2020 IEEE International Conference on Data Mining (ICDM)

Order-Preserving Metric Learning for Mining Multivariate Time Series

Jie Xu*, Zhenxing Xu*, Bin Yu† and Fei Wang*
*Department of Population Health Sciences
Weill Cornell Medicine, New York, USA
Email: {jix4002, zhx2005, few2001}@med.cornell.edu
†American Air Liquide, Newark, USA
Email: bin.yu@airliquide.com

***Abstract*—Multivariate time series (MTS) analysis is an increasingly popular research topic in recent years due to the vast amount of MTS data that are being generated in numerous fields such as genomics research, health informatics, finance and abnormal detection. The particularity of the data makes it a challenging task, *e.g.*, missing data, different sampling frequencies, and random noise. Moreover, each instance depends not only on its past values but also has some dependency on other instances, and there exist discriminatory order-dependent characteristics. To address these challenges, in this paper, we introduce an order-preserving metric learning framework for multivariate time series prediction. Specifically, we adopt quadruplet-wise constraints which can encompass pair-wise and triplet-wise constraints to model similarity from complex label relations. To preserve the inherent temporal relationships of the instances in MTS, order-preserving Wasserstein distance is integrated to the framework to measure dissimilarity between MTS data, where the inverse difference moment regularization enforces flow-network with local homogeneous structures and the KL-divergence with a prior distribution regularization prevents flow-network between instances with faraway temporal locations. Besides the regularizations on flow-network, the ground measurement of the Wasserstein distance is replaced by Mahalanobis distance to increase its discrimination capability. An alternating iteration strategy is proposed to jointly optimize the Mahalanobis distance matrix in the ground measurement and the flow-network of Wasserstein distance. Extensive experiments on real-world clinical data from critical care are provided to demonstrate the effectiveness of the proposed method on sepsis prediction task.**

***Keywords*-metric learning; Wasserstein distance; order-preserving; sepsis;**

I. Introduction

Nowadays temporal data are being generated from almost every application domain at an unprecedented rate, *e.g.*, daily fluctuations of stock market, network monitoring, sensor readings, patient data captured from medical device, *etc*. The analysis of such time series data has increasingly become to a critical task in applications including stock trend analysis, mechanical structure reliability assessment, air quality prediction, clinical risk prediction and other practical projects, *etc* [1]–[4].

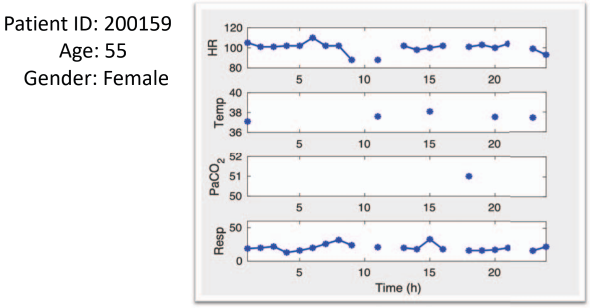


Figure 1: An example of a patient's ICU data.

Time series are discrete or continuous sequences of real-valued elements collected over a period of time. The data capturing entity (*e.g.*, a mobile device or a computer) is continuously sampled at fixed or varying temporal resolutions. Usually these time series data are multivariate, *i.e.*, the sampled readings are originated from multiple sensors, which are not necessarily synchronized in any fashion [5]–[7]. Figure 1 is an example of a patients' ICU data. Multiple variables (*e.g.*, heart rate (HR), temperature (Temp)) were recorded at different time points. The phenomena of missing data, different sampling frequencies and random noise exist. Predictive modeling with such asynchronous MTS data is a challenging task.

Existing time series classification methods can be roughly divided into three large categories, *i.e.*, feature based, model based and sequence distance based methods [8]–[10]. Feature based methods usually transform the sequence into a feature vector through feature engineering and then apply conventional classification methods such as a decision tree or a neural network [11], [12]. Model based methods assume sequences in a class are generated by an underlying model such as naive Bayes sequence classifier [13] and hidden Markov model [14]. Sequence distance based methods define a distance function to measure the similarity between a pair of sequences, where the nearest neighbor (NN) algorithm is typically coupled with such a distance function. In this paper, we focus on the sequence distance based methods.

Among all distance measures, Euclidean distance is a widely adopted option because of its simplicity and good interpretability. However, it requires two sequences to have the same length and is sensitive to the misalignment on temporal locations. To address this problem, Berndt *et*

2374-8486/20/$31.00 ©2020 IEEE
DOI 10.1109/ICDM50108.2020.00080

al. [15] proposed dynamic time wrapping (DTW), which allowed one time series to be "stretched" to provide a better match with another time series [16]. However, the alignments determined by DTW remain strictly sequential, *i.e.*, instances (sampled points) in one sequence are not allowed to be aligned with instances in another sequence unless all points before them in both sequences have already been aligned [17]. By viewing the instances in a sequence as variable-size descriptions of distributions, Wasserstein distance provides another flexible way to measure the dissimilarity between two sequences [18]. It is a powerful metric based on the theory of optimal transport and defined between probability distributions on a ground distance matrix. Although Wasserstein distance can solve the problem of local rank inversion and different starting points, it completely ignores the temporal dependencies of instances. Therefore, Su *et al.* [19] proposed the order-preserving Wasserstein distance which incorporates the advantages of flexibility of optimal transport and order preserving alignments. Two regularization terms including the inverse difference moment regularization and KL-divergence with a prior distribution regularization were imposed on the flow-network to preserve the inherent temporal relationships.

Besides the flow-network, the ground distance matrix between two sequences in the Wasserstein definition also plays an important role in the flow optimization procedure. A well-designed ground distance matrix can generate better flow-network and thus lead to better Wasserstein distance measurements. For this reason, researchers have proposed to adopt the metric learning framework [20] to further improve the flexibility of Wasserstein distance. Xu *et al.* [21] proposed a multi-level metric learning method using a smoothed Wasserstein distance to characterize the errors between any sample pairs, where the ground distance is considered as a Mahalanobis distance. Su *et al.* [22] unified a wide range of distance measures for sequences into a unified framework as a function of the ground metric for elements in sequences. The final distances are meta-distances built upon the ground metric by inferring the temporal alignments among the element pairs.

In this paper, we propose an order-preserving metric learning framework for multivariate time series prediction. We model similarity from label relations with quadruplet-wise constraint since it can encompass pair-wise and triplet-wise constraint. Order-preserving Wasserstein distance is then integrated into the framework to preserve the inherent temporal relationships of the instances in sequences. Two temporal regularizations are imposed to penalize the flow-network between instances with distant temporal positions. Besides the regularizations on flow-network, the ground measurement of the Wasserstein distance is set to Mahalanobis distance to increase its discrimination capability. An alternating iteration strategy is proposed to jointly optimize the Mahalanobis distance matrix in the ground distance and Wasserstein distance flow-network. Results on predicting the risk of sepsis in critical care demonstrate the effectiveness of the proposed method.

The rest of this paper is organized as follows. Section II detailedly introduces the most related work including smoothed Wasserstein distance and order-preserving Wasserstein distance. In Section III, we first formulate the general metric learning framework and then propose our order-preserving metric learning model. Then, we introduce an alternating iteration strategy to optimize the proposed model in section IV. Experimental results are presented and discussed in section V. Section VI concludes the paper.

