Uncorrelated Patient Similarity Learning

Mengdi Huai* Chenglin Miao* Qiuling Suo* Yaliang Li† Jing Gao* Aidong Zhang*

Abstract

Patient similarity learning aims to derive a clinically meaningful similarity metric to measure the similarity between a pair of patients according to their historical clinical information, which could help to predict the clinical outcomes of the patient of interest. However, the patient clinical data are usually complex, and contain much irrelevant and redundant information, which makes it difficult to learn the similarity metric with high accuracy. Although some methods have been proposed to address the complex nature of patient data, they overemphasize sparsity-based relevant feature selection and fail to take into consideration the redundant features that are highly correlated with each other, and this heavily degrades the accuracy of the learned results. To address the above challenges, we propose a novel uncorrelated patient similarity learning approach, which can not only select the most relevant features for the learning task, but also guarantee that the selected features have low correlations with each other. Additionally, to address the scenarios where the patient data are distributed across different sites, we extend the proposed approach and design a distributed mechanism, based on which the similarity metric can be accurately learned without directly accessing the raw patient data at each site. The desirable performance of the proposed methods are verified through extensive experiments conducted on both real-world and synthetic datasets.

Keywords: Patient similarity, Feature selection, Distributed similarity learning

1 Introduction

With the prevalence of the adoption of Electronic Health Records (EHRs), various clinical information are becoming available for a large number of patients. These wealth clinical information make it possible to perform patient similarity analysis, which is a fundamental problem in healthcare informatics. The goal of patient similarity learning is to measure the similarity between a pair of patients according to their clinical information, which could help to retrieve similar reference cases for predicting the clinical outcome of interest. As patient similarity learning is capable of improving clinical decision making without incurring additional effort from physicians, it has been widely used in various applications, such as target patient retrieval [20], medical prognosis [22] and clinical pathway analysis [9].

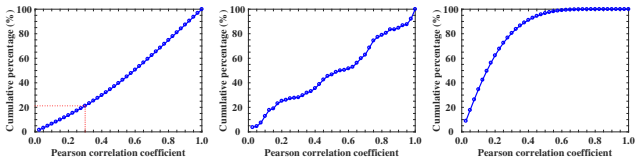
The key part of patient similarity learning is to learn a clinically meaningful and precise similarity metric that can be used to measure the similarity between a pair of patients. However, since the patient data collected from real-world applications are usually high dimensional, complex and noisy, the features extracted from these data to characterize patient profiles may contain much irrelevant and redundant information, which can hide the relationship between the learning task and the most relevant features. Thus, it is essential to remove such irrelevant and redundant information when conducting patient similarity learning so that an accurate distance metric can be learned based on the extracted features.

To address the complex nature of patient data, some sparse feature selection methods [28, 16, 8, 26, 15, 17] have been proposed to select a subset of relevant features that are highly correlated with the learning task, but these methods assume that the input features are nearly independent, and they ignore the correlations among the selected features. However, in reality, most features in patient data exhibit strong correlations, which can be seen from the example shown in Fig. 1. Those correlated features may share similar properties and thus reveal overlapped or redundant information, which makes the knowledge discovery process much difficult [11]. Especially when the number of selected features is very limited, it is desirable to remove the overlapped and redundant features so that more discriminative information can be extracted from the data to conduct the patient similarity learning task.

Towards this end, we propose a novel Uncorrelated Patient Similarity Learning approach(UnPSL), which can not only select the most relevant features for the learning task, but also guarantee that the selected features have low correlations with each other. More specifically, we first formulate the patient similarity learning problem as a maximum likelihood estimation problem, and then introduce two regularization terms to control the feature selection process such that only the most relevant and uncorrelated features can be selected. The advantages of the proposed approach are threefold:

*State University of New York at Buffalo, {mengdihu, cmiao, qiulings, jing, azhang}@buffalo.edu.

†Baidu Research Big Data Lab, yaliangli@baidu.com.



(a) (b) (c)

Figure 1: Feature correlations on real-world patient datasets. (a): Colon cancer dataset [1]. (b): Parkinson's disease dataset [14]. (c): Leukemia dataset [7]. The horizontal axis represents the absolute pearson correlation coefficient, and the vertical axis represents the cumulative percentage of the feature pairs whose absolute pearson correlation coefficients are less than or equal to a specific value, e.g., for the Colon cancer dataset, the cumulative percentage of the feature pairs whose pearson correlation coefficients are less than or equal to 0.3 is only around 20%, which means that the remaining 80% feature pairs are highly correlated.

Firstly, it increases the interpretability of the learned models, as the selected features are discriminative and can well represent the feature space covered by the entire dataset. Secondly, it is more robust to the noisy and redundant information, and the learned result can be achieved with high accuracy. Last but not least, the learning process in the proposed approach is not dependent on any particular data, and it can be generally used to other applications, e.g., multi-class regression.

Additionally, we take into consideration the scenarios where the patient data are distributed across different sites (e.g., hospitals, medical centers). Due to various reasons such as the concern of privacy leakage, the consumption of time or other resources when uploading data, these sites usually are not willing to provide the raw patient data to the learner who aims to learn the clinical similarity metric from the collected data. In such cases, it is difficult for the learner to achieve satisfactory accuracy based on the above proposed approach. To address this challenge, we extend the aforementioned learning approach and propose a novel distributed patient similarity learning mechanism, based on which the similarity metric can be accurately learned without directly accessing the raw patient data at each site.

