Multi-Task Sparse Metric Learning for Monitoring Patient Similarity Progression

Qiuling Suo*, Weida Zhong*, Fenglong Ma*, Ye Yuan†, Mengdi Huai*, Aidong Zhang*
*Department of Computer Science and Engineering, State University of New York at Buffalo, NY, USA
Email: {qiulings, weidazho, fenglong, mengdihu, azhang}@buffalo.edu
†College of Information and Communication Engineering,
Beijing University of Technology, Beijing, China
Email: yuanye91@emails.bjut.edu.cn

***Abstract*—A clinically meaningful distance metric, which is learned from measuring patient similarity, plays an important role in clinical decision support applications. Several metric learning approaches have been proposed to measure patient similarity, but they are mostly designed for learning the metric at only one time point/interval. It leads to a problem that those approaches cannot reflect the similarity variations among patients with the progression of diseases. In order to capture similarity information from multiple future time points simultaneously, we formulate a multi-task metric learning approach to identify patient similarity. However, it is challenging to directly apply traditional multi-task metric learning methods to learn such similarities due to the high dimensional, complex and noisy nature of healthcare data. Besides, the disease labels often have clinical relationships, which should not be treated as independent. Unfortunately, traditional formulation of the loss function ignores the degree of labels' similarity. To tackle the aforementioned challenges, we propose *mtTSML*, a multi-task triplet constrained sparse metric learning method, to monitor the similarity progression of patient pairs. In the proposed model, the distance for each task can be regarded as the combination of a common part and a task-specific one in the transformed low-rank space. We then perform sparse feature selection for each individual task to select the most discriminative information. Moreover, we use triplet constraints to guarantee the margin between similar and less similar pairs according to the ordered information of disease severity levels (i.e. labels). The experimental results on two real-world healthcare datasets show that the proposed multi-task metric learning method significantly outperforms the state-of-the-art baselines, including both single-task and multi-task metric learning methods.**

***Index Terms*—Metric Learning; Patient Similarity; Multi-task Learning; Sparse Regularization**

I. Introduction

In healthcare domain, understanding a patient's health status and providing a reasonable treatment plan are important for both doctors and patients. An effective and efficient way for doctors is to find similar patients and refer to their successful treatment plans. However, it is difficult to identify similar patients from a large number of healthcare data even according to rich experience of doctors. Thus, automatically recognizing patients with similar symptoms and disease history is a challenging problem. To solve this challenge, we need to learn a clinical meaningful metric which measures the relative similarities between a pair of patients based on their medical indicators. A proper similarity measure enables various downstream clinical applications, such as personalized medicine [1, 2], medical diagnoses [3], trajectory analysis [4] and cohort study [5].

In fact, we can use some simple metrics such as Euclidean distance to measure the similarity among patients. However, the obtained distances cannot reflect the statistical regularities specified by the supervised information from a desired task. To address this issue, metric learning algorithms are proposed and shown their superiority under various scenarios [6], such as image recognition and document retrieval. In healthcare domain, several studies are proposed to measure patient similarity based on the metric learning methods [5, 7–12]. The working process of these models includes two steps: (1) Transforming samples in the original space to a new space, either through a linear [7, 8, 10] or non-linear [5, 9, 11] operation, and (2) calculating their distances in the new space based on the label information. The label information can be binary with control and disease cohorts [10], multiple independent diseases [7, 9] or constructed from medical knowledge [11].

However, all the aforementioned methods are designed for measuring patient similarity at only one time point/interval. Actually, *patients' health condition is ever changing with the progression of diseases*. For example, in the study of Alzheimer's disease (AD), a patient with a current diagnosis "mild cognitive impairment (MCI)" is considered to be similar with the cohort of mild diseased patients, but s/he may gradually become more similar to the severely diseased group after a few months, as the disease status changes. On the contrary, the status of another MCI patient may remain stable for a long time. Therefore, it is better to **predict patient similarity on multiple future time points** (i.e. multi-task metric learning), which can provide a more comprehensive study to enable personalized healthcare compared with performing a one-time prediction.

Existing work for multi-task metric learning [13–16] treats the Mahalanobis matrix as the combination of a common part and a task-specific one, which results in a convex optimization problem. However, these approaches cannot be directly applied to measure patient similarity due to the characteristics of healthcare data. Healthcare data collected from real-world clinical systems are usually high dimensional, sparse and noisy, i.e., containing a lot of redundant and irrelevant information.

Incorporating these features directly may hide the discriminative information, resulting in poor performance of existing multi-task metric learning models. Extracting informative features manually by experts will cost huge expenses and efforts. Therefore, a multi-task metric learning model should be able to perform sparse feature selection to remove the effect from redundant and irrelevant features and capture the most relevant information for each individual task. Nevertheless, it is not easy to explicitly perform sparse feature selection for existing multi-task metric learning models. Moreover, as the feature dimension increases, the dimensions of the Mahalanobis matrix increase, which may cause the overfitting issue.

Additionally, in the medical field, the label information of different classes often has clinical relations. For example, there are multiple stages of Alzheimer's disease, indicating the severity levels during the disease progression process. The symptoms of "severe" stage should be more similar with "moderate" than with the "mild" stage. In supervised metric learning, a common way to obtain the similarity label of a sample pair is to denote them as similar if they belong to the same class, otherwise dissimilar. In this way, the similarity degree among labels will be ignored, which may not provide sufficient information to the learner. Therefore, the multi-task metric learning approach should be able to provide the similarity degree relationship, as well as the pairwise label information.

To tackle all the aforementioned problems, in this paper, we propose a multi-task Triplet constrained Sparse Metric Learning method (mtTSML) to measure the progression of patient similarity over time. The input data are the attributes measured at the screening or baseline time, and each task is to learn the distance metric at a future time point. To select informative features from the high dimensional inputs, we first decompose the common and task-specific Mahalanobis matrix by low-rank transformation matrices. In this way, the Mahalanobis distance can be formulated as the combination of a common distance shared by all tasks and a task-specific one. In the transformation matrix, each column can be regarded as a vector which measures the importance of the corresponding features. We then perform feature selection on the transformation matrix for each task, through introducing the $\ell_{2,1}$ [17] regularization terms, which sets a number of non-informative feature columns to zero. To consider the similarity degree among disease labels, we construct triplet constraints by forcing a margin between similar pairs and less similar pairs. The designed triplet constraints not only force patients with the same disease labels to be similar and different labels to be dissimilar, but also successfully models the ordered label relationship.

Our main contributions can be summarized as follows:

- We propose a multi-task sparse metric learning method based on triplet constraints to monitor patient similarity progression over time, which is capable of learning patient distances at multiple future time points of interest simultaneously. The multi-task formulation improves the generalization performance for both diagnosis and prognosis tasks. With learned distance metrics, clinical studies can be performed to monitor the trend of similarity variations.
- We perform sparse feature selection during the multi-task learning process, which removes non-discriminative features from the high dimensional input space for each individual task.
- We incorporate the similarity degree information by considering the ordered relationship of disease labels in formulating the distance constraints. In this way, severity levels of the disease can be well reflected.
- Experimental results on two real healthcare datasets show that our proposed method significantly outperforms state-of-the-art single task learning and multi-task learning methods.

II. Methodology

In this section, we first review a typical traditional framework of multi-task metric learning, and then introduce the proposed method.