Notations Let boldface lowercase letters like $\mathbf{z} \in \mathbb{R}^d$ be vectors with dimension d and boldface uppercase letters like $\mathbf{Z} \in \mathbb{R}^{d\times c}$ be matrices with size $d \times c$. The transpose of $\mathbf{Z}$ is denoted as $\mathbf{Z}^\top$ and the real numbers are denoted as uppercase letters like $Z \in \mathbb{R}$.

II. Related Work

We start this section by first describing the smoothed Wasserstein distance, and then revisiting the order-preserving Wasserstein distance [19].

A. Smoothed Wasserstein Distance

Wasserstein distance is essentially the optimal solution for transportation problems in linear programming. It can be thought of as a minimum amount of work required to move the entire earth from source to destination [23]. Given two sequences $P = \{(\mathbf{p}_1, w_{p_1}), (\mathbf{p}_2, w_{p_2}), \cdots, (\mathbf{p}_m, w_{p_m})\}$ and $Q = \{(\mathbf{q}_1, w_{q_1}), (\mathbf{q}_2, w_{q_2}), \cdots, (\mathbf{q}_n, w_{q_n}),$ the Wasserstein distance between them is defined as:

$$W(P,Q) = \min_{\mathbf{F}\in\mathbb{F}(P,Q)} Tr(\mathbf{D}^\top \mathbf{F}) \tag{1}$$

where $\mathbf{D} = \{d(i,j)\}, i = 1, \cdots, m, j = 1, \cdots, n$ is the ground distance matrix, and $d(i,j)$ defines the cost of moving one unit of earth from the source $\mathbf{p}_i$ to the target $\mathbf{q}_j$. $\mathbf{F} = \{f(i,j)\}, i = 1, \cdots, m, j = 1, \cdots, n$ is a flow-network matrix, and $f(i,j)$ denotes the amount of earth moved from the source $\mathbf{p}_i$ to the target $\mathbf{q}_j$.

Let $\mathbf{w_p} = [w_{p_1}, w_{p_2}, \cdots, w_{p_m}] \in \mathbb{R}^m$, $\mathbf{w_q} = [w_{q_1}, w_{q_2}, \cdots, w_{q_n}] \in \mathbb{R}^n$, then $\mathbb{F}(P,Q)$ can be written as:

$$\mathbb{F}(P,Q) = \{\mathbf{F} \in \mathbb{R}_+^{m\times n} | \mathbf{F}^\top \mathbf{1}_m = \mathbf{w_q}, \mathbf{F}\mathbf{1}_n = \mathbf{w_p}\}. \tag{2}$$

Optimizing the Wasserstein distance problem is actually solving several expensive optimal transportation problems. In addition, due to the minimum value of affine function, Wasserstein itself is not a smooth function of its parameters, which limits the application of Wasserstein distance. In order to overcome the above problems, some researchers have proposed the smoothed optimal transport problem under entropy constraint [24]:

$$W_\gamma(P,Q) = \min_{\mathbf{F}\in\mathbb{F}(P,Q)} Tr(\mathbf{D}^T \mathbf{F}) - \gamma h(\mathbf{F}), \tag{3}$$

where h is the (strictly concave) entropy function

$$h(\mathbf{F}) = -\langle \mathbf{F}, \mathrm{log}\mathbf{F} \rangle, \tag{4}$$

and $\gamma > 0$ is a balance parameter. In this paper, we call Eq. (3) as smoothed Wasserstein distance [21]. The flow-network $\mathbf{F}^*$ solution of the problem in Eq. (3) is unique and can be found by computing two vectors $\boldsymbol{\kappa}_1 \in \mathbb{R}_+^m$, $\boldsymbol{\kappa}_2 \in \mathbb{R}_+^n$ such that $\mathrm{diag}(\boldsymbol{\kappa}_1)\mathbf{e}^{-\mathbf{D}/\gamma}\mathrm{diag}(\boldsymbol{\kappa}_2) \in \mathbb{F}(P, Q)$. The optimal solution is then

$$\mathbf{F}^* = \mathrm{diag}(\boldsymbol{\kappa}_1)\mathbf{e}^{-\mathbf{D}/\gamma}\mathrm{diag}(\boldsymbol{\kappa}_2). \tag{5}$$

Typically, $\boldsymbol{\kappa}_1$ and $\boldsymbol{\kappa}_2$ can be solved using Sinkhorn's algorithm [25]. This algorithm has linear convergence, and requires $\mathcal{O}(mn)$ operations at each iteration. However, if two sequences differ only in the order of elements, then the Wasserstein distance is incapable of distinguishing them.

B. Order-Preserving Wasserstein Distance

To preserve the inherent temporal relationships of the instances in sequences, Su *et al.* [19] proposed the order-preserving Wasserstein distance. Specifically, two novel temporal regularizations are imposed to punish the flow-network between instances with distant temporal positions, so that the learned optimal flow can maintain the temporal dependencies of instances in sequences.

To encourage temporally approached elements to match, the first regularization favors $\mathbf{F}$ with large inverse difference moment which is calculated as

$$I(\mathbf{F}) = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{f_{ij}}{\left(\frac{i}{m} - \frac{j}{n}\right)^2 + 1}. \tag{6}$$

If the large values of $\mathbf{F}$ are distributed mainly along the diagonal, then the value of $I(\mathbf{F})$ will be large.

To encourage the flow-network to be smooth and reasonable, the second regularization favors the distribution of $\mathbf{F}$ to be similar to a prior distribution $\mathbf{P}$:

$$p_{ij} = \mathbf{P}(i, j) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{l^2(i,j)}{2\sigma^2}}, \tag{7}$$

where $l(i, j) = \frac{|i/m - j/n|}{1/m^2 + 1/n^2}$ is the distance from the position (i, j) to the diagonal line.

By imposing two regularizations to the feasible set $\mathbb{F}(P, Q)$, we have

$$\begin{aligned}\mathbb{F}_{\lambda_1,\lambda_2}(P, Q) = \{\mathbf{F} \in \mathbb{R}_+^{m\times n} | \mathbf{F}^\top \mathbf{1}_m = \mathbf{w_q}, \mathbf{F}\mathbf{1}_n = \mathbf{w_p},\\ I(\mathbf{F}) \geq \lambda_1, KL(\mathbf{F}\|\mathbf{P}) \leq \lambda_2\},\end{aligned} \tag{8}$$

where $KL(\mathbf{F}\|\mathbf{P}) = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} \log \frac{f_{ij}}{p_{ij}}$ is the Kullback-Leibler (KL) divergence between two matrices. Then, order-preserving Wasserstein (OPW) distance between two sequences P and Q can be defined as:

$$W_{\lambda_1,\lambda_2}(P, Q) = \min_{\mathbf{F}\in\mathbb{F}_{\lambda_1,\lambda_2}(P,Q)} Tr(\mathbf{D}^T\mathbf{F}). \tag{9}$$

Similar to the smoothed Wasserstein distance, the additional constraints greatly reduce the computational complexity of calculating the optimal flow-network [19].