The main contributions of this paper are summarized as follows. Firstly, we propose a novel patient similarity learning method which can select the most relevant and uncorrelated features so that the learned metric can be more accurate. Secondly, we are the first to address the patient similarity learning problem in the distributed scenarios, and we propose a novel distributed learning mechanism, based on which the similarity metric can be accurately learned without directly accessing the raw patient data at each site. Thirdly, extensive experiments based on both real-world and synthetic datasets are conducted to verify the desirable performance of the proposed approaches.

2 Problem Formulation

In this section, we describe the problem setting of this paper. Suppose there are n patients and the data samples of them are denoted as $\mathcal{X} = \{x_i \in \mathcal{R}^d\}_{i=1}^n$, where $x_i \in \mathcal{R}^d$ is the data sample of the i-th patient and it is a d-dimensional vector (i.e., contains d features). For each pair of samples (x_i, x_j), we assume that a label y_{ij} is given and it denotes whether the two samples are similar (e.g., the two patients belong to the same cohort or have the same disease) or not. If x_i and x_j are similar, y_{ij} is equal to 1, otherwise it is equal to -1. We use $\mathcal{I}_s = \{(x_i, x_j) : y_{ij} = 1\}$ to denote the set of sample pairs whose labels are equal to 1 and $\mathcal{I}_d = \{(x_i, x_j) : y_{ij} = -1\}$ to denote the set of pairs whose labels are equal to -1.

Given the two sets of sample pairs (i.e., $\mathcal{I}_s$ and $\mathcal{I}_d$), our goal in this paper is to learn a similarity function

$$s(x_i, x_j) = x_i^T M x_j = x_i^T W^T W x_j = (W x_i)^T W x_j, \tag{2.1}$$

which can measure the similarity between any two inputs (samples) x_i and x_j. Here, $s(x_i, x_j)$ is parameterized by a positive semidefinite matrix M, which can be decomposed as $M = W^T W (W \in \mathcal{R}^{d*d})$ and W is the similarity metric that needs to be learned in a supervised manner.

3 Methodology

In this section, we provide the details of the proposed patient similarity learning approach. We first formulate the patient similarity learning process as an optimization problem, and then discuss how to effectively solve this optimization problem through adopting the alternating direction method of multipliers (ADMM) [2].

3.1 Learning Framework. In our proposed approach, the similarity learning process is formulated as an optimization problem. Specifically, we first model the learning process based on maximum likelihood method, i.e., estimating the parameters which maximize the likelihood of the data samples in the training dataset. Then, we introduce a term to select the most relevant features through conducting sparse feature selection. Last but not least, we introduce another term which can effectively reduce the correlations among the selected features. The details are described as follows.

Maximum likelihood estimation. As described in Section 2, our goal in this paper is to learn the similarity function $s(x_i, x_j)$, which is parameterized by W, based on the given two sets of sample pairs (i.e.,

$\mathcal{I}_s$ and $\mathcal{I}_d$). To achieve the goal, we adopt maximum likelihood estimation here. That is to say, we need to find a matrix W which can maximize the likelihood of the given sample pairs in the training dataset (i.e., $\mathcal{I}_s$ and $\mathcal{I}_d$). We model the probability for each sample pair (x_i, x_j) and the corresponding label y_{ij} as

$$\Pr(y_{ij}|x_i, x_j; M, b) = \frac{1}{(1 + exp(-y_{ij}(s(x_i, x_j) - b)))}, \tag{3.2}$$

where $y_{ij} \in \{-1, 1\}$ and b is the bias, which also works as a threshold. The two patients x_i and x_j are treated as similar (i.e., $y_{ij} = 1$) only when the similarity measure $s(x_i, x_j)$ is greater than or equal to b, otherwise they are treated as dissimilar (i.e., $y_{ij} = -1$). In this paper, we set b as 1 by following existing works. Since Eqn. (3.2) is capable of providing the probability that a pair of patients are similar or not, it is particularly suitable for disease prediction and decision making in medical diagnosis and prognosis. Then the log likelihood of the sample pairs in the two sets $\mathcal{I}_s$ and $\mathcal{I}_d$ is:

$$\begin{aligned}\mathcal{L}(W, b) &= \log \Pr(\mathcal{I}_s) + \log \Pr(\mathcal{I}_d) \\ &= - \sum_{(x_i, x_j) \in \mathcal{I}_s} \log(1 + \exp(-(s(x_i, x_j) - b))) \\ &\quad - \sum_{(x_i, x_j) \in \mathcal{I}_d} \log(1 + \exp((s(x_i, x_j) - b))),\end{aligned} \tag{3.3}$$

where $s(x_i, x_j) = (Wx_i)^T Wx_j$. Thus, the problem of the patient similarity learning can be transformed to the maximum likelihood optimization problem. Since maximizing the log likelihood is equivalent to minimizing the negative log likelihood, we can get the following optimization problem

$$\begin{aligned}\min_{W \in \mathcal{R}^{d*d}} \ f_1(W, b) &= -\log \Pr(\mathcal{I}_s) - \log \Pr(\mathcal{I}_d) \\ &= \sum_{(x_i, x_j) \in \mathcal{I}_s} \log(1 + \exp(-(s(x_i, x_j) - b))) \\ &\quad + \sum_{(x_i, x_j) \in \mathcal{I}_d} \log(1 + \exp((s(x_i, x_j) - b)))\end{aligned} \tag{3.4}$$

Sparse feature selection. Clinical data collected from real-world applications are usually high-dimensional, and contain many irrelevant features, which makes the patient similarity learning process very difficult. Thus, it is essential to design a feature selection method such that the most relevant features can be selected when conducting patient similarity learning.