A. Preliminary

Most metric learning methods [18–21] aim at finding the Mahalanobis distance metric, due to its simplicity and flexibility. The Mahalanobis distance can be seen as the Euclidean distance after performing a linear transformation on the input space. The Mahalanobis distance between two vectors $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^D$ is defined as,

$$d^2(\mathbf{x}_i - \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M}(\mathbf{x}_i - \mathbf{x}_j), \tag{1}$$

where $\mathbf{M} \in \mathbb{R}^{D\times D}$ is a positive semidefinite (PSD) matrix to be learned, and D is the size of feature dimension. A popular way to construct a multi-task metric learning model is by sharing the composition of Mahalanobis matrices [13, 14, 22]. The Mahalanobis matrix of each task is assumed to be composed of a common part $\mathbf{M}_0$ shared by all the tasks and a task-specific part $\mathbf{M}_t$ preserving its specific properties. Thus, the distance between two points defined by the metric of the t-th task is,

$$d_t^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top (\mathbf{M}_0 + \mathbf{M}_t)(\mathbf{x}_i - \mathbf{x}_j), \tag{2}$$

where $\mathbf{M}_0, \mathbf{M}_t \in \mathbb{R}^{D\times D}$. Intuitively, the metric defined by $\mathbf{M}_0$ picks up general trends across multiple data sets and $\mathbf{M}_t$ specializes the metric for each particular task. The objective function is formulated as,

$$\min_{\mathbf{M}_0,\ldots,\mathbf{M}_T} \mathcal{L}(\mathbf{M}_0, \mathbf{M}_t) + \gamma_0 \text{Reg}(\mathbf{M}_0) + \sum_{t=1}^{T} \gamma_t \text{Reg}(\mathbf{M}_t), \tag{3}$$

where $\mathcal{L}(\cdot)$ is the loss function, T is the number of tasks, $\text{Reg}(\cdot)$ is a regularization function, and γ_0 and γ_t are the trade-off parameters. $\mathcal{L}(\cdot)$ can be contrastive loss or triplet loss, and L-2 norm is usually used in $\text{Reg}(\cdot)$.

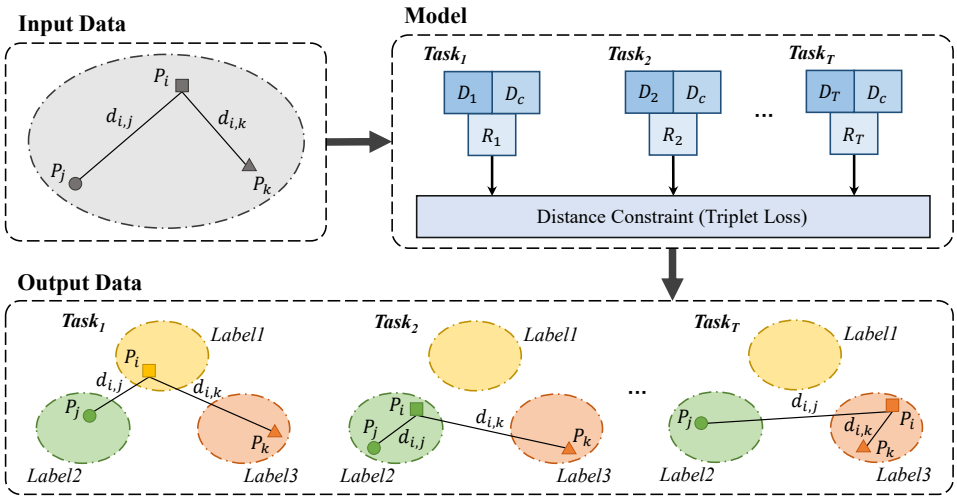


Fig. 1: The training process of the proposed method on measuring patient similarity progression. We use the combination of a common distance D_c and task-specific ones $D_1, D_2, ..., D_T$ to represent the desired distances, and then formulate the constraint function based on the distances. $R_1, R_2, ..., R_T$ are the sparse regularization terms.

B. The Proposed Multi-task Metric Learning Method

Healthcare data collected from real-world applications often tend to be high dimensional, complex and noisy. Therefore, a proper model on learning patient similarity should be able to work on such kind of data. In this paper, we propose a multi-task learning framework that learns the distance through a coupled low-rank sparse transformation. Moreover, we consider the similarity degree among patient groups using a distance constrict based on triplet loss. The training process of the proposed method can be seen in Fig. 1.

In Fig. 1, there are three groups of patients labeled as 1, 2 and 3. Taking a triplet of patients P_i, P_j and P_k for example, the distances shown in the input space do not reflect the label information. Through metric learning, we can obtain their relative distances in a new space according to their labels. As patients' health conditions are ever changing, we monitor the changes of their distances in several future time points of interest. Learning patient similarity at each timestamp is regarded as one task. As illustrated in Fig. 1, through multi-task metric learning, the condition of the anchor patient P_i gradually changes. In the following subsections, we introduce details of the proposed method.

1) Low-Rank Metric Formulation: The formulation in Eq. (3) is often limited to low dimensional features, which may not be suitable for real-world healthcare data. We seek a low rank matrix $\mathbf{L} \in \mathbb{R}^{r \times D}$ with $r \ll D$ by decomposing $\mathbf{M}$ as $\mathbf{M} = \mathbf{L}^\top \mathbf{L}$. With this decomposition, the PSD constraint of $\mathbf{M}$ can be automatically satisfied. Although the optimization problem becomes non-convex in terms of $\mathbf{L}$, it is usually not an issue, and very good results can be achieved [23–25]. Through decomposing $\mathbf{M}_0$ and $\mathbf{M}_t$, the distance function for task t as defined in Eq. (2) can be written as,

$$\begin{aligned} d_t^2(\mathbf{x}_i, \mathbf{x}_j) &= (\mathbf{x}_i - \mathbf{x}_j)^\top (\mathbf{L}_0^\top \mathbf{L}_0 + \mathbf{L}_t^\top \mathbf{L}_t)(\mathbf{x}_i - \mathbf{x}_j) \\ &= \|\mathbf{L}_0 \mathbf{x}_i - \mathbf{L}_0 \mathbf{x}_j\|^2 + \|\mathbf{L}_t \mathbf{x}_i - \mathbf{L}_t \mathbf{x}_j\|^2, \end{aligned} \tag{4}$$

where $\mathbf{L}_0 \in \mathbb{R}^{r \times D}$ is a global transformation matrix and $\mathbf{L}_t \in \mathbb{R}^{r \times D}$ is the task-specific one. The individual distance turns out to be the combination of the common distance and a task-specific one.

2) Sparse Feature Selection: The real-world healthcare data may contain a lot of irrelevant and redundant information. Blindly incorporating these information may hide the relationship between labels and informative features. Therefore, it is essential to select the most discriminative features during the learning process.

We use sparse regularization for feature selection and dimension reduction. Suppose $\mathbf{l}_j$ is the j-th column vector of the transformation matrix $\mathbf{L}$, i.e., $\mathbf{L} = [\mathbf{l}_1, \mathbf{l}_2, ..., \mathbf{l}_D]$, and $\|\mathbf{l}_j\| = (\sum_{i=1}^{r} \mathbf{L}_{ij}^2)^{1/2}$. Then $\mathbf{l}_j$ can be regarded as a measurement of the importance of the j-th feature. If $\|\mathbf{l}_j\| = 0$, then the j-th feature in the transformed space becomes zero. We expect that only a small amount of $\mathbf{l}_j$s are non-zero, which transforms the discriminative features to a low dimensional representation. Therefore, the sparse regularization on $\mathbf{L}$ can be written as,

$$\|\mathbf{L}\|_{2,1} = \sum_{j=1}^{D} \|\mathbf{l}_j\| = \sum_{j=1}^{D} (\sum_{i=1}^{r} \mathbf{L}_{ij}^2)^{1/2}. \tag{5}$$

This regularization term enforces some $\mathbf{l}_j$s to be zero vectors, and the corresponding features will not be selected, yielding feature sparsity.