The ground distance in Eq. (9) is usually Euclidean, cosine or sparse L$_1$-norm distances. However, these distances cannot allow arbitrary linear scaling and rotation of the feature space, nor can they take advantage of the discriminative information that exists in the data space. Therefore, in this paper, we use the Mahalanobis distance as the ground measurement to improve the discrimination capability of W_{λ_1,λ_2}.

III. Order-Preserving Metric Learning Model

In this section, we first describe the general metric learning formulation and then propose our order-preserving metric learning model for multivariate time series analysis.

A. General Metric Learning Formulation

Metric learning targets at learning an adaptive distance, such as widely used Mahalanobis distance $d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M}(\mathbf{x}_i - \mathbf{x}_j)}$, to effectively reflect the similarity between data. Generally, it can be formulated as an optimization problem of the form [26]–[28],

$$\min_{\mathbf{M}} \quad \mathrm{Loss}(\mathbf{M}, \mathcal{A}) + \lambda \mathrm{Reg}(\mathbf{M}), \tag{10}$$

where $\mathrm{Reg}(\mathbf{M})$ is a regularization term on the metric matrix $\mathbf{M}$, and $\lambda > 0$ is the regularization parameter. $\mathrm{Loss}(\mathbf{M}, \mathcal{A})$ is a loss function that penalizes constraints that are not satisfied. It usually measures the ability of the matrix $\mathbf{M}$ to satisfy some distance constraints given in the training set. Commonly used constraints include pairwise constraint [29] and triplet constraint [30].

The pairwise constraint contains information whether two objects in a pair are similar or dissimilar, sometimes positive pairs or negative pairs. It can be represented by $\mathcal{D}$ and $\mathcal{S}$ as

$$\begin{aligned}\mathcal{S} &= \{(\mathbf{x}_i, \mathbf{x}_j), \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are similar}\},\\ \mathcal{D} &= \{(\mathbf{x}_i, \mathbf{x}_j), \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are dissimilar}\}.\end{aligned} \tag{11}$$

The information-Theoretic Metric Learning (ITML) is one of many methods using pairwise constraint training examples in metric learning field [29], which is formulated as follows

$$\begin{aligned}\min_{\mathbf{M}\in\mathbb{S}_+^d} \quad & \gamma \sum_{i,j} \xi_{ij} + D_{ld}(\mathbf{M}, \mathbf{M}_0)\\ s.t. \quad & d_{\mathbf{M}}^2(\mathbf{x}_i, \mathbf{x}_j) \leq u + \xi_{ij}, \forall(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S},\\ & d_{\mathbf{M}}^2(\mathbf{x}_i, \mathbf{x}_j) \geq l - \xi_{ij}, \forall(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D},\end{aligned} \tag{12}$$

where u and l are upper bound and lower bound for similar samples and dissimilar samples respectively. ξ_{ij} is a safety margin distance for each pair. $D_{ld}(\mathbf{M}, \mathbf{M}_0)$ is LogDet divergence.

Triplet constraint is also widely used in metric learning, and it is denoted by $\mathcal{R}$ as

$$\mathcal{R} = \{(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) : \mathbf{x}_i \text{ is more similar to } \mathbf{x}_j \text{ than to } \mathbf{x}_k\}. \tag{13}$$

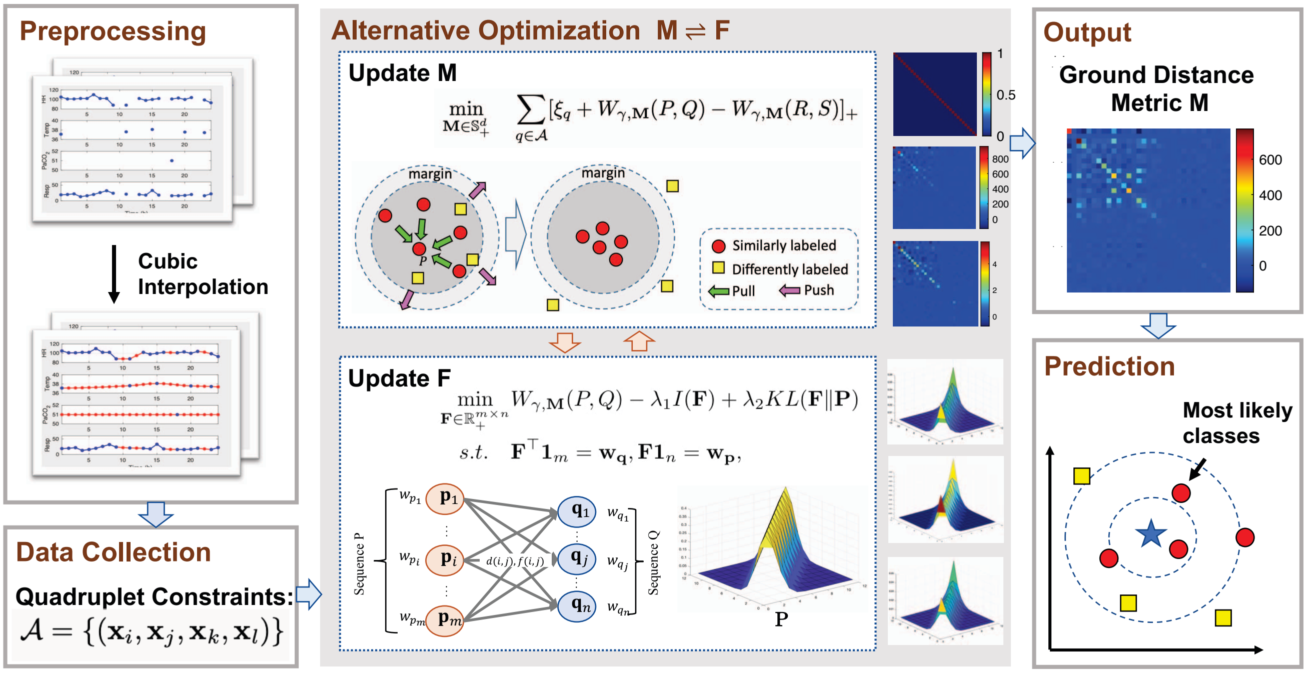

Preprocessing
Cubic Interpolation
Data Collection
Quadruplet Constraints:
$\mathcal{A} = \{(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l)\}$
Alternative Optimization M ⇌ F
Update M
$\min_{\mathbf{M}\in\mathbb{S}_+^d} \sum_{q\in\mathcal{A}} [\xi_q + W_{\gamma,\mathbf{M}}(P,Q) - W_{\gamma,\mathbf{M}}(R,S)]_+$
margin
margin
p
Similarly labeled
Differently labeled
Pull
Push
1
0.5
0
800
600
400
200
0
4
2
0
Update F
$\min_{\mathbf{F}\in\mathbb{R}_+^{m\times n}} W_{\gamma,\mathbf{M}}(P,Q) - \lambda_1 I(\mathbf{F}) + \lambda_2 KL(\mathbf{F}\|\mathbf{P})$
$s.t.\ \mathbf{F}^\top \mathbf{1}_m = \mathbf{w_q}, \mathbf{F}\mathbf{1}_n = \mathbf{w_p},$
Sequence P
Sequence Q
w_{p_1} $\mathbf{p}_1$
w_{p_i} $\mathbf{p}_i$
w_{p_m} $\mathbf{p}_m$
$\mathbf{q}_1$ w_{q_1}
$\mathbf{q}_j$ w_{q_j}
$\mathbf{q}_n$ w_{q_n}
$d(i,j), f(i,j)$
P
Output
Ground Distance Metric M
600
400
200
0
Prediction
Most likely classes

Figure 2: Overview of multivariate time series prediction with order-preserving metric learning model.