To achieve the goal, we conduct sparse feature selection during the learning process. Suppose w_i is the i-th row vector of W, i.e., $W = [w_1, w_2, ..., w_d]^T$. Then w_i can be regarded as a vector that measures the importance of the i-th feature in the data samples. For the purpose of selecting the most relevant information, we expect that only a few number of w_i are non-zero and the selected features are enough to embed the original data to its low dimensional representation. When we adopt l_2-norm of w_i as a metric to measure its contribution, the following optimization problem should be taken into consideration during the learning process.

$$\min_{W \in \mathcal{R}^{d*d}} \ \sum_{i=1}^{d} \|w_i\|_2 = \sum_{i=1}^{d} (\sum_{j=1}^{d} W_{ij}^2)^{1/2} \tag{3.5}$$

This optimization framework will enforce some rows of W to be all zero, and the corresponding features in the data samples will not be selected.

Uncorrelated feature selection. Although sparse feature selection can remove much irrelevant information from the data samples, it does not take feature correlation into account. In reality, the features in patients' data are often highly correlated, and these correlated features may share similar properties and contain redundant information. It is obvious that the accuracy of the learned results will be degraded if all the correlated features are adopted at the same time, because the features that are highly correlated with others can not provide much more information for the learning process. Thus, it is desirable that the selected features are uncorrelated as much as possible (especially when the number of selected features is limited) such that more information can be extracted from the data samples when conducting patient similarity learning.

In order to select the uncorrelated features, we first normalize the data samples $\mathcal{X} \in \mathcal{R}^{n*d}$ such that the values of each feature in the data have zero-mean and unit-variance. Then we calculate the correlation coefficient matrix $R = (r_{kl}) \in [-1, 1]^{d*d}$ for the d features. Here r_{kl} is the correlation coefficient between the k-th and l-th features and it is calculated as

$$r_{kl} = \frac{\sum_{i=1}^{n} x_{ki} x_{li}}{\sqrt{\sum_{i=1}^{n} x_{ki}^2} \sqrt{\sum_{i=1}^{n} x_{li}^2}}. \tag{3.6}$$

To select features which are uncorrelated as much as possible, we should let the similarity metric W satisfy

$$\min_{W \in \mathcal{R}^{d*d}} \ Trace((W \cdot W^T)(R \odot R)^T), \tag{3.7}$$

where $\odot$ is Hadamard product of matrices. In Eqn.(3.7), we use the square of the correlation coefficient matrix (i.e., $(R \odot R)^T$) instead of the original correlation coefficient matrix (i.e., R) to eliminate the effect of anti-correlation. The objective function in Eqn. (3.7) guarantees that when the k-th feature and the l-th feature are highly correlated, i.e., the value of $(r_{kl})^2$ is large, the two features are not treated as equally important and they can not be selected simultaneously

during the feature selection process. Thus, the goal of selecting as many uncorrelated features as possible can be achieved.

The learning framework. Taking both sparse and uncorrelated feature selection into consideration, we formalize the patient similarity learning problem as the following optimization problem

$$\min_{W\in\mathcal{R}^{d*d}} \mathcal{L}_1(W) = \sum_{(x_i,x_j)\in\mathcal{I}_s} \log(1+\exp(-(x_i^T W^T W x_j - b))) + \sum_{(x_i,x_j)\in\mathcal{I}_d} \log(1+\exp(x_i^T W^T W x_j - b))) + \lambda_1 \sum_{i=1}^{d} \|w_i\|_2 + \lambda_2 Trace((W\cdot W^T)(R\odot R)^T), \tag{3.8}$$

where λ_1,$\lambda_2 \geq 0$ are the tuning parameters, and the correlation coefficient matrix R is positive semidefinite. Next, we will discuss how to solve this optimization problem to learn the similarity metric W.

3.2 Optimization. Firstly, we transform the optimization problem into the following equivalent problem:

$$\begin{aligned}\min_{W,V,Z} \mathcal{L}_2(W) = &\sum_{(x_i,x_j)\in\mathcal{I}_s} \log(1+\exp(-(x_i^T W^T W x_j - b))) \\ &+ \sum_{(x_i,x_j)\in\mathcal{I}_d} \log(1+\exp(x_i^T W^T W x_j - b))) \\ &+ \lambda_1 \sum_{i=1}^{d} \|V_i\|_2 + \lambda_2 Trace((Z\cdot Z^T)(R\odot R)^T) \\ s.t.\quad & W = V = Z.\end{aligned} \tag{3.9}$$