3) Distance Constraint Construction: As mentioned in Section I, there are multiple class labels to describe the status of a disease, and there exists ordered relationship among labels according to the severity levels [26]. We believe that incorporating such similarity degree information could help the model to better capture the relative distance information. Recent work on measuring patient similarity in [11] incorporates the fine-grained label information by using a quadruplet which consists of an anchor patient, a positive/negative patient and a partial

similar patient. However, this approach does not work in our scenario, as one partial similarity label ignores the multiple levels of severities.

We formulate triplet constraints to incorporate the similarity degree among classes. We denote the instance set as $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$ and its associated label set $\mathcal{C} = \{c_i\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^D$, $c_i \in \mathbb{R}$ reflects the disease stage, and N is the number of patients. Considering a triplet of patients $\mathbf{x}_i$, $\mathbf{x}_j$ and $\mathbf{x}_k$ from $\mathcal{X}$, the following sets of constraints can be constructed,

• $\mathcal{R}^1 = \{(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k), \quad c_i < c_j < c_k\}$. In this set, although $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to different classes, we can say that $\mathbf{x}_i$ is more similar to $\mathbf{x}_j$ than $\mathbf{x}_k$, according to the ordered label information that c_i is closer to c_j than to c_k.

• $\mathcal{R}^2 = \{(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k), \quad c_i > c_j > c_k\}$. Similarly, in this set, $\mathbf{x}_i$ should be more similar to $\mathbf{x}_j$ than to $\mathbf{x}_k$.

• $\mathcal{R}^3 = \{(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k), \quad c_i = c_j \neq c_k\}$. In this set, $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same class, while $\mathbf{x}_k$ has a different class label. Obviously, $\mathbf{x}_i$ should be more similar to $\mathbf{x}_j$ than to $\mathbf{x}_k$.

The above sets pose distance constraints that describe the relative relationships among the class labels. In the combined set $\mathcal{R} = \mathcal{R}^1 \cup \mathcal{R}^2 \cup \mathcal{R}^3$, the distance between x_i and x_j should not be larger than that between $\mathbf{x}_i$ and $\mathbf{x}_k$ (i.e., $d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, \mathbf{x}_k)$). The distance constraints need to be satisfied during the training process. Following the large margin principle, we formulate the constraint as,

$$d^2(\mathbf{x}_i, \mathbf{x}_j) \leq d^2(\mathbf{x}_i, \mathbf{x}_k) - g, \tag{6}$$

where $g > 0$ is a margin parameter to ensure a sufficiently large difference. This inequality constraint ensures that the distance between an anchor sample and its similar sample should be smaller than that between the less similar one with some fixed margin. Therefore, metric learning needs to optimize the following hinge loss,

$$\sum_{(i,j,k)\in\mathcal{R}} [d^2(\mathbf{x}_i, \mathbf{x}_j) - d^2(\mathbf{x}_i, \mathbf{x}_k) + g]_+, \tag{7}$$

where $[\cdot]_+ = \max(\cdot, 0)$. If the constraint in Eq. (6) does hold, the corresponding triplet makes no contribution to the loss function. The large number of triplets by fully considering the constraints in Eq. (6) will result in heavy computational cost and slow convergence. In practice, we calculate local constraints between each sample point and its neighborhood instead of all the samples.

4) Learning Framework: Suppose there are T tasks, we need to learn the shared transformation $\mathbf{L}_0$, and the specific transformations $\{\mathbf{L}_1, \mathbf{L}_2, ..., \mathbf{L}_T\}$ jointly for all the tasks. We formulate the overall loss function as below,

$$\min_{\mathbf{L}_0,\ldots,\mathbf{L}_T} \mathcal{F} = \sum_{t=1}^{T} \left[\mathcal{L}_t(\mathbf{L}_0, \mathbf{L}_t) + \gamma_t \|\mathbf{L}_t\|_{2,1}\right] + \gamma_0 \|\mathbf{L}_0\|_F^2, \tag{8}$$

where $\mathcal{L}_t(\cdot) = \frac{1}{|\mathcal{R}_t|} \sum_{(i,j,k)\in\mathcal{R}_t} [d_t^2(\mathbf{x}_i, \mathbf{x}_j) - d_t^2(\mathbf{x}_i, \mathbf{x}_k) + g]_+$, $d(\mathbf{x}_i, \mathbf{x}_j)$ is formulated in Eq. (4), $\mathcal{R}_t$ is the set of triplets for the t-th task, $\|\cdot\|_F$ is the Frobenius norm, and γ_0 and γ_t are the trade-off parameters. On one hand, If $\gamma_0 \to \infty$, then $\mathbf{L}_0$ approaches a zero matrix, which means that there is no shared metric among tasks, i.e., Eq. (8) reduces to T independent models. On the other hand, if $\gamma_t \to \infty$, the task-specific transformations $\mathbf{L}_t$ become irrelevant zero matrices, and we learn a single metric $\mathbf{L}_0$ across all the tasks. In Eq. (8), we use the Frobenius norm to penalize large value of elements in common metric $\mathbf{L}_0$, and make the transformed space to be low dimensional by setting $r \ll D$. Then we use the regularization term $\|\mathbf{L}_t\|_{2,1}$ to perform sparse feature selection for each individual task. Since the $\ell_{2,1}$ term is non-differentiable, inspired by [7, 10], we apply the alternating direction method of multipliers (ADMM) [27] to optimize the loss function.

5) Optimization: We first transform Eq. (8) to the following equivalent problem:

$$\begin{aligned} \min_{\mathbf{L}_0,\mathbf{L}_t,\mathbf{W}_t} \mathcal{F} &= \sum_{t=1}^{T} \left[\mathcal{L}_t(\mathbf{L}_0, \mathbf{L}_t) + \gamma_t \|\mathbf{W}_t\|_{2,1}\right] + \gamma_0 \|\mathbf{L}_0\|_F^2, \\ s.t. \quad \mathbf{L}_t &= \mathbf{W}_t, \quad \text{for} \quad t = 1, 2, ..., T. \end{aligned} \tag{9}$$