Large Margin Nearest Neighbor (LMNN) [30] is one of the most widely used metric learning methods which uses triplet constraint on training examples. LMNN is to solve

$$
\begin{aligned}
\min_{\mathbf{M}\in\mathbb{S}_+^d} \quad & (1-\mu)\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}} d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j) + \mu \sum_{(i,j,k)\in\mathcal{R}} \xi_{ijk} \\
s.t. \quad & d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_k) - d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j) \geq 1 - \xi_{ijk}, \\
& \forall(\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_i) \in \mathcal{R},
\end{aligned}
\tag{14}
$$

where $\mu \in [0,1]$ controls relative weight between two terms. Suppose y_i is the label of $\mathbf{x}_i$, then $\mathcal{S} = \{(\mathbf{x}_i,\mathbf{x}_j) : y_i = y_j$ and $\mathbf{x}_j$ belongs to the k-neighborhood of $\mathbf{x}_i\}$, $\mathcal{R} = \{(\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_k) : (\mathbf{x}_i,\mathbf{x}_j) \in \mathcal{S}, y_i \neq y_k\}$. It is proved to be very effective to learn Mahalanobis distance in practice, which is extended to many methods for different applications.

B. Order-preserving Metric Learning Model

In this paper, we focus on quadruplet constraint since it can encompass pairwise and triplet constraints. Quadruplet constraint [31] was proposed to model similarity from complex semantic label relations, for example, the degree of presence of smile, from least smiling to most smiling. The scheme of quadruplet constraint is

$$
\mathcal{A} = \{(\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_k,\mathbf{x}_l) : d^2_{\mathbf{M}}(\mathbf{x}_k,\mathbf{x}_l) \geq d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j) + \xi_q\},
\tag{15}
$$

where ξ_q is a soft margin. Thus, quadruplet constraint based metric learning model can be easily written as

$$
\min_{\mathbf{M}\in\mathbb{S}_+^d} \sum_{q\in\mathcal{A}} [\xi_q + d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j) - d^2_{\mathbf{M}}(\mathbf{x}_k,\mathbf{x}_l)]_+ \tag{16}
$$

where $[\cdot] = \max(0,\cdot)$. Note that pairwise constraint can be represented as $(\mathbf{x}_i,\mathbf{x}_i,\mathbf{x}_i,\mathbf{x}_j)$, and set $\xi_q = l$, so $d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j) \geq l$ and $\mathbf{x}_i,\mathbf{x}_j$ are from dissimilar set; or $(\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_i,\mathbf{x}_i)$, and set $\xi_q = -u$, then $d^2_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j) \leq u$, $\mathbf{x}_i,\mathbf{x}_j$ are from similar set. Similarly, triple constraint can also be represented as $(\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_i,\mathbf{x}_k)$.

This paper is to deal with multivariate time series data. Considering the inherent temporal relationships of the instances in MTS data, we replace the Mahalanobis distance with order-preserving Wasserstein distance, and use Mahalanobis distance as the ground measurement of OPW distance. That is, the squared Mahalanobis distance between the i-th bin of P and the j-th bin of Q can be expressed as

$$
d_{\mathbf{M}}(i,j) = (\mathbf{p}_i - \mathbf{q}_j)^\top \mathbf{M}(\mathbf{p}_i - \mathbf{q}_j), \tag{17}
$$

where $\mathbf{M}$ is a global linear transformation of the underlying space, and $\mathbf{D_M} = \{d_{\mathbf{M}}(i,j)\}$ is the ground distance.

On the basis of Eq. (17), we can construct a Mahalanobis OPW distance ($W_{\lambda,\mathbf{M}}$ for short). Let the OPW distance between signatures P and Q be $W_{\lambda,\mathbf{M}}(P,Q)$ and the quadruplet be $(P,Q,R,S) \in \mathcal{A}$. Then, we replace

$d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j)$ with $W_{\lambda,\mathbf{M}}(P,Q)$ in Eq. (16) and obtain:

$$\min_{\mathbf{M}\in\mathbb{S}^d_+} \sum_{q\in\mathcal{A}} [\xi_q + W_{\gamma,\mathbf{M}}(P,Q) - W_{\gamma,\mathbf{M}}(R,S)]_+ - \sum_{q\in\mathcal{A}} \mathcal{L}_{order}, \tag{18}$$

where $\mathcal{L}_{order} = \lambda_1 (I(\mathbf{F}_{PQ}) + I(\mathbf{F}_{RS})) - \lambda_2(KL(\mathbf{F}_{PQ}\|\mathbf{P}) + KL(\mathbf{F}_{RS}\|\mathbf{P}))$ and $\mathbf{F}_{PQ} \in \mathbb{F}(P,Q)$, $\mathbf{F}_{RS} \in \mathbb{F}(R,S)$.

Mode (18) inherits the advantages of OPW distance, *i.e.*, imposes two regularization terms to the flow-network to preserve the global temporal information. Mahalanobis distance is as the ground measurement of OPW to improve the discriminative capability. Once such a distance function is obtained, a simple k-nearest neighbor classifier (KNN) is used to finish the prediction task.

Figure 2 illustrates the overview of multivariate time series prediction with our order-preserving metric learning model. Specifically, first, we impute missing values in multivariate time series data (*e.g.*, cubic interpolation). Second, quadruplet constraints are constructed from complex label relations. Third, An alternating iteration strategy is used to jointly optimize the Mahalanobis distance matrix and the flow-network of Wasserstein distance. After we obtain the optimal ground distance metric, 1NN is used to do classification.

IV. Optimization Algorithm

There are two groups of variables that need to be learned in model (18), *i.e.*, flow-network $\mathbf{F}$ and metric matrix $\mathbf{M}$. When flow-network $\mathbf{F}$ is fixed, model (18) turns into the Mahalanobis-like metric learning problem, *i.e.*,

$$\min_{\mathbf{M}\in\mathbb{S}^d_+} \sum_{q\in\mathcal{A}} [\xi_q + W_{\gamma,\mathbf{M}}(P,Q) - W_{\gamma,\mathbf{M}}(R,S)]_+, \tag{19}$$

The subgradient of problem (19) with respect to $\mathbf{M}$ is:

$$\begin{aligned} \nabla_{\mathbf{M}} &= \sum_{q\in\mathcal{A}^+} (G_{P,Q} - G_{R,S}), \\ G_{P,Q} &= P\text{diag}(\mathbf{F}_{PQ}\mathbf{e})P^\top + Q\text{diag}(\mathbf{e}^\top\mathbf{F}_{PQ})Q^\top \\ &\quad - (P\mathbf{F}_{PQ}Q^\top + Q\mathbf{F}^\top_{PQ}P^\top), \\ G_{R,S} &= R\text{diag}(\mathbf{F}_{RS}\mathbf{e})R^\top + S\text{diag}(\mathbf{e}^\top\mathbf{F}_{RS})S^\top \\ &\quad - (R\mathbf{F}_{RS}S^\top + S\mathbf{F}^\top_{RS}R^\top), \end{aligned} \tag{20}$$

where $\mathcal{A}^+$ denotes the subset of constraints in $\mathcal{A}$ that is larger than 0 in function (15). Then, we can update $\mathbf{M}$ using subgradient method, *i.e.*,

$$\mathbf{M} \Leftarrow \mathcal{P}_{\mathbb{S}_+}(\mathbf{M} - \eta\nabla_{\mathbf{M}}), \tag{21}$$

where $\mathcal{P}_{\mathbb{S}_+}(\cdot)$ denotes the projection operator and $\eta > 0$ is a step size.