Since $W = V$ and $V = Z$, the above problem is equivalent to problem (3.8). Here we adopt the ADMM algorithm to solve Eqn.(3.9). Through introducing the Lagrange multipliers $\Delta_V \in \mathcal{R}^{d*d}, \Delta_W \in \mathcal{R}^{d*d}$, we get the Lagrange form of the above optimization problem:

$$\begin{aligned}&\mathcal{L}_\rho(W,V,Z,\Delta_W,\Delta_V) = \\ &\sum_{(x_i,x_j)\in\mathcal{I}_s} \log(1+\exp(-(x_i^T W^T W x_j - b))) \\ &+ \sum_{(x_i,x_j)\in\mathcal{I}_d} \log(1+\exp(x_i^T W^T W x_j - b))) + \lambda_1 \sum_{i=1}^{d} \|V_i\|_2 \\ &+ \lambda_2 Trace((Z\cdot Z^T)(R\odot R)^T) + < \Delta_W, W - Z >_F \\ &+ \frac{\rho}{2}\|W-Z\|_F^2 + < \Delta_V, V - Z >_F + \frac{\rho}{2}\|V-Z\|_F^2\end{aligned} \tag{3.10}$$

where $\rho > 0$ is a nonnegative penalty parameter, and $\|\cdot\|_F$ denotes the Frobenius norm. Then we solve the optimization problem with an iterative procedure based on the ADMM algorithm [2]. More specifically, in the t-th iteration, the parameters are updated as follows:

$$\begin{aligned}W^{t+1} \leftarrow \operatorname*{argmin}_W \mathcal{L}_{\rho 1}(W) = &\sum_{(x_i,x_j)\in\mathcal{I}_s} \log(1+\exp(-(x_i^T W^T W x_j - b))) \\ &+ \sum_{(x_i,x_j)\in\mathcal{I}_d} \log(1+\exp(x_i^T W^T W x_j - b)) \\ &+ \frac{\rho}{2}\|W - (Z^t - U_W^t)\|_F^2,\end{aligned} \tag{3.11}$$

$$V^{t+1} \leftarrow \operatorname*{argmin}_V \mathcal{L}_{\rho 2}(V) = \lambda_1 \sum_{i=1}^{d} \|V_i\|_2 + \frac{\rho}{2}\|V - (Z^t - U_V^t)\|_F^2, \tag{3.12}$$

$$Z^{t+1} \leftarrow \operatorname*{argmin}_Z \mathcal{L}_{\rho 3}(Z) = \lambda_2 Trace((Z\cdot Z^T)(R\odot R)^T) + \rho\|Z - \frac{1}{2}(W^{t+1} + V^{t+1} + U_W^t + U_V^t)\|_F^2, \tag{3.13}$$

$$U_W^{t+1} \leftarrow U_W^t + W^{t+1} - Z^{t+1}, \tag{3.14}$$

$$U_V^{t+1} \leftarrow U_V^t + V^{t+1} - Z^{t+1}, \tag{3.15}$$

where $U_W = \frac{1}{\rho}\Delta_W$ and $U_V = \frac{1}{\rho}\Delta_V$. Z^t, U_W^t and U_V^t are calculated in the previous iteration (If it is the first iteration, these parameters are randomly initialized). The above procedure will be iteratively conducted until the convergence criterion is satisfied. Then the optimal similarity metric W can be achieved. Next, we will discuss how to update each parameter according to the above equations.

W-update. We first fix V, Z, Δ_W and Δ_V, and then update W according to Eqn. (3.11). Since the objective function in Eqn. (3.11) is convex and differentiable, we adopt gradient decent method to update W, and the gradient is calculated as follows:

$$\begin{aligned}\frac{\partial \mathcal{L}_{\rho 1}}{\partial W} = &\sum_{(x_i,x_j)\in\mathcal{I}_s} \frac{(-W)(x_i x_j^T + x_j x_i^T)}{1+\exp((s(x_i,x_j)-b))} \\ &+ \sum_{(x_i,x_j)\in\mathcal{I}_d} \frac{W(x_i x_j^T + x_j x_i^T)}{1+\exp(-(s(x_i,x_j)-b))} \\ &+ \rho(W - (Z^t - U_W^t))\end{aligned} \tag{3.16}$$

V-update. In order to update V, we fix W, Z, Δ_W and Δ_V, and adopt the proximal operator [13]. Then Eqn.(3.12) can be efficiently computed via the following element-wise thresholding operation

$$V_{ij}^{t+1} = (Z^t - U_V^t)_{ij}[1 - \frac{\lambda_1}{\rho\|(Z^t - U_V^t)_{i,:}\|_2}]_+, \tag{3.17}$$

where $(Z^t - U_V^t)_{i,:}$ is the i-th row of $(Z^t - U_V^t)$, and the operator $[\cdot]_+$ means taking the maximum of zero and the argument inside.

Z-update. In this step, we fix W, V, Δ_W and Δ_V and update Z through minimizing $\mathcal{L}_{\rho 3}$ in Eqn. (3.13). Let the partial derivate of $\mathcal{L}_{\rho 3}$ with respect to Z be 0, and we can get

$$
\begin{aligned}
(3.18)\quad Z^{t+1} =& \rho(\lambda_2(R \odot R)^T \\
&+ \lambda_2(R \odot R) + 2\rho I)^{-1}(W^{t+1} + V^{t+1} + U_W^t + U_V^t).
\end{aligned}
$$

U_W, U_V-update. U_W and U_V are just updated according to Eqn. (3.14) and Eqn. (3.15), respectively.

The proposed patient similarity learning method is summarized in Algorithm 1. The convergence criterion can be a predetermined number of iterations or a threshold of the change in the estimated parameters in two consecutive iterations. Based on the convergence analysis for the ADMM algorithm in paper [2], we can know the proposed algorithm could produce a globally optimal solution for the optimization problem in Eqn. (3.8).

