir, Araz Rawshani, Dana Dabelea, Ezio Bonifacio, Barbara J. Anderson, Laura M. Jacobsen, Desmond A. Schatz, Åke Lernmark, Type 1 diabetes mellitus, Nat. Rev. Disease Primers 3 (2017) 17016.
[28] Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, Ioanna Chouvarda, Machine learning and data mining methods in diabetes research, Comput. Struct. Biotechnol. J. 15 (2017) 104–116.
[29] Chuyang Ke, Yan Jin, Heather Evans, Bill Lober, Xiaoning Qian, Ji Liu, Shuai Huang, Prognostics of surgical site infections using dynamic health data, J. Biomed. Informat. 65 (2017) 22–33.
[30] Eamonn Keogh, Chotirat Ann Ratanamahatana, Exact indexing of dynamic time warping, Knowl. Inform. Syst. 7 (3) (2005) 358–386.
[31] Eamonn Keogh, Kaushik Chakrabarti, Michael Pazzani, Sharad Mehrotra, Dimensionality reduction for fast similarity search in large time series databases, Knowl. Inform. Syst. 3 (3) (2001) 263–286.
[32] Eamonn Keogh, Kaushik Chakrabarti, Michael Pazzani, Sharad Mehrotra, Locally adaptive dimensionality reduction for indexing large time series databases, ACM Sigmod Record 30 (2) (2001) 151–162.
[33] Eamonn J. Keogh, Michael J. Pazzani, Derivative dynamic time warping, Proceedings of the 2001 SIAM International Conference on Data Mining, SIAM, 2001, pp. 1–11.
[34] Abed Khorasani, Mohammad Reza Daliri, HMM for classification of Parkinson's disease based on the raw gait data, J. Med. Syst. 38 (12) (2014) 147.
[35] Ralph L. Kodell, Bruce A. Pearce, Songjoon Baek, Hojin Moon, Hongshik Ahn, John F. Young, James J. Chen, A model-free ensemble method for class prediction with application to biomedical decision making, Artif. Intell. Med. 46 (3) (2009) 267–276.
[36] Mostafa Langarizadeh, Fateme Moghbeli, Applying naive bayesian networks to disease prediction: a systematic review, Acta Informat. Medica 24 (5) (2016) 364.
[37] Maksim Lapin, Matthias Hein, Bernt Schiele, Learning using privileged information: SVM+ and weighted SVM, Neural Netw. Off. J. Int. Neural Netw. Soc. 53C (2014) 95–108.
[38] Sandra Lee, Hui Huang, Marvin Zelen, Early detection of disease and scheduling of screening examinations, Statist. Methods Med. Res. 13 (6) (2004) 443–456.
[39] Fan Li, Yiming Yang, Eric P. Xing, From lasso regression to feature vector machine, Adv. Neural Inform. Process. Syst. (2006) 779–786.
[40] Yu-Ying Liu, Shuang Li, Fuxin Li, Le Song, James Rehg, Efficient learning of continuous-time hidden markov models for disease progression, Adv. Neural Inform. Process. Syst. 28 (2015) 3599–3607.
[41] Zitao Liu, Milos Hauskrecht, Learning linear dynamical systems from multivariate time series: a matrix factorization based framework, Proceedings of the 2016 SIAM International Conference on Data Mining, SIAM, 2016, pp. 810–818.
[42] Michael E. Matheny, Frederic S. Resnic, Nipun Arora, Lucila Ohno-Machado, Effects of SVM parameter optimization on discrimination and calibration for post-procedural pci mortality, J. Biomed. Inform. 40 (6) (2007) 688–697.
[43] Karl Øyvind Mikalsen, Filippo Maria Bianchi, Cristina Soguero-Ruiz, Robert Jenssen, Time series cluster kernel for learning similarities between multivariate time series with missing data, Pattern Recogn. 76 (2018) 569–581.
[44] Oh. Wonsuk, M. Era Kim, Regina Castro, Pedro J Caraballo, Vipin Kumar, Michael S. Steinbach, Gyorgy J. Simon, Type 2 diabetes mellitus trajectories and associated risks, Big Data 4 (1) (2016) 25–30.
[45] Frans J. Oort, Three-mode models for multivariate longitudinal data, Br. J. Math. Stat. Psychol. 54 (1) (2001) 49–78.

[46] Graziella Orr, William Pettersson-Yeo, Andre Marquand, Giuseppe Sartori, Andrea Mechelli, Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review, Neurosci. Biobehav. Rev. 36 (2012) 1140–1152.
[47] Akin Ozcift, Arif Gulten, Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms, Comput. Methods Programs Biomed. 104 (3) (2011) 443–451.
[48] Devi Parikh, Kristen Grauman, Relative attributes, IEEE International Conference on Computer Vision, 2011, pp. 503–510.