When fixing $\mathbf{M}$ in Eq. (18), problem can be splitted into several order-preserving Wasserstein distance subproblems [19]. We take the pair (P,Q) in the quadruplet $(P,Q,R,S) \in \mathcal{A}$ as an example, then we have

$$\begin{aligned} &\min_{\mathbf{F}\in\mathbb{R}^{m\times n}_+} W_{\gamma,\mathbf{M}}(P,Q) - \lambda_1 I(\mathbf{F}) + \lambda_2 KL(\mathbf{F}\|\mathbf{P}) \\ &\quad s.t. \quad \mathbf{F}^\top\mathbf{1}_m = \mathbf{w_q}, \mathbf{F}\mathbf{1}_n = \mathbf{w_p}, \end{aligned} \tag{22}$$

To obtain the optimal $\mathbf{F}$, we start from the Lagrangian function of Eq. (22)

$$\begin{aligned} L(\mathbf{F},\boldsymbol{\alpha},\boldsymbol{\beta}) =& W_{\gamma,\mathbf{M}}(P,Q) - \lambda_2 \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{f_{ij}}{\left(\frac{i}{m} - \frac{j}{n}\right)^2 + 1} \\ &+ \lambda_3 KL(\mathbf{F}\|\mathbf{P}) + \boldsymbol{\alpha}^\top(\mathbf{F}^\top\mathbf{1}_m - \mathbf{w_q}) + \boldsymbol{\beta}^\top(\mathbf{F}\mathbf{1}_n - \mathbf{w_p}), \end{aligned} \tag{23}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the dual variables for the two equality constraints. The derivative of $L(\mathbf{F},\boldsymbol{\alpha},\boldsymbol{\beta})$ with respect to f_{ij} is:

$$\begin{aligned} \frac{\partial L(\mathbf{F},\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial f_{ij}} =& d_{\mathbf{M}}(i,j) - \lambda_2/((i/m - j/n)^2 + 1) \\ &+ \lambda_3 (\log(f_{ij}/p_{ij}) + 1) + \alpha_i + \beta_j. \end{aligned} \tag{24}$$

Setting $\frac{\partial L(\mathbf{F},\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial f_{ij}}$ to zero, we get:

$$f_{ij} = p_{ij} e^{-\frac{1}{\lambda_3}(d_{\mathbf{M}}(i,j) - \lambda_2/((i/m - j/n)^2 + 1) + \alpha_i + \beta_j) - 1} \tag{25}$$

We denote $k_{ij} = p_{ij} e^{-\frac{1}{\lambda_3}(d_{\mathbf{M}}(i,j) - \lambda_2/((i/m - j/n)^2 + 1))}$, then $f_{ij} = e^{-1/2 - \alpha_i/\lambda_3} k_{ij} e^{-1/2 - \beta_i/\lambda_3}$. Thus, we have

$$\mathbf{F} = \mathbf{e}^{\text{diag}(-\frac{1}{2} - \frac{\boldsymbol{\alpha}}{\lambda_3})} \mathbf{K} \mathbf{e}^{\text{diag}(-\frac{1}{2} - \frac{\boldsymbol{\beta}}{\lambda_3})}. \tag{26}$$

Lemma 1: According to [25], for any $N \times M$ matrix $\mathbf{A}$ with all positive elements, there exist diagonal matrices $\mathbf{B}_1$ and $\mathbf{B}_2$ such that $\mathbf{B}_1\mathbf{A}\mathbf{B}_2$ belongs to $\mathbb{F}(P,Q)$. $\mathbf{B}_1$ and $\mathbf{B}_2$ have strictly positive diagonal values, and are unique up to a positive scalar factor.

Via Lemma 1 and follow [19], we know the optimal $\mathbf{F}$ in Eq. (26) has the same form with Eq. (5), and hence is exactly the unique matrix in $\mathbb{F}(P,Q)$ which is a rescaled version of $\mathbf{K}$. The flow-network of each pair (*i.e.*, (P,Q) or (R,S)) in the quadruplet $(P,Q,R,S) \in \mathcal{A}$ can be similarly solved using the above process. The specific procedures are summarized in Algorithm 1.

Algorithm 1 Algorithm to solve Eq. (18)

```
Input: $\mathcal{A}$, $\mathbf{X} \in \mathbb{R}^{d\times n}$
Output: $\mathbf{M} \in \mathbb{S}^d_+$
Initialization: $\mathbf{M}$, $\mathbf{L}$
repeat
    Calculate each $\mathbf{F}_{PQ}$ and $\mathbf{F}_{RS}$ in the quadruplet $(P,Q,R,S) \in \mathcal{A}$ using Eq. (25) $\sim$ Eq. (26);
    Calculate $G_{P,Q}$ and $G_{R,S}$ using Eq. (20);
    Calculate $\nabla_{\mathbf{M}} = \sum_{q\in\mathcal{A}^+}(G_{P,Q} - G_{R,S})$;
    Update $\mathbf{M} \Leftarrow \mathcal{P}_{\mathbb{S}_+}(\mathbf{M} - \eta\nabla_{\mathbf{M}})$;
until Converge
```

V. Experiments

In this section, we evaluate the proposed model on a real-world clinical data set from critical care.

A. Datasets

We evaluate the proposed model over electronic health record (EHR) data acquired from Medical Information Mart for Intensive Care III (MIMIC-III) [32]. MIMIC-III is a freely and publicly-available database which encompasses a diverse and very large population of Intensive Care Unit (ICU) patients. Of 15309 ICU hospitalized patients aged 18 years and older, 1221 patients who met the criteria of sepsis were case patients [33]. The remaining stays were controls. All patients diagnosed with sepsis at admission or within 7 hours of admission were excluded from the analysis. In addition, we also excluded some patients with insufficient records. Table I shows the final dataset summary statistics.

Table I: Dataset summary statistics

Observation Window	# Case	# Control
6 hours	586	6408
12 hours	586	6408
24 hours	509	5899

B. Experimental Setup

1) Setting: Our experiments are designed to identify sepsis patients at the time of onset and 3, 6, 12 and 24 hours prior to onset. The prediction window setting is shown in Fig. 3. The patients without sepsis (*i.e.*, controls) were randomly assigned "onset time" according to a continuous and uniform probability distribution as a negative sample. The observation window is the sequence of values that is used to predict if there will occur a sepsis onset or not. In this paper, we examined three observation windows, *i.e.*, 6, 12 and 24 hours.