Algorithm 1 Uncorrelated patient similarity learning

Input: The sets of sample pairs $\mathcal{I}_s$ and $\mathcal{I}_d$, λ_1, λ_2, ρ
Output: The similarity metric W

```
Initialize V^0, Z^0, U_W^0, U_V^0;
repeat
    Update W according to Eqn. (3.11);
    Update V according to Eqn. (3.12);
    Update Z according to Eqn. (3.13);
    U_W^{t+1} ← U_W^t + W^{t+1} − Z^{t+1};
    U_V^{t+1} ← U_V^t + V^{t+1} − Z^{t+1};
until Convergence criterion is satisfied;
return The similarity metric W.
```

4 Distributed Patient Similarity Learning

In the above proposed approach, we assume that the learner has got the patient samples (i.e., training data) when conducting the similarity learning algorithm. However, in reality, the patient data are usually distributed across different sites (e.g., hospitals or medical centers), and these sites might not be willing to provide the raw data due to privacy concerns or resource consumption. This makes it difficult for the learner to achieve satisfactory learning accuracy. To address this challenge, we extend the aforementioned learning approach and propose a distributed patient similarity learning mechanism, based on which the learner could learn an accurate similarity metric without directly accessing the raw data at each site.

Suppose there is a set of parties (i.e., sites) $\mathcal{P} = \{1, 2, 3, ..., P\}$. Each party p has two sets of sample pairs: $\mathcal{I}_s^p = \{(x_{ip}, x_{jp}) : y_{ij}^p = 1)\}$ and $\mathcal{I}_d^p = \{(x_{ip}, x_{jp}) : y_{ij}^p = -1)\}$. Here x_{ip} and x_{jp} are two patient samples of party p, and y_{ij}^p is equal to 1 if the two patients are similar, otherwise y_{ij}^p is equal to -1. The main idea of the proposed distributed mechanism can be described as follows: each party p first learns local parameters based on the local data (i.e., $\mathcal{I}_s^p$ and $\mathcal{I}_d^p$) and uploads them to the learner. Then the learner combines all the received parameters and derives a global similarity metric W, which is then sent to each party p and used for updating the local parameters. This procedure is iteratively conducted until the convergence criterion is satisfied and the final W is what the learner wants to learn. In order to achieve the goal, we formulate the following optimization problem by considering the data samples from all the parties:

$$
\begin{aligned}
(4.19)\quad \min_{W,\{W_p\}_{p=1}^P} & \sum_{p\in P}\sum_{(x_{ip},x_{jp})\in\mathcal{I}_s^p} \log(1+\exp(-(s_p(x_{ip},x_{jp})-b))) \\
&+ \sum_{p\in P}\sum_{(x_{ip},x_{jp})\in\mathcal{I}_d^p} \log(1+\exp((s_p(x_{ip},x_{jp})-b))) \\
&+ \lambda_2 \sum_{p\in P} Trace((W_p \cdot W_p^T)(R_p \odot R_p)^T) \\
&+ \lambda_1 \|W\|_{2,1} \\
s.t.\ & W = W_p, \quad p = 1, 2, 3, ..., P,
\end{aligned}
$$

where W_p and R_p represent the local similarity metric and pearson correlation matrix of party p, respectively, and $s_p(x_{ip}, x_{jp}) = x_{ip}^T W_p^T W_p x_{jp}$. All optimal $W, W_1, W_2, ..., W_P$ are the same as the solution of the original problem. By adopting the scaled augmented Lagrangian multiplier, the optimization problem in Eqn. (4.19) can be formulated as

$$
\begin{aligned}
(4.20)\quad \min_{W,\{W_p,U_p\}_{p=1}^P} & \mathcal{L}_3(W, W_1, ..., W_P, U_1, ..., U_P) = \\
&\sum_{p\in P}\sum_{(x_{ip},x_{jp})\in\mathcal{I}_s^p} \log(1+\exp(-(s_p(x_{ip},x_{jp})-b))) \\
&+ \sum_{p\in P}\sum_{(x_{ip},x_{jp})\in\mathcal{I}_d^p} \log(1+\exp((s_p(x_{ip},x_{jp})-b))) \\
&+ \lambda_2 \sum_{p\in P} Trace((W_p \cdot W_p^T)(R_p \odot R_p)^T) + \lambda_1 \|W\|_{2,1} \\
&+ \frac{\lambda_3}{2} \sum_{p\in\mathcal{P}} \|W_p - W + U_p\|_{Frob}^2,
\end{aligned}
$$

where $\lambda_3 > 0$ is the penalty parameter, and $U_1, U_2, ..., U_P$ are the dual variables. We solve the optimization problem in Eqn. (4.20) through conducting an

iterative procedure. More specifically, in the t-th iteration, $\{W_p\}_{p=1}^P$, W and $\{U_p\}_{p=1}^P$ are updated as follows.