[49] David Pi, Monika Hudoba, Mike de Badyn, Rick White Nimmo, Jason Pal, Patrick Wong, Carmen Phoon, Deidre O'connor, Steven Pi, Kam Shojania, Application of linear discriminant analysis in performance evaluation of extractable nuclear antigen immunoassay systems in the screening and diagnosis of systemic autoimmune rheumatic diseases, Am. J. Clin. Pathol. 138 (4) (2012) 596–603.
[50] K. Wojtek Przytula, Don Thompson, Construction of bayesian networks for diagnostics, Aerospace Conference Proceedings, 2000 IEEE, vol. 5, IEEE, 2000, pp. 193–200.
[51] Nilam Ram, Kevin Grimm, Growth mixture modeling: a method for identifying differences in longitudinal change among unobserved groups, Int. J. Behav. Develop. 33 (2009) 565–576.
[52] Pradeep Ravikumar, John Lafferty, Han Liu, Larry Wasserman, Sparse additive models, J. Roy. Stat. Soc. Ser. B (Stat. Methodol.) 71 (5) (2009) 1009–1030.
[53] Beth A. Reboussin, David M. Reboussin, Kun-Yee Liang, James C. Anthony, Latent transition modeling of progression of health-risk behavior, Multivar. Behav. Res. 33 (4) (1998) 457–478.
[54] Jost Reinecke, Daniel Seddig, Growth mixture models in longitudinal research, AStA Adv. Stat. Anal. 95 (2011) 415–434.
[55] Mark D. Robinson, Davis J. McCarthy, Gordon K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics 26 (1) (2010) 139–140.
[56] Abid Sarwar, Vinod Sharma, Intelligent naïve bayes approach to diagnose diabetes type-2, Int. J. Comput. Appl. (2012) 14–16 (IJCA Special Edition Nov).
[57] Donald Stull, Ingela Wiklund, Rupert Gale, Gorana Capkun, Katherine Houghton, Paul Jones, Application of latent growth and growth mixture modeling to identify and characterize differential responders to treatment for copd, Contemp. Clin. Trials 32 (2011) 818–828.
[58] Rafid Sukkar, Elyse Katz, Yanwei Zhang, David Raunig, Bradley T Wyman, Disease progression modeling using hidden markov models, Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, IEEE, 2012, pp. 2845–2848.
[59] Tudor Toma, Robert Bosman, Arno Siebes, Niels Peek, Ameen Abu-Hanna, Learning predictive models that use pattern discovery–a bootstrap evaluative approach applied in organ functioning sequences, J. Biomed. Informat. 43 (2010) 578–586.
[60] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, Yasemin Altun, Large margin methods for structured and interdependent output variables, J. Mach. Learn. Res. 6 (Sep) (2005) 1453–1484.
[61] Vladimir Vapnik, Akshay Vashist, A new learning paradigm: learning using privileged information, Neural Netw. Off. J. Int. Neural Netw. Soc. 22 (2009) 544–557.
[62] Geert Verbeke, Steffen Fieuws, Geert Molenberghs, Marie Davidian, The analysis of multivariate longitudinal data: a review, Statist. Methods Med. Res. 23 (1) (2014) 42–59.
[63] Michail Vlachos, George Kollios, Dimitrios Gunopulos, Discovering similar multidimensional trajectories, Data Engineering, 2002. Proceedings. 18th International Conference on, IEEE, 2002, pp. 673–684.
[64] Xiang Wang, David Sontag, Fei Wang, Unsupervised learning of disease progression models, Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2014, pp. 85–94.
[65] Xiaoyue Wang, Abdullah Mueen, Hui Ding, Goce Trajcevski, Peter Scheuermann, Eamonn Keogh, Experimental comparison of representation methods and distance measures for time series data, Data Min. Knowl. Disc. 26 (2) (2013) 275–309.
[66] Makoto Yamada, Wittawat Jitkrittum, Leonid Sigal, Eric P Xing, Masashi Sugiyama, High-dimensional feature selection by feature-wise kernelized lasso, Neural Comput. 26 (1) (2014) 185–207.
[67] Lexiang Ye, Eamonn Keogh, Time series shapelets: a new primitive for data mining, Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2009, pp. 947–956.
[68] Daoqiang Zhang, Yaping Wang, Luping Zhou, Hong Yuan, Dinggang Shen, Multimodal classification of Alzheimer's disease and mild cognitive impairment, NeuroImage 55 (3) (2011) 856–867.
[69] Jinghe Zhang, Haoyi Xiong, Yu Huang, Hao Wu, Kevin Leach, Laura E. Barnes, M-SEQ: Early detection of anxiety and depression via temporal orders of diagnoses in electronic health data, in: 2015 IEEE International Conference on Big Data (Big Data), 2015, pp. 2569–2577.
[70] Jiayu Zhou, Jun Liu, Vaibhav A. Narayan, Jieping Ye, Modeling disease progression via multi-task learning, NeuroImage 78 (2013) 233–248.