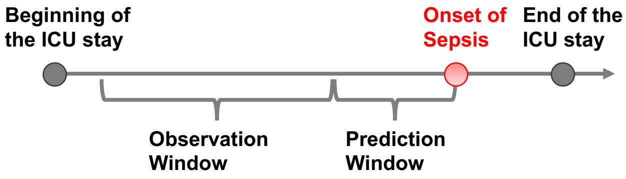


Figure 3: Prediction window setting.

2) Preprocessing: We extracted variables hourly [33] and kept those variables that more than 95% patients have, including age, gender, heart rate, potassium, temperature, pH, PaCO2 (partial pressure of carbon dioxide from arterial blood), systolic blood pressure, blood urea nitrogen, mean arterial pressure, chloride, creatinine, glucose, diastolic blood pressure, respiration rate, white blood cells count, platelet count [34], [35]. These values, as well as the hourly differences in each value (except for age and gender) were concatenated into a feature vector. If a patient does not have at least one new measurement per hour prior to the onset time, the missing value is determined by cubic interpolation with the patient's existing recorded value.

3) Compared Methods: We compared our method with most commonly used distances for time series data and most related work, including DTW [15], OPW [19], RVSML-DTW [22], RVSML-OPW [22]. DTW and OPW are two baseline methods. RVSML-DTW and RVSML-OPW are metric learning methods, where RVSML-OPW is based on Wasserstein distance. Since we want to evaluate the ability of using a specific distance of time series in the classification process rather than the classifier itself, all methods are coupled with 1NN to do the final prediction task.

4) Evaluation Criteria: We use mean accuracy with standard deviation as our main evaluation criterion, and the method with the highest mean accuracy is the most accurate one. The dataset is splitted into 5 folds based on sample proportion, where 4 folds are used for training and 1 fold for testing.

C. Experimental Results

Recall that we use quadruplet constraints and our metric learning model (18) is:

$$\min_{\mathbf{M}\in\mathbb{S}_+^d} \sum_{q\in\mathcal{A}} [\xi_q + W_{\gamma,\mathbf{M}}(P,Q) - W_{\gamma,\mathbf{M}}(R,S)]_+ - \sum_{q\in\mathcal{A}} \mathcal{L}_{order}.$$

where $\mathcal{L}_{order} = \lambda_1 \left(I(\mathbf{F}_{PQ}) + I(\mathbf{F}_{RS}) \right) - \lambda_2 (KL(\mathbf{F}_{PQ}\|\mathbf{P}) + KL(\mathbf{F}_{RS}\|\mathbf{P}))$ and $\mathbf{F}_{PQ} \in \mathbb{F}(P,Q)$, $\mathbf{F}_{RS} \in \mathbb{F}(R,S)$. Because of the setting of the experiments, instead of quadruplet constraints, we only construct the triplet constraints according to whether two samples are from same class or not. The core idea is that samples from same class are more similar than samples from different classes. Let $\mathcal{R}$ denote the triplet constraints, then the above model degenerates to

$$\begin{aligned} \min_{\mathbf{M}\in\mathbb{S}_+^d} & \sum_{(P,Q,R)\in\mathcal{R}} [\xi_q + W_{\gamma,\mathbf{M}}(P,Q) - W_{\gamma,\mathbf{M}}(P,R)]_+ \\ & - \sum_{(P,Q,R)\in\mathcal{R}} \mathcal{L}_{order}, \end{aligned} \tag{27}$$

where $\mathcal{L}_{order} = \lambda_1 \left(I(\mathbf{F}_{PQ}) + I(\mathbf{F}_{PR}) \right) - \lambda_2 (KL(\mathbf{F}_{PQ}\|\mathbf{P}) + KL(\mathbf{F}_{PR}\|\mathbf{P}))$ and $\mathbf{F}_{PQ} \in \mathbb{F}(P,Q)$, $\mathbf{F}_{PR} \in \mathbb{F}(P,R)$. In our setting, $(P,Q) \in \mathcal{S}$ and $(P,R) \in \mathcal{D}$. $\mathcal{S}$ and $\mathcal{D}$ denote similar pair set and dissimilar pair set.

1) Sepsis Prediction: At the time of sepsis onset, our method demonstrated a relatively higher mean accuracy using different lengths of sequences, *i.e.*, 84.42%, 83.39% and 85.18% using 6, 12 and 24 hours data, respectively. Another metric learning method RVSML-DTW also achieved comparable performance using 12 and 24 hours data. Both our method and RVSML-OPW were based on OPW distance, but we got a relatively better results than RVSML-OPW.

Table II: Classification accuracy (%) of different methods on sepsis prediction task using 6 hours data.

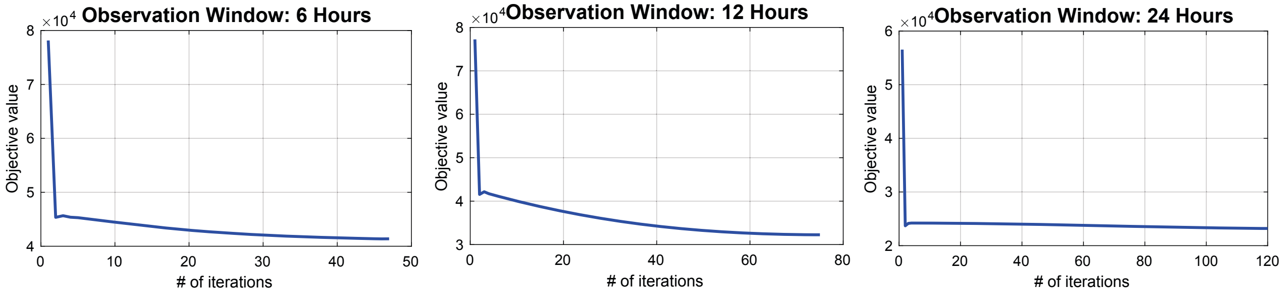
	Prediction Window									
	Onset		**3 hours**		**6 hours**		**9 hours**		**12 hours**	
Methods	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
DTW [15]	80.36	0.796	78.74	1.700	76.70	1.388	75.14	0.817	75.44	1.505
OPW [19]	80.46	0.499	78.70	1.396	76.71	1.274	75.14	0.823	75.58	1.399
RVSML-DTW [22]	78.28	1.971	76.71	1.960	75.11	1.712	75.85	1.131	72.27	0.894
RVSML-OPW [22]	80.47	5.089	71.04	7.073	77.77	8.129	74.99	4.784	65.19	6.723
Our	84.42	1.690	81.72	1.760	79.33	0.760	78.07	1.550	76.87	1.470

Table III: Classification accuracy (%) of different methods on sepsis prediction task using 12 hours data.

	Prediction Window									
	Onset		**3 hours**		**6 hours**		**9 hours**		**12 hours**	
Methods	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
DTW [15]	80.77	1.269	79.08	0.625	78.23	1.162	77.80	1.162	76.01	1.007
OPW [19]	81.00	1.107	79.18	1.008	78.45	1.126	77.37	1.126	76.22	0.741
RVSML-DTW [22]	78.88	1.671	76.44	1.741	75.14	1.476	71.77	1.476	69.41	1.987
RVSML-OPW [22]	80.01	2.427	82.27	0.926	75.49	3.112	75.51	3.112	64.66	2.563
Our	83.39	1.350	80.90	1.340	79.54	1.390	78.51	1.390	77.63	1.350

Table IV: Classification accuracy (%) of different methods on sepsis prediction task using 24 hours data.