Through introducing the Lagrange multipliers $\{\mathbf{Y}_t \in \mathbb{R}^{r\times D}\}$, the following function can be obtained,

$$\begin{aligned} \min_{\mathbf{L}_0,\ldots,\mathbf{L}_T} \mathcal{F} = \sum_{t=1}^{T} \Big[&\mathcal{L}_t(\mathbf{L}_0, \mathbf{L}_t) + \gamma_t \|\mathbf{W}_t\|_{2,1}^2 + \langle \mathbf{Y}_t, \mathbf{L}_t - \mathbf{W}_t \rangle \\ &+ \frac{\rho}{2} \|\mathbf{L}_t - \mathbf{W}_t\|_F^2 \Big] + \gamma_0 \|\mathbf{L}_0\|_F^2, \end{aligned} \tag{10}$$

where $\langle \mathbf{A}, \mathbf{B} \rangle$ is the inner product of matrices $\mathbf{A}$ and $\mathbf{B}$, and ρ is a nonnegative penalty hyper-parameter. We solve Eq. (10) via an iterative procedure based on ADMM. In the s-th iteration, the parameters $\mathbf{L}_0$, $\mathbf{L}_t$, $\mathbf{W}_t$ and $\mathbf{U}_t$ are updated as follows,

$$\begin{aligned} \mathbf{L}_0^{s+1} &\leftarrow \operatorname*{argmin}_{\mathbf{L}_0} \mathcal{L}_1(\mathbf{L}_0) + \gamma_0 \|\mathbf{L}_0\|_F^2, \\ \mathbf{L}_t^{s+1} &\leftarrow \operatorname*{argmin}_{\mathbf{L}_t} \mathcal{L}_2(\mathbf{L}_t) + \frac{\rho}{2} \|\mathbf{L}_t - \mathbf{W}_t^s + \mathbf{U}_t^s\|_F^2, \\ \mathbf{W}_t^{s+1} &\leftarrow \operatorname{argmin} \gamma_t \|\mathbf{W}_t\|_{2,1}^2 + \frac{\rho}{2} \|\mathbf{L}_t^{s+1} - \mathbf{W}_t + \mathbf{U}_t^s\|_F^2, \\ \mathbf{U}_t^{s+1} &\leftarrow \mathbf{U}^s + (\mathbf{L}_t^{s+1} - \mathbf{W}_t^{s+1}), \end{aligned} \tag{11}$$

where $\mathbf{U}_t = \frac{1}{\rho}\mathbf{Y}_t$. To update the matrices, we calculate the partial derivatives of the distance function $\mathcal{F}$ with respect to $\mathbf{L}_0$ and $\{\mathbf{L}_t\}_{t=1}^T$. The gradient of $\mathbf{L}_0$ is,

$$\frac{\partial \mathcal{F}}{\partial \mathbf{L}_0} = \gamma_0 \mathbf{L}_0 + \sum_{t=1}^{T} \sum_{(i,j,k)\in\mathcal{R}} \beta_{ijk}^t \left(\mathbf{L}_0 \mathbf{X}_{ij} - \mathbf{L}_0 \mathbf{X}_{ik}\right), \tag{12}$$

where $\mathbf{X}_{ij} = (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top$, and β_{ijk} is an indicator that controls whether the distance constraint is satisfied or not,

$$\beta_{ijk}^t = \begin{cases} 1, & \text{if} \quad d_t^2(\mathbf{x}_i, \mathbf{x}_j) + 1 - d_t^2(\mathbf{x}_i, \mathbf{x}_k) > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

The gradient of $\mathbf{L}_t$ can be calculated as,

$$\frac{\partial \mathcal{F}}{\partial \mathbf{L}_t} = \sum_{(i,j,k)\in\mathcal{R}} \beta_{ijk}^t \left(\mathbf{L}_t\mathbf{X}_{ij} - \mathbf{L}_t\mathbf{X}_{ik}\right) + \rho(\mathbf{L}_t - \mathbf{W}_t + \mathbf{U}_t). \tag{14}$$

Fixing $\mathbf{L}_t$ and $\mathbf{U}_t$, we use the proximal operator [28] to update $\mathbf{W}_t$ via an element-wise thresholding operation,

$$(\mathbf{W}_t^{s+1})_{ij} = (\mathbf{L}_t^{s+1} + \mathbf{U}_t^s)_{ij} \left[1 - \frac{\gamma_t}{\rho\|(\mathbf{L}_t^{s+1} + \mathbf{U}_t^s)_{i,:}\|_2}\right]_+. \tag{15}$$

Eq. (15) provides a closed form solution of $\mathbf{W}_t$. The proposed method is summarized in Algorithm 1.

Algorithm 1 Multi-task Sparse Metric Learning

Input: Triplet samples $\mathcal{R}_t$, γ_0, γ_t, ρ, α, α_t
Output: Transformation matrices $\mathbf{L}_0, \{\mathbf{L}_t\}_{t=1}^T$

```
Initialize L_0, {L_t}_{t=1}^T, {W_t}_{t=1}^T, {U_t}_{t=1}^T
repeat
    for t = 1 : T do
        update L_0 according to Eq. (12);
        update L_t according to Eq. (14);
        update W_t according to Eq. (15);
        update U_t according to Eq. (11);
    end for
until The stopping criterion is satisfied.
```

III. Experiments

We conduct experiments on two real-word datasets, and evaluate the performance of the proposed approach compared with existing state-of-the-art metric learning methods. Moreover, we use case studies to further demonstrate the intuition behind the proposed method.

A. Experimental Setup

1) Datasets: We use two healthcare datasets to validate the proposed approach.

• Alzheimer's Disease Neuroimaging Initiative (ADNI) [1] is a longitudinal project which aims to track the progression of the disease using biomarkers and clinical measures. Following the variables given in [29, 30], we extract 40 meta features, and combine them with 323 MRI features. The diagnosis of Alzheimer's disease is recorded every few months over years. There are three cohorts of patients: normal controls (NL), mild cognitive impairment (MCI), and Alzheimer's disease (AD) patients. We use the features at the baseline visit to study patient similarity at current and future time points, i.e., the relative distances among patients six months later, twelve months later, and so on.

• The study of osteoporotic fracture (SOF) [2] is a comprehensive study focused on bone diseases. It includes longitudinal visit records about osteoporosis of elder Caucasian women over 20 years. Potential risk factors and confounders belong to several categories such as demographics, family history, and lifestyle. We process the bone health status using the bone mineral density (BMD) values by comparison with young healthy references [31], resulting in three severity levels: normal, osteopenia and osteoporosis. Similarly, we use features measured at the first visit to predict status of bone disease progression in future visits.

Note that some diagnosis results are missing due to several reasons, such as patient's absence of visits and incomplete recording. Therefore, the number of patients in different tasks is not the same. Following [29], if there is no label for a patient in certain tasks, we then remove the data of the patient when training the corresponding tasks. Moreover, since the disease status is ever changing, the diagnosis results of one patient can be different over different visits. Data statistics are shown in Table I, where each task corresponds to one future visit.

TABLE I: Statistics of the ADNI and SOF datasets.

Dataset	# of samples					# of features
	Task1	Task2	Task3	Task4	Task5	
ADNI	732	693	615	425	105	364
SOF	539	544	542	230	540	200

2) Baseline Approaches: We compare our proposed multi-task sparse metric learning model with state-of-the-art single-task (ST) metric learning models and multi-task (MT) metric learning models, respectively.

• Single-task Learning Methods. In this set of methods, each task is trained separately, and there is no shared information among tasks. Therefore, we obtain different metrics for different tasks. LMNN [18] is a classical metric learning method, which pulls the k-nearest neighbors belonging to the same class closer, and separates examples from different classes by a large margin. ITML [20] learns the Mahalanobis distance by minimizing the differential relative entropy under the pairwise constraints between two multivariate Gaussians. GMML [19] formulates the learning process as an unconstrained smooth and convex optimization problem. SCML [21] learns a sparse combination of locally discriminative metrics, which regards the Mahalanobis matrix as a nonnegative weighted sum of multiple low-dimensional basis. LowRank [7] encodes a low-rank structure for the Mahalanobis matrix of bilinear similarity and performs sparse feature selection.