(4.21)
$$\begin{aligned} W_p^{t+1} = \underset{W_p}{\operatorname{argmin}} & \sum_{(x_{ip},x_{jp})\in\mathcal{I}_s^p} \log(1+\exp(-(s_p(x_{ip},x_{jp})-b))) \\ & + \sum_{(x_{ip},x_{jp})\in\mathcal{I}_d^p} \log(1+\exp((s_p(x_{ip},x_{jp})-b))) \\ & + \lambda_2 Trace((W_p\cdot W_p^T)(R_p\odot R_p)^T) \\ & + \frac{\lambda_3}{2}\|W_p - W^t + U_p^t\|_F^2 \end{aligned}$$

(4.22)
$$W^{t+1} = \underset{W}{\operatorname{argmin}} \ \frac{\lambda_3}{2}\sum_{p\in P}\|W_p^{t+1} - W + U_p^t\|_F^2 + \lambda_1\|W\|_{2,1}$$

(4.23)
$$U_p^{t+1} = U_p^t + W_p^{t+1} - W^{t+1}$$

That is to say, in each iteration the p-th party first updates its local similarity metric W_p based on the parameters W and U_p which are calculated in the previous iteration, then uploads W_p and U_p to the learner. After receiving the parameters $\{W_p, U_p\}_{p=1}^P$ from all the parties, the learner updates the global similarity metric W and sends it to each party to update U_p and W_p. This procedure is iteratively conducted until convergence and final W is the learned result. As we can see, each party only communicates with the learner with a considerably small amount of messages and the raw data are not uploaded to the learner. Therefore, the privacy issues as well as the communication and time cost are all addressed. Meanwhile, the proposed distributed mechanism always converges to the global optimal solution W^* which is learned from the whole dataset according to Algorithm 1. The convergence proof can be easily derived from paper [2].

5 Experiments

In this section, we conduct experiments on both real-world and synthetic datasets to evaluate the performance of the proposed approaches.

Baseline Methods. We compare the proposed approach with the following state-of-the-art similarity learning methods. **LowRank** [28] conducts patient similarity learning by selecting sparse features, and a low-rank structure is encoded to the learning process to implement the sparse feature selection. **R2ML** [8] learns the local distance metric by introducing a sparse-inducing matrix norm to control the rank of the involved mappings. **LMNN** [25] is a classical distance metric learning method, and its goal is to let the k-nearest neighbors always belong to the same class while examples from different classes are separated by a large margin. **ITML** [4] aims to learn the Mahalanobis distance by minimizing the differential relative entropy between two multivariate Gaussians. **GMML** [27] formulates distance metric learning process as an unconstrained smooth and convex optimization problem. Additionally, we also take **Cosine** and **Euclidean** as baselines, which adopt cosine similarity and l^2-norm distance to measure the similarity between two samples.

5.1 Experiments on Real-world Datasets. In this experiment, the three real-world datasets mentioned in Fig. 1 are used to measure the performance of the proposed UnPSL.

- **Colon cancer dataset** [1]. This dataset contains 40 tumor and 22 normal colon tissues which are characterized by 2000 genes. The samples were collected from 40 different colon cancer patients, in which 22 patients provided both normal and tumor samples. Since this dataset does not contain test set, we use the leave-one-out cross validation to evaluate the performance of the proposed method.
- **Parkinson's disease dataset** [14]. This dataset contains 22 features and 195 biomedical voice samples collected from 31 people, in which 23 were diagnosed with Parkinson's Disease (PD). Here we randomly select 98 samples as the training dataset, and the remaining samples are taken as the test set.
- **Leukemia dataset** [7]. This dataset consists of 72 samples and each of them is characterized by 7129 genes. There are 38 training data samples of which 27 are acute lymphoblastic leukaemia (ALL) and 11 are acute myeloid leukaemia (AML). The test set consists of 34 samples of which 20 are ALL and 14 are AML. We process this dataset with the method adopted in [5]: (i) thresholding: floor of 100 and ceiling of 16,000; (ii) filtering: exclusion of genes with $(\max - \min) \leq 500$ or $\max / \min \leq 5$, where max and min represent the maximum and minimum expression levels of a particular gene across the samples respectively; (iii) base 10 logarithmic transformation.

Performance comparison. We first compare the performance of UnPSL with that of the baselines on the above real-world patient datasets. In this paper, the KNN classifier is adopted to evaluate the performance of the methods, and we use *Accuracy*, *Recall*, *F1-score* and *F2-score* as the performance measures. For each method, we repeat the experiment 5 times and report the average results in Table 1. From Table 1, we can see UnPSL performs much better than the baselines on all of the datasets. The reason is that we take both sparse and uncorrelated feature selection into consideration,

Table 1: Performance comparison on real-world datasets

	Colon cancer dataset				Parkinson's disease dataset				Leukemia dataset			
	Accuracy	Recall	F1-score	F2-score	Accuracy	Recall	F1-score	F2-score	Accuracy	Recall	F1-score	F2-score
Cosine	0.6452	0.6452	0.7843	0.6944	0.6433	0.7810	0.7448	0.7597	0.7059	0.8000	0.7619	0.7843
Euc	0.7419	0.7541	0.8519	0.7904	0.7753	0.7720	0.8416	0.7983	0.7941	0.9000	0.8372	0.8739
ITML	0.8197	0.5714	0.6857	0.6122	0.7938	0.8393	0.8528	0.8437	0.8824	0.8462	0.9167	0.8730
Low-Rank	0.7903	0.8167	0.8829	0.8419	0.7979	0.8742	0.7389	0.7199	0.8824	0.9091	0.9375	0.9202
GMML	0.7903	0.6818	0.6977	0.6881	0.7443	0.8712	0.8358	0.8564	0.8235	0.8500	0.8500	0.8500
LMNN	0.7581	0.8393	0.8624	0.8484	0.8151	0.8621	0.8721	0.8661	0.8824	0.8261	0.9048	0.8559
R2ML	0.8065	0.6818	0.7143	0.6944	0.7835	0.8784	0.8609	0.8713	0.8889	0.9091	0.9091	0.9091
UnPSL	**0.8226**	**0.9444**	**0.9027**	**0.9273**	**0.8222**	**0.8823**	**0.8797**	**0.8802**	**0.9412**	**0.9500**	**0.9500**	**0.9500**

which guarantees that more relevant information can be extracted from the data samples when conducting patient similarity learning. Additionally, the results also show that Cosine and Euc have relatively poor performance when compared with other methods. This is mainly because Cosine and Euc can not well capture the statistical regularity of the data that needs to be learned from a large set of training samples.