	Prediction Window									
	Onset		**3 hours**		**6 hours**		**9 hours**		**12 hours**	
Methods	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
DTW [15]	80.54	1.945	79.08	1.898	80.42	1.457	79.40	1.652	78.26	2.647
OPW [19]	80.83	2.261	78.30	2.186	80.16	1.850	78.97	1.900	77.47	2.546
RVSML-DTW [22]	80.58	2.159	80.65	1.657	78.61	1.555	79.10	2.537	76.87	3.607
RVSML-OPW [22]	84.47	1.379	80.16	2.300	83.38	1.340	77.95	3.773	75.42	1.804
Our	85.18	0.630	82.30	1.850	81.81	1.370	80.78	1.920	79.48	1.060



(a) 6 hours data

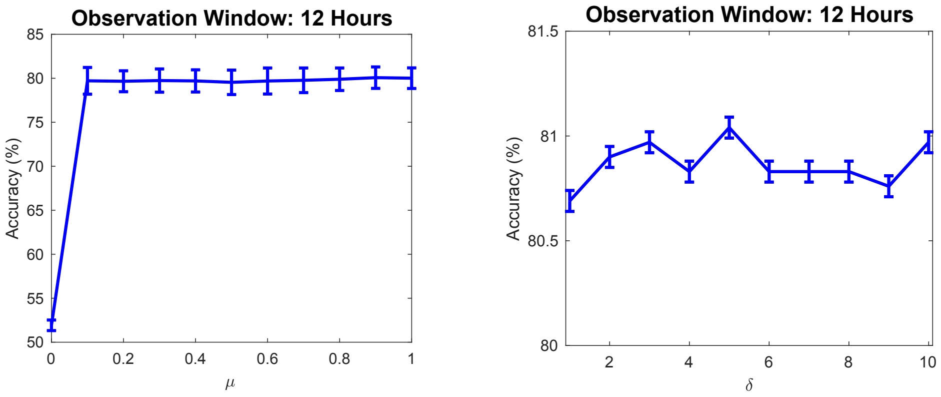
(b) 12 hours data

(c) 24 hours data

Figure 4: Objective function value vs iterations.

Tables II, III and IV show the performance measures before sepsis onset upon different observation windows. At 3, 6, 9 and 12 hours before onset, our method achieved better results at most of times.

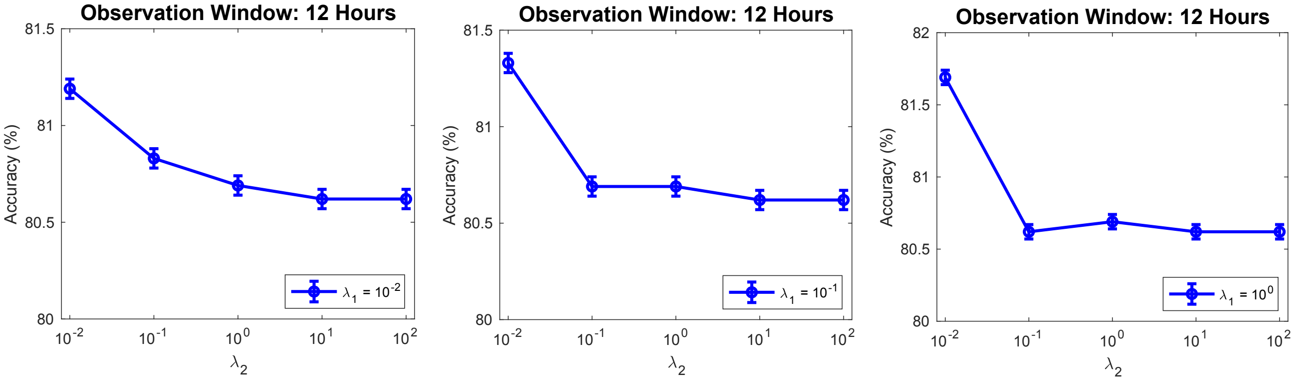
2) Effect of Sequence Length: We examined the effect of the sequence length (*i.e.*, observation window) on the predictive performance at different time points before sepsis onset, as shown in Table, II, III and IV. In most cases, longer observation window corresponds to better performance. Because the data are usually with problems of missing variables, different sampling frequencies and sampling frequencies, longer observation window is sometimes benefit



(a) μ

(b) δ

Figure 5: Accuracy vs (a) μ and (b) δ.



(a) $\lambda_1 = 0.01$

(b) $\lambda_1 = 0.1$

(c) $\lambda_1 = 1$

Figure 6: Accuracy vs λ_1 and λ_2.

for extracting more useful information.

3) Effect of Constraints: As we stated before, LMNN is one of the most widely used metric learning methods which uses triplet constraint [30]. It combines two terms into a single loss function, *i.e.*, one is to attract target neighbors and the other one is to repel impostors. If we also construct the loss function as LMNN, we will have

$$\min_{\mathbf{M}\in\mathbb{S}_+^d} \quad \mu \sum_{(P,Q,R)\in\mathcal{R}} [\xi_q + W_{\gamma,\mathbf{M}}(P,Q) - W_{\gamma,\mathbf{M}}(P,R)]_+ \\ + (1-\mu) \sum_{(P,Q)\in\mathcal{S}} W_{\gamma,\mathbf{M}}(P,Q) - \sum_{(P,Q,R)\in\mathcal{R}} \mathcal{L}_{order}, \tag{28}$$

where μ balances the first two terms. The first term is to repel impostors, and the second term to attract neighbors. In this experiment, we tested whether the additional term improved the results compared to (27). We fixed λ_1, λ_2 to 1 and δ to 5. Both observation window and prediction window are set to 12 hours. The performance with respect to different μ are shown in Fig. 5(a). From the figure, we can find that except for $\mu = 0$, other results did not depend sensitively on the value of μ. In practice, the value $\mu = 1$ worked well. That is, instead of having an addition term as in Eq (28), model (27) is good enough in our experiments.

4) Convergence Analysis: To illustrate the convergence of our method, we set prediction window as 12 hours and conduct several experiments on MIMIC-III dataset. The objective function values versus number of iterations are shown in Fig. 4. From the figure we can see that the objective values reduce reasonably well. Especially at first several iterations, the objective value is drastically reduced. That is, we can obtain a relatively good metric matrix with very few iterations.

D. Influence of Parameters

The proposed metric learning model has three hyper-parameters due to the introduction of order-preserving Wasserstein distance: λ_1, λ_2 and standard deviation δ of the prior distribution. λ_1 controls the weight of the inverse difference moment regularization, λ_2 controls the balance of the regularization in terms of the KL-divergence with prior distribution and δ controls the expected bandwidth of warping [19]. In this experiment, we test the influences of the three parameters.

1) Effect of δ: We fix λ_1, λ_2 to 1, and evaluate the performances with respect to different δ. The results are shown in Fig. 5(b). We can find that results are not sensitive to δ. We set δ to 5 in other experiments.