• Multi-task Learning Methods. This set of methods capture both shared and task-specific information. mtMLCS [24] assumes that all the tasks share a common low-dimensional subspace, and then exploits a task-specific projection. mt-SCML [21] is a multi-task version of SCML, which uses $\ell_{2,1}$ norm to perform group feature selection of the combination matrix. mt-LMNN [13] extends LMNN and optimizes the convex formulation on common Mahalanobis metric and task-specific one. CP-mtML [25] decomposes different tasks into a common projection and a task-specific projection, and learns the metric by optimizing on the transformation matrix.

In our problem setting, the disease stages of some patients may change during the monitored visits. Therefore, global

[1] https://adni.loni.usc.edu/

[2] https://sofonline.epi-ucsf.org/interface/

TABLE II: Performance comparison on the ADNI dataset in terms of MSE.

		ADNI (20% as training data)						ADNI (40% as training data)					
		Task1	Task2	Task3	Task4	Task5	Avg	Task1	Task2	Task3	Task4	Task5	Avg
ST	Euclidean	0.3795	0.4938	0.6369	0.6912	0.8583	0.6119	0.3617	0.4828	0.5818	0.6367	0.6470	0.5420
	Cosine	0.3654	0.5022	0.6508	0.6632	0.8052	0.5974	0.3443	0.4611	0.6017	0.6436	0.6134	0.5328
	GMML	0.2392	0.3616	0.5858	0.4376	0.4121	0.4072	0.2503	0.4221	0.5112	0.5450	0.3998	0.4257
	SCML	0.2096	0.3894	0.4481	0.5681	0.5147	0.4260	0.1995	0.3615	0.3864	0.4126	0.5217	0.3763
	LowRank	0.1699	0.2636	0.3104	0.4003	0.7528	0.3794	0.2027	0.3302	0.3646	0.3895	0.2568	0.3088
	ITML	0.1271	0.2227	0.3108	0.3481	0.3509	0.2719	0.1180	0.2290	0.2929	0.3615	0.3294	0.2662
	LMNN	0.1818	0.3257	0.3673	0.4471	0.5246	0.3693	0.1333	0.3128	0.3797	0.4516	0.4037	0.3362
	TSML*	**0.0957**	0.2057	0.2687	0.3274	0.4469	0.2689	0.1098	0.1892	0.2300	0.2830	0.3513	0.2327
MT	mtSCML	0.2301	0.3053	0.3798	0.4981	0.5147	0.3850	0.1913	0.2799	0.3117	0.3592	0.3478	0.2980
	mtLMNN	0.2137	0.2681	0.3270	0.3972	0.3425	0.3097	0.1470	0.2358	0.2935	0.3247	0.2560	0.2514
	CP-mtML	0.1317	0.2075	0.3823	0.3637	0.3336	0.2838	0.1098	0.2092	0.2965	0.3656	0.3234	0.2609
	mtMLCS	0.1777	0.2500	0.2800	0.3917	0.3734	0.2946	0.1355	0.2325	0.2379	0.3410	0.3237	0.2541
	mtTSML*	0.1011	**0.2018**	**0.2367**	**0.2976**	**0.2536**	**0.2182**	**0.0962**	**0.1766**	**0.2170**	**0.2652**	**0.2093**	**0.1929**

TABLE III: Performance comparison on the SOF dataset in terms of MSE.

		SOF (20% as training data)						SOF (40% as training data)					
		Task1	Task2	Task3	Task4	Task5	Avg	Task1	Task2	Task3	Task4	Task5	Avg
ST	Euclidean	0.5951	0.6154	0.6409	0.8671	0.8031	0.7043	0.5417	0.5576	0.5207	0.6067	0.7256	0.5905
	Cosine	0.5552	0.5262	0.5666	0.8392	0.7600	0.6494	0.5691	0.5664	0.5786	0.5926	0.7392	0.6092
	GMML	0.4417	0.4369	0.4737	0.4685	0.5138	0.4669	0.4398	0.4424	0.4332	0.5730	0.5023	0.4782
	SCML	0.3558	0.4185	0.3560	0.5315	0.4677	0.4259	0.3333	0.3410	0.3628	0.6279	0.5000	0.4330
	LowRank	0.4847	0.4954	0.4892	0.5455	0.5846	0.5199	0.4352	0.4286	0.4240	0.6629	0.5116	0.4925
	ITML	0.4417	0.4185	0.5232	0.4825	0.5446	0.4821	0.4306	0.4147	0.4747	0.5281	0.6279	0.4952
	LMNN	0.3834	0.4369	0.4582	0.6503	0.7108	0.5279	0.3796	0.3318	0.3963	0.5506	0.4884	0.4293
	TSML	0.3282	0.3631	0.4180	0.4615	0.4462	0.4034	0.3287	0.3180	**0.3410**	**0.3708**	0.4558	0.3629
MT	mtSCML	0.3037	0.3846	0.3467	0.5245	0.4185	0.3944	0.2824	0.3226	0.3721	0.5814	0.4393	0.3995
	mtLMNN	0.3804	0.3938	0.4334	0.5245	0.5108	0.4486	0.3426	0.3410	0.4055	0.4270	0.4698	0.3972
	CP-mtML	0.3589	0.3538	0.3994	0.4895	0.4554	0.4114	0.3056	0.3456	0.3548	0.4270	0.4465	0.3759
	mtMLCS	0.3776	0.4055	0.4022	0.4814	0.5196	0.4372	0.3102	0.3088	0.3502	0.3933	0.4651	0.3655
	mtTSML	**0.3067**	**0.3415**	**0.3407**	**0.3497**	**0.4062**	**0.3490**	**0.2824**	**0.2811**	**0.3410**	0.3820	**0.4279**	**0.3429**

setting where all the tasks only share the common projection without the task-specific parts is not applicable here.

3) Proposed Approaches: mtTSML is our proposed multi-task metric learning method with sparse feature selection and triplet constraints. Through removing the impact from the common part, we can obtain the corresponding single-task method TSML. TSML shares some similarities with LowRank [7], as they both utilize $\ell_{2,1}$ norm to perform sparse feature selection. However, LowRank learns the bilinear similarity by optimizing the logistic loss for binary classification, whereas TSML learns the Mahalanobis distance using the hinge loss on triplet constraints. Moreover, LowRank is not designed to incorporate the fine-grained similarity degree information.

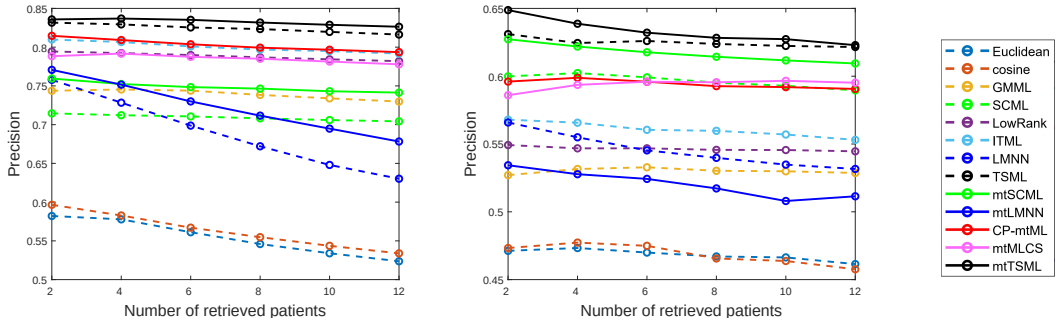
4) Implementation Details & Measurement: KNN classifier is adopted to evaluate the performance on disease diagnosis and prognosis. Since the class labels are ordinal, we use mean squared error (MSE) to measure the classification results. The smaller the MSE values, the better the results. For example, if the true label is "0", predicting it to be "1" can be regarded relatively better than predicting it to be "2". We set $k = 3$ in KNN for the comparison of all the metric learning methods. In the following experiments, the MSE results are averaged over 5 random trials, i.e., we randomly split the dataset and conduct experiments five times. For all the tasks, the training, validation and testing data have no overlapped patient samples.