Convergence. To evaluate the convergence of UnPSL, we calculate the primal residual $r_t = \|W^t - Z^t\|_F$ in each iteration t [2]. Figure 2 reports the results on the Parkinson's Disease dataset. Here we vary t from 2 to 60 and conduct the experiment for three times (i.e., Trial 1, Trial 2 and Trial 3). Each time we randomly select 98 samples from the dataset as the training data. From this figure we can see the primal residuals gradually converge to 0 with the increase of the number of iterations. This confirms that the convergence can be guaranteed in our proposed algorithm.

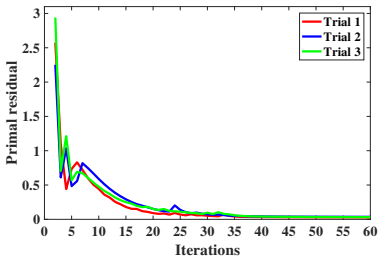


Figure 2: Primal residuals w.r.t Iterations on the Parkinson's disease dataset.

5.2 Experiments on Synthetic Datasets. Here, we evaluate the performance of UnPSL on synthetic datasets. We first introduce the data generation procedure, and then report the experimental results.

Data generation. We adopt the method mentioned in paper [11] to generate the synthetic datasets which have linear-correlated features. We first generate two data $X^1 = [x_1^1, x_2^1, ..., x_n^1] \in \mathcal{R}^{d*n}$ and $X^2 = [x_1^2, x_2^2, ..., x_n^2] \in \mathcal{R}^{d*n}$, where $x_i^1 \sim \mathcal{N}(0_{d*1}, I_{d*d})$ and $x_i^2 \sim \mathcal{N}(0_{d*1}, I_{d*d})$. Here I is an identity matrix. Then we generate data X^3 based on a linear combination of X^1 and X^2, i.e., $X^3 = 0.5(X^1 + X^2) + \epsilon$, where $\epsilon \sim \mathcal{N}(-0.1e, 0.1I_{d*d})$. Finally, we construct the data $X = [X^1; X^2; X^3] \in \mathcal{R}^{3d*n}$. It is obvious that the features in dimension $[2d+1, 3d]$ of X are highly correlated with the features in dimension $[1, d]$ and $[d+1, 2d]$. Let $W^1 \in \mathcal{R}^{3d*3d}$, where $W_{i,j}^1 \sim$ Uniform$(-0.5, 0.5)$. Suppose $W^2 = [1, ..., 1, 0, ..., 0]$ is a $3*d$ dimensional vector, of which the first $2*d$ elements are ones and the remaining d elements are zeros. Let $W = \text{diag}(W^2) * W^1 \in \mathcal{R}^{3d*3d}$. Then for any sample pair $(x_i, x_j) \in X$, we can use the similarity function $s(x_i, x_j)$ that is parameterized by W to generate the similarity label $y_{ij} \in \{-1, 1\}$.

Performance comparison. In this experiment, we assume that there are 400 data samples (i.e., $n = 400$), and generate three synthetic datasets by setting d as 20, 40 and 60, respectively. For each dataset, we randomly select 200 samples as the training data, and the remaining samples are taken as the testing data. Then we calculate *Accuracy*, *Recall*, *F1-score* and *F2-score* for each method. Here we repeat the experiment 10 times and report the average results in Table 2, from which we can see UnPSL performs much better than the baselines in all cases, and this further confirms that the similarity metric learned by our proposed method is more accurate than that learned by the baselines. Additionally, the results also show that the performance of the methods goes worse when d varies from 20 to 60. This is mainly because the higher dimensional (when d is larger) data usually has much more noisy and redundant information.

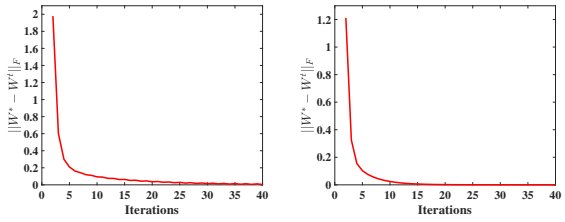
5.3 Experiments for Distributed Patient Similarity Learning. In this section, we evaluate the performance of the proposed distributed learning approach on two datasets, i.e. the Parkinson's disease dataset and a synthetic dataset ($n = 600$, $d = 10$). For simplicity, we consider a distributed scenario with 3 parties (i.e., $P = 3$) in this experiment. For each dataset, we equally divide it into three parts, and then assign them to the 3 parties respectively.