2) Effect of λ_1 and λ_2: We fix δ to 5 and evaluate the performances with different λ_1 and λ_2. We set λ_1 to 10^{-2}, 10^{-1} and 1 and tuned λ_2 from range $\{10^{-2}, 10^{-1}, 1, 10, 10^2\}$. The results are shown in Fig. 6. Our method is not sensitive to λ_1 when $\lambda_2 >= 1$. We set $\lambda_1 = 1$ and $\lambda_2 = 1$ in other experiments. Both observation window and prediction window are set to 12 hours for all the influence of parameters experiments.

VI. Conclusion

In this paper, we introduce an order-preserving metric learning framework for multivariate time series prediction. Quadruplet constraints are used to model similarity between sequences. Besides, we integrate the order-preserving Wasserstein distance into the framework where two temporal regularization terms including inverse difference moment regularization and KL-divergence are enforced to the flow-network to preserve the inherent temporal relationships of the variables in MTS. In addition, Mahalanobis distance is used as the ground measurement of Wasserstein distance to increase its flexibility. Finally, extensive experiments on a real-world clinical data set demonstrate the effectiveness and convergence of the proposed method.

Acknowledgment

The work is partially supported by National Science Foundation under grant number 1750326 and 2027970, and Office of Naval Research under grant number N00014-18-1-2585. The authors would also like to acknowledge the support from American Air Liquide Inc.

References

[1] D. Ding, M. Zhang, X. Pan, M. Yang, and X. He, "Modeling extreme events in time series prediction," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 1114–1122.

[2] C. Fan, Y. Zhang, Y. Pan, X. Li, C. Zhang, R. Yuan, D. Wu, W. Wang, J. Pei, and H. Huang, "Multi-horizon time series forecasting with temporal attention learning," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2527–2535.

[3] M. Sun, S. Zhao, C. Gilvary, O. Elemento, J. Zhou, and F. Wang, "Graph convolutional networks for computational drug development and discovery," *Briefings in bioinformatics*, vol. 21, no. 3, pp. 919–935, 2020.

[4] S. Zhao, C. Su, Z. Lu, and F. Wang, "Recent advances in biomedical literature mining," *Briefings in Bioinformatics*, 2020.

[5] E. Sheetrit, N. Nissim, D. Klimov, and Y. Shahar, "Temporal probabilistic profiles for sepsis prediction in the icu," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2961–2969.

[6] J. Xu, Z. Xu, P. Walker, and F. Wang, "Federated patient hashing." in *AAAI*, 2020, pp. 6486–6493.

[7] S. Zhao, T. Liu, S. Zhao, and F. Wang, "A neural multi-task learning framework to jointly model medical named entity recognition and normalization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 817–824.

[8] Z. Xing, J. Pei, and E. Keogh, "A brief survey on sequence classification," *ACM Sigkdd Explorations Newsletter*, vol. 12, no. 1, pp. 40–48, 2010.

[9] A. Abanda, U. Mori, and J. A. Lozano, "A review on distance based time series classification," *Data Mining and Knowledge Discovery*, vol. 33, no. 2, pp. 378–412, 2019.

[10] A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh, "The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances," *Data Mining and Knowledge Discovery*, vol. 31, no. 3, pp. 606–660, 2017.

[11] N. A. Chuzhanova, A. J. Jones, and S. Margetts, "Feature selection for genetic sequence classification." *Bioinformatics (Oxford, England)*, vol. 14, no. 2, pp. 139–143, 1998.

[12] X. Ji, J. Bailey, and G. Dong, "Mining minimal distinguishing subsequence patterns with gap constraints," *Knowledge and Information Systems*, vol. 11, no. 3, pp. 259–286, 2007.

[13] D. D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," in *European conference on machine learning*. Springer, 1998, pp. 4–15.

[14] P. K. Srivastava, D. K. Desai, S. Nandi, and A. M. Lynn, "Hmm-mode–improved classification using profile hidden markov models by optimising the discrimination threshold and modifying emission probabilities with negative training sequences," *BMC bioinformatics*, vol. 8, no. 1, p. 104, 2007.

[15] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series." in *KDD workshop*, vol. 10, no. 16. Seattle, WA, 1994, pp. 359–370.

[16] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, no. 1, pp. 43–49, 1978.

[17] B. Ni, P. Moulin, and S. Yan, "Order-preserving sparse coding for sequence classification," in *European Conference on Computer Vision*. Springer, 2012, pp. 173–187.

[18] M. Muskulus and S. Verduyn-Lunel, “Wasserstein distances in the analysis of time series and dynamical systems,” *Physica D: Nonlinear Phenomena*, vol. 240, no. 1, pp. 45–58, 2011.

[19] B. Su and G. Hua, “Order-preserving wasserstein distance for sequence matching,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1049–1057.

[20] F. Wang and J. Sun, “Survey on distance metric learning and dimensionality reduction in data mining,” *Data mining and knowledge discovery*, vol. 29, no. 2, pp. 534–564, 2015.

[21] J. Xu, L. Luo, C. Deng, and H. Huang, “Multi-level metric learning via smoothed wasserstein distance.” in *IJCAI*, 2018, pp. 2919–2925.

[22] B. Su and Y. Wu, “Learning distance for sequences by learning a ground metric,” in *International Conference on Machine Learning*, 2019, pp. 6015–6025.

[23] R. Sandler and M. Lindenbaum, “Nonnegative matrix factorization with earth mover’s distance metric for image analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1590–1602, 2011.

[24] M. Cuturi, “Sinkhorn distances: Lightspeed computation of optimal transport,” in *NIPS*, 2013, pp. 2292–2300.

[25] R. Sinkhorn, “Diagonal equivalence to matrices with prescribed row and column sums,” *The American Mathematical Monthly*, vol. 74, no. 4, pp. 402–405, 1967.

[26] M. T. Law, N. Thome, and M. Cord, “Fantope regularization in metric learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1051–1058.

[27] J. Xu, L. Luo, C. Deng, and H. Huang, “New robust metric learning model using maximum correntropy criterion,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 2555–2564.

[28] ——, “Bilevel distance metric learning for robust image recognition,” in *Advances in Neural Information Processing Systems*, 2018, pp. 4198–4207.

[29] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, “Information-theoretic metric learning,” in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 209–216.

[30] K. Q. Weinberger and L. K. Saul, “Distance metric learning for large margin nearest neighbor classification,” *Journal of Machine Learning Research*, vol. 10, no. Feb, pp. 207–244, 2009.

[31] M. T. Law, N. Thome, and M. Cord, “Quadruplet-wise image similarity learning,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 249–256.

[32] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, “Mimic-iii, a freely accessible critical care database,” *Scientific data*, vol. 3, p. 160035, 2016.

[33] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith *et al.*, “The third international consensus definitions for sepsis and septic shock (sepsis-3),” *Jama*, vol. 315, no. 8, pp. 801–810, 2016.

[34] C. W. Seymour, J. N. Kennedy, S. Wang, C.-C. H. Chang, C. F. Elliott, Z. Xu, S. Berry, G. Clermont, G. Cooper, H. Gomez *et al.*, “Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis,” *Jama*, vol. 321, no. 20, pp. 2003–2017, 2019.

[35] S. Nemati, A. Holder, F. Razmi, M. D. Stanley, G. D. Clifford, and T. G. Buchman, “An interpretable machine learning model for accurate prediction of sepsis in the icu.” *Critical care medicine*, vol. 46, no. 4, pp. 547–553, 2018.