B. Experimental Results

1) Disease Prediction Results: Table II shows the experimental results on the ADNI dataset. Table III shows the experimental results on the SOF dataset. We conduct experiments using 20% and 40% portions of the whole dataset as the training data, respectively.

In the tables, we list the MSE results obtained from a KNN classifier for all the tasks and the averaged results. Since in our multi-task learning setting, all the tasks in one dataset are optimized simultaneously, and there is no main/auxiliary task, we care more about the averaged MSE than that of a specific task. The averaged measurement reflects the general performance during the disease progression process.

From the comparison with Euclidean and cosine distance, we can see that the supervised metric learning methods can better capture the statistical regularity from the original data. Among the single-task learning methods, TSML achieves the best results, as it can better capture the characteristics from the dataset. The triplet-based distance constraints provide more similarity information about the progression stages of the disease, while other methods ignore the label relationships. The low-rank formulation with sparse feature selection ensures that the transformed space is low dimensional, and also reduces the noisy information existing in the high dimensional inputs. However, since different metrics are learned separately, the relationships among tasks are totally ignored, resulting in the overfitting issue. Therefore, single-task learning methods cannot obtain good performance for each of the tasks.



(a) ADNI **(b)** SOF

Fig. 2: Patient retrieval results using different metrics on ADNI and SOF datasets using 20% as training data. X-axis is the number of retrieved patients, and Y-axis is the average precision of each method.

As shown in Table II and Table III, the overall performance of multi-task learning methods is better than that of single-task learning methods. Specially, the multi-task methods significantly improve the results compared to their corresponding single-task versions (e.g. mt-SCML vs. SCML, mt-LMNN vs. LMNN, and mtTSML vs. TSML). Multi-task learning captures the information from multiple tasks simultaneously, and transfers useful information among different tasks. The shared information benefits each individual task, especially the weakly trained tasks. On the ADNI dataset, for example, Task 5 has limited number of training samples due to missing labels. We can see that the single-task learning methods cannot perform well using such small amount of training data samples. However, the multi-task learning methods can significantly improve the classification accuracy. This owes to the fact that multi-task learning transfers shared information which helps for training Task 5. Similar observations can also be found in Task 4 on the SOF dataset.

Among the multi-task metric learning methods, the proposed method mtTSML achieves the best results. mtLMNN simultaneously optimizes the common Mahalanobis matrix and task-specific one as described in Section II-A. However, it cannot reduce the rank of the Mahalanobis matrix. CP-mtML and mt-MLCS are the two methods that try to transform the original data into a low-dimensional space, and hence they have some capability of handling data with high dimensions. mtMLCS transforms all the tasks to a common subspace by sharing a common projection matrix, but it may underestimate the heterogeneous information of different tasks. CP-mtML has both common projection and task-specific ones, without feature selection and the distance constraints for similarity degree of the labels. mtSCML regards the Mahalanobis matrix as a weighted sum of several rank-1 matrices and is computationally efficient. However, the matrices cannot be optimized jointly with the combination weights, which may reduce the generalization ability of the model.

Results obtained using 40% data as training are generally better than those of using 20%. As the training size increases, the models are able to capture more statistical information. However, it is usually hard to obtain sufficient amount of healthcare samples with labels in practice. Therefore, the ability of learning from small amount of data samples is important. In Table III, the MSE values of each method are higher than those in Table II. This is due to the fact that the features on the SOF dataset are very noisy and not informative enough. The features on the SOF dataset are mostly risk factors which implicitly affect the disease, while those on the ADNI dataset are mostly biomarkers. Nevertheless, the proposed method mtTSML achieves the best performance.

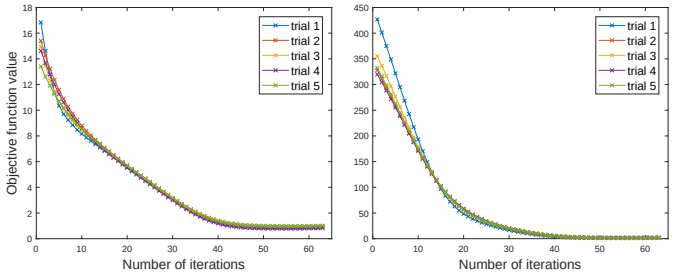
2) Patient Retrieval: Identifying similar patients is an important practice in the healthcare domain. We perform the task using metric learning to retrieve patients with similar health conditions. Given a query from the testing pool at some future time point, we retrieve the top k most similar patients from the training set. Then we compare the disease status labels of retrieved samples at the time point with the testing patient, and obtain the precision which is the percentage of the correctly retrieved patients among k patients. We repeat the query process for all the tasks and calculate the average precision. Note that in the query, we only compare whether the retrieved patients have the same label as the testing patient, without considering the ordered label relationships in the measurement.

Fig. 2 illustrates the comparison of different methods on precision@k trained using 20% data on these two datasets. We can see that the proposed method mtTSML outperforms all the other baselines on both datasets. Also, multi-task learning methods perform generally better than single-task learning methods.

C. Experimental Analysis

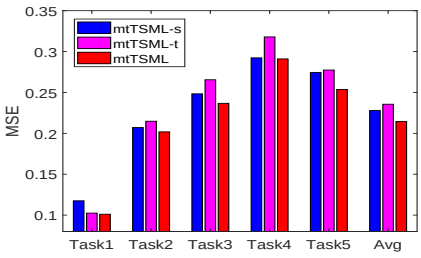
1) Convergence: We empirically show that our method converges to a sub-optimal solution. Fig. 3(a) and Fig 3(b) show the variation of objective function values with respect to the number of iterations on the two datasets, respectively. We conduct the experiments five times (i.e., trial 1, trial 2, etc). At each time, there are 20% patients randomly selected as the training data. From the two figures, we can see that for both of the two datasets, the values of the objective function gradually converge with the increase of the number of iterations.

2) Reduced Model Comparison: To validate the importance of sparse feature selection and proposed triplet constraints,

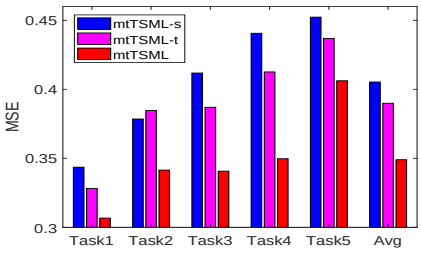


(a) ADNI **(b)** SOF

Fig. 3: Illustration of convergence, i.e., objective function values with respect to iterations on two datasets.



(a) ADNI



(b) SOF

Fig. 4: Comparison of MSE of mtTSML-s, mtTSML-t and mtTSML on ADNI and SOF datasets. The smaller the MSE values, the better the results.

we conduct experiments with two reduced models separately. We remove the sparse regularization term $\|\mathbf{L}_t\|_{2,1}$ in Eq. (8), and denote the reduced method as mtTSML-s. We remove $\mathcal{R}^1$ and $\mathcal{R}^2$ sets and keep only the $\mathcal{R}^3$ set when constructing distance constraints (see Section II-B3) for Eq. (8), and denote the reduced method as mtTSML-t. The MSE results on two datasets (20% as training data) are shown in Fig. 4. From the two figures, we can see that the proposed mtTSML performs better than the reduced versions mtTSML-s and mtTSML-t. This demonstrates the effectiveness of considering feature sparsity for parameter regularization, and the relative similarity of labels (reflected in $\mathcal{R}^1$ and $\mathcal{R}^2$) for distance constraints.