Suppose W^* is the similarity metric learned from the original dataset in the centralized scenario (according to Algorithm 1), and W^t is the similarity metric

Table 2: Performance comparison on synthetic datasets

	$d = 20$				$d = 40$				$d = 60$			
	Accuracy	Recall	F1 score	F2 score	Accuracy	Recall	F1 score	F2 score	Accuracy	Recall	F1 score	F2 score
Cosine	0.5200	0.5103	0.5183	0.5133	0.4976	0.4960	0.4983	0.4965	0.4896	0.5565	0.5390	0.5485
Euc	0.6144	0.3489	0.4617	0.3864	0.5872	0.3309	0.4365	0.3661	0.5384	0.3351	0.4379	0.3697
ITML	0.5408	0.4956	0.5237	0.5022	0.5228	0.4720	0.5102	0.4845	0.5092	0.5749	0.5682	0.5699
Low-Rank	0.8808	0.8574	0.8751	0.8644	0.8520	0.5440	0.7046	0.5986	0.8267	0.8560	0.6489	0.7386
GMML	0.7248	0.7424	0.7152	0.7311	0.6184	0.6017	0.5937	0.5978	0.5704	0.6766	0.6148	0.6493
LMNN	0.8960	0.8704	0.8785	0.5022	0.8080	0.7967	0.8033	0.7993	0.7840	0.7681	0.7970	0.7794
R2ML	0.9107	0.6573	0.6533	0.6358	0.8144	0.8227	0.8171	0.8199	0.8080	0.8150	0.8187	0.8164
UnPSL	**0.9464**	**0.9371**	**0.9441**	**0.9398**	**0.8712**	**0.8807**	**0.8735**	**0.8778**	**0.8413**	**0.8626**	**0.8594**	**0.8613**

learned in the t-th iteration of the distributed learning approach (according to Eqn. (4.22)). Then we calculate $\|W^* - W^t\|_F$ to measure the deviation between the two similarity metrics. Figure 3 shows the results on the Parkinson's disease dataset and synthetic dataset when iteration t varies from 2 to 40. From this figure we can see the value of $\|W^* - W^t\|_F$ converges to 0 gradually as the number of iterations increases. That is to say, the proposed distributed learning approach can achieve high accuracy as the learned similarity metric is almost the same with that learned in the centralized scenario. Additionally, this figure also shows that the proposed distributed learning approach converges faster on the synthetic dataset than that on the Parkinson's disease dataset. This is due to each party has more data samples for the synthetic dataset such that fewer iterations are needed to get the optimal solution.



(a) (b)

Figure 3: $\|W^* - W^t\|_F$ w.r.t Iterations on two datasets. (a): Parkinson's Disease dataset. (b): the synthetic dataset.

6 Related Work

Patient Similarity. LSML proposed in [20] aims to learn the similarity metric by using physician feedback as the supervision. [6, 19, 18] incorporate LSML into their proposed methods to address clinically relevant issues. Given that obtaining physicians' input is difficult and expensive in reality, Wang et al. [23] propose a weak supervised learning method. Due to the fact that patient similarity is highly context sensitive, Sun et al. [24] use both statistical and wavelet based features to capture the characteristics of patients. [30, 21] propose different deep learning frameworks to learn patient representations for similarity measuring. Based on patient similarity learning, [29] proposes a method to identify which drug is the most effective for a given patient. Under the medical social networks, [10] presents a method to calculate similarities of patient profiles for recommending people to other members. Considering the high dimensionality of medical data, Zhan et al. [28] propose a sparse feature selection method for patient similarity learning. However, these papers assume that the input features are nearly independent, while in real-world, the features in patients' data are usually highly correlated.

Distance Metric Learning. Distance metric learning has been studied in many works [25, 8, 4, 27]. Given some data pairs labeled as similar or dissimilar, these works try to learn similarity metrics to make similar pairs close to each other and separate dissimilar pairs apart. However, they do not consider feature correlations when conducting the learning algorithms.

Uncorrelation in other settings. [3, 11] address the uncorrelated lasso, but these methods can not be directly adopted here. In the context of multi-class regression, [12] proposes a method to eliminate the feature correlation via putting correlated features into the same group. However, it requires to know the groups in advance, which is also the drawback of group lasso. Additionally, it is time-costing and not suitable for the distributed settings.

7 Conclusions

In this paper, we propose a novel uncorrelated patient similarity learning approach which can extract more discriminative information from the patient data so that the learned similarity metric can be more accurate. To address the scenarios where the patient data are distributed across different sites, we also propose a distributed patient similarity learning approach, based on which the similarity metric can be accurately learned without directly accessing the raw data at each site. Experiments on both real-world patient datasets and synthetic datasets demonstrate the advantages of the proposed approaches.

8 Acknowledgments

This work was supported in part by the US National Science Foundation under grants IIS-1218393, IIS-1514204, IIS-1747614 and IIS-1553411. 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] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*, in Proc. of the National Academy of Sciences, 96(1999), pp. 6745–6750.

[2] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Foundations and Trends® in Machine Learning, 3(2011), pp. 1–122.

[3] S. Chen, C. H. Ding, B. Luo, and Y. Xie, *Uncorrelated Lasso*, in Proc. of AAAI, 2013.

[4] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, *Information-theoretic metric learning*, in Proc. of ICML, 2007.

[5] S. Dudoit, J. Fridlyand, and T. P. Speed, *Comparison of discrimination methods for the classification of tumors using gene expression data*, Journal of the American statistical association, 97(2002), pp. 77–87.

[6] 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 Proc. of AMIA, 2010.

[7] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, and others, *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*, Science, 286(1999), pp. 531–537.

[8] Y. Huang, C