3) Sparse Feature Selection: To indicate that the proposed method mtTSML is able to select informative features from the high dimensional feature space for each individual task, we use color map to illustrate the learned transformation matrix $\mathbf{L}_t$. We append 100 dimension Gaussian noises on the SOF dataset to expand its feature dimension to 300. The model is trained using 20% portion of the data.

Fig. 5 shows the color map of the transformation matrix $\mathbf{L}_t$ from Task 1 to Task 5, respectively, on the SOF dataset with noise. The horizontal axis is the feature dimension, and the vertical axis is the dimension of transformed space. We can see that the last 100 columns which correspond to the noisy features are set to zero vectors by the model. At the same time, a number of columns have zero values for original features, which indicates that the original dataset itself is noisy and contains many redundant features. Among the selected informative features, the left columns are assigned large weights in all the matrices. In fact, these features are the dual x-ray absorptiometry measurement values at the baseline visit, which are clinically important for bone disease diagnosis.

4) Disease Progression Visualization: Multi-task metric learning provides us a way to analyze and visualize patient disease progression. Fig. 6 shows the neighbor distributions of a given sample patient over time. In each timestamp, we use the nearest 15 neighbor patients. Multidimensional Scaling (MDS) [32] is used to visualize the relative distances. We can see that in the first month, the anchor patient is surrounded by neighbors with MCI. Twelve months later, although the patient is still correctly predicted as MCI by KNN, more impostors labeled AD become nearer to him/her, indicating the risk of disease deterioration. In the 24th month, the patient becomes more similar to AD patients and diagnosed as AD, which means that his/her disease stage becomes worse. As the Alzheimer's disease is neurodegenerative, the patients are more and more close to ADs later on.

On the contrary, Fig. 7 illustrates the surroundings of another patient diagnosed as MCI. From its neighbors, we can see that the situation of this patient is better than the first one, as there are normal patients in his/her top-15 nearest neighbors. As expected, the status of this MCI patient remains stable later on.

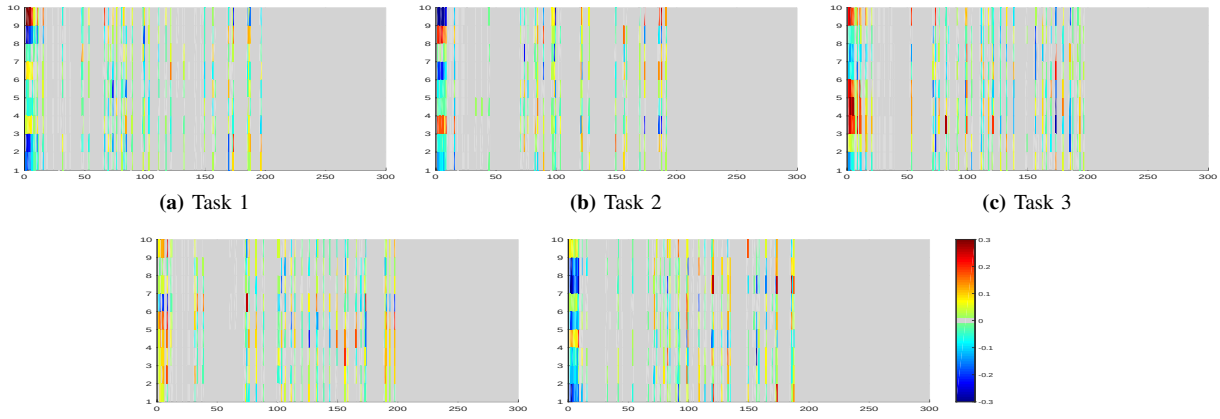
IV. Related Work

A. Patient Similarity Learning

Measuring and identifying similar patients is a crucial component for clinical decision support. LSML proposed in [8] is a locally supervised metric learning method which incorporates the physician feedback as the supervision. Zhang et al. [1] propose a label propagation method to learn patient similarity and drug similarity jointly. Wang et al. [33] propose a weakly supervised patient similarity learning by using small amount of labels. Sun et al. [34] use both statistical and wavelet based features to capture the characteristics of patients. Zhan et al. [7] learns the bilinear similarity while perform feature selection. Considering the longitudinal records, [9] uses convolution neural network and [5] uses modified recurrent neural network to learn nonlinear representation for the original inputs.

B. Disease Progression Modeling

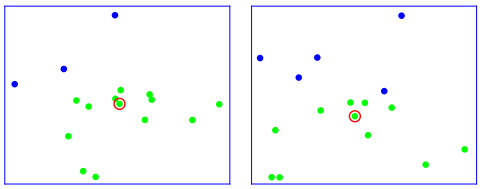
Monitoring the progression stages of a specific disease helps to promote disease control and prevention. Models based on probabilistic models [35] and neural networks [36–39] are



(a) Task 1 **(b)** Task 2 **(c)** Task 3

(d) Task 4 **(e)** Task 5

Fig. 5: Illustration of the transformation matrices. Elements colored gray are zeros. Other colors indicate the positive and negative values. X-axis is the feature dimension and Y-axis is the dimension of the transformed low-rank space.



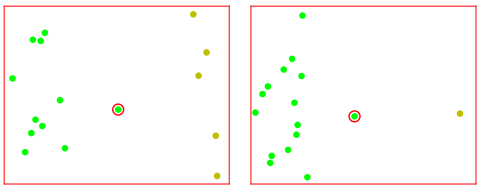
(a) Task 1: M00 **(b)** Task 2: M12

(c) Task 3: M24 **(d)** Task 4: M36

Fig. 6: Visualization of disease progression for one patient on the ADNI dataset. The red circle highlights the given anchor patient. Green dots indicate samples labeled as "MCI", and blue dots are samples labeled as "AD". M00, M12, M24 and M36 denote the baseline time, 12 month, 24 month and 36 month after baseline time respectively.

proposed to detect the occurrence of disease or the risk of potential disease. To monitor the disease stages over time, models of multi-task formulations are developed. [29, 30, 40] employ multi-task learning on regression model with fused group lasso to predict cognitive scores in future timestamps for Alzheimer's disease and Parkinson's disease, respectively. [41, 42] develop multi-task survival models to predict the survival/transition time. The above work has a similar problem setting as our work: features measured at the baseline time are used as the input, and each task is performed at one future time point.

The above multi-task methods provide a way to explicitly predict the values of cognitive score or disease transition status. Differently, we focus on learning the variation of patient



(a) Task 1: M00 **(b)** Task 4: M36

Fig. 7: Visualization of disease progression for one patient on the ADNI dataset. The red circle highlights the given anchor patient. Green dots indicate samples labeled as "MCI", and yellow dots are samples labeled as "NL".

similarity over time through multi-task metric learning. Besides performing the classification task on predicting disease labels of patients, metric learning can also be used to retrieve similar patients and visualize patient cohort distributions.

V. Conclusion

In this paper, we proposed a multi-task metric learning method to measure patient similarity at multiple future time points of interest on two real-world healthcare datasets. Specifically, we aim to resolve two challenges caused by the uniqueness of healthcare data when performing metric learning: (1) the high-dimension and noisy nature of the data collected from real world systems, and (2) the clinical relationships among disease labels. To remove the noisy information, feature selection is exploited in the proposed model. We first decompose the Mahalanobis distance for each task to a common and task-specific part through low-rank transformation matrices, and then perform sparse feature selection using $\ell_{2,1}$ norm for each task. In such a way, non-informative features can be set to zero while discriminative features can be selected for each task. Triplet constraints are further designed to incorporate the information of class labels. The constraints force the patients with similar labels to get closer and less similar patients to move far away by a fixed margin. Finally, experimental

results on two real world healthcare datasets demonstrate the effectiveness of the proposed model.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their valuable comments and helpful suggestions. This work was supported in part by the US National Science Foundation under grants NSF IIS-1218393 and IIS-1514204. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] P. Zhang, F. Wang, J. Hu, and R. Sorrentino, "Towards personalized medicine: leveraging patient similarity and drug similarity analytics," *AMIA Summits on Translational Science Proceedings*, 2014.

[2] J. Lee, D. M. Maslove, and J. A. Dubin, "Personalized mortality prediction driven by electronic medical data and a patient similarity metric," *PloS one'15*.

[3] A. Gottlieb, G. Y. Stein, E. Ruppin, R. B. Altman, and R. Sharan, "A method for inferring medical diagnoses from patient similarities," *BMC medicine*, 2013.

[4] S. Ebadollahi, J. Sun, D. Gotz, J. Hu, D. Sow, and C. Neti, "Predicting patients trajectory of physiological data using temporal trends in similar patients: A system for near-term prognostics," in *AMIA annual symposium proceedings*, 2010.

[5] C. Che, C. Xiao, J. Liang, B. Jin, J. Zho, and F. Wang, "An rnn architecture with dynamic temporal matching for personalized predictions of parkinson's disease," in *SDM'17*. SIAM.

[6] A. Bellet, A. Habrard, and M. Sebban, "A survey on metric learning for feature vectors and structured data," *arXiv:1306.6709*, 2013.

[7] M. Zhan, S. Cao, B. Qian, S. Chang, and J. Wei, "Low-rank sparse feature selection for patient similarity learning," in *Data Mining (ICDM), 2016 IEEE International Conference on*.

[8] J. Sun, F. Wang, J. Hu, and S. Edabollahi, "Supervised patient similarity measure of heterogeneous patient records," *ACM SIGKDD Explorations Newsletter*, 2012.

[9] Z. Zhu, C. Yin, B. Qian, Y. Cheng, J. Wei, and F. Wang, "Measuring patient similarities via a deep architecture with medical concept embedding," in *Data Mining (ICDM), 2016 IEEE International Conference on*.

[10] M. Huai, C. Miao, Q. Suo, Y. Li, J. Gao, and A. Zhang, "Uncorrelated patient similarity learning," in *SDM*. SIAM, 2018.

[11] J. Ni, J. Liu, C. Zhang, D. Ye, and Z. Ma, "Fine-grained patient similarity measuring using deep metric learning," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2017.

[12] Q. Suo, F. Ma, Y. Yuan, M. Huai, W. Zhong, J. Gao, and A. Zhang, "Deep patient similarity learning for personalized healthcare," *IEEE Transactions on NanoBioscience*, 2018.

[13] S. Parameswaran and K. Q. Weinberger, "Large margin multi-task metric learning," in *Advances in neural information processing systems*, 2010.

[14] Y. Zheng, J. Fan, J. Zhang, and X. Gao, "Hierarchical learning of multi-task sparse metrics for large-scale image classification," *Pattern Recognition*, 2017.

[15] Y. Zhang and D.-Y. Yeung, "Transfer metric learning by learning task relationships," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010.

[16] L. Ma, X. Yang, and D. Tao, "Person re-identification over camera networks using multi-task distance metric learning," *IEEE Transactions on Image Processing*, 2014.

[17] J. Vogt and V. Roth, "A complete analysis of the l_1, p group-lasso," *arXiv:1206.4632*, 2012.

[18] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, 2009.

[19] P. Zadeh, R. Hosseini, and S. Sra, "Geometric mean metric learning," in *International Conference on Machine Learning*, 2016.

[20] 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.

[21] Y. Shi, A. Bellet, and F. Sha, "Sparse compositional metric learning." in *AAAI*, 2014.

[22] P. Yang, K. Huang, and C.-L. Liu, "Geometry preserving multi-task metric learning," *Machine learning*, 2013.

[23] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012.

[24] P. Yang, K. Huang, and C.-L. Liu, "A multi-task framework for metric learning with common subspace," *Neural Computing and Applications*, 2013.

[25] B. Bhattarai, G. Sharma, and F. Jurie, "Cp-mtml: Coupled projection multi-task metric learning for large scale face retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[26] M. Huai, C. Miao, Y. Li, Q. Suo, L. Su, and A. Zhang, "Metric learning from probabilistic labels," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018.

[27] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine learning*, 2011.

[28] M. Kowalski, "Sparse regression using mixed norms," *Applied and Computational Harmonic Analysis*, 2009.

[29] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via fused sparse group lasso," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012.

[30] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A multi-task learning formulation for predicting disease progression," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011.

[31] J. Compston, C. Cooper, and J. Kanis, "Bone densitometry in clinical practice." *BMJ: British Medical Journal*, 1995.

[32] G. P. Borg, I., "Modern multidimensional scaling: Theory and applications," 2005.

[33] F. Wang and J. Sun, "Psf: a unified patient similarity evaluation framework through metric learning with weak supervision," *IEEE journal of biomedical and health informatics*, 2015.

[34] F. Wang, J. Sun, and S. Ebadollahi, "Integrating distance metrics learned from multiple experts and its application in patient similarity assessment," in *SDM'11*.

[35] X. Wang, D. Sontag, and F. Wang, "Unsupervised learning of disease progression models," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014.

[36] Q. Suo, F. Ma, G. Canino, J. Gao, A. Zhang, P. Veltri, and G. Agostino, "A multi-task framework for monitoring health conditions via attention-based recurrent neural networks," in *AMIA annual symposium proceedings*. American Medical Informatics Association, 2017.

[37] Y. Yuan, G. Xun, Q. Suo, K. Jia, and A. Zhang, "Wave2vec: Learning deep representations for biosignals," in *Data Mining (ICDM), 2017 IEEE International Conference on*.

[38] Y. Yuan, X. Guangxu, F. Ma, Y. Wang, N. Du, K. Jia, L. Su, and A. Zhang, "Muvan: A multi-view attention network for multivariate temporal data," in *Data Mining (ICDM), 2018 IEEE International Conference on*.

[39] F. Ma, J. Gao, Q. Suo, Q. You, J. Zhou, and A. Zhang, "Risk prediction on electronic health records with prior medical knowledge," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018.

[40] S. Emrani, A. McGuirk, and W. Xiao, "Prognosis and diagnosis of parkinson's disease using multi-task learning," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.

[41] Y. Li, J. Wang, J. Ye, and C. K. Reddy, "A multi-task learning formulation for survival analysis," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[42] Y. Li, T. Yang, J. Zhou, and J. Ye, "Multi-task learning based survival analysis for predicting alzheimer's disease progression with multi-source block-wise missing data," in *SDM*. SIAM, 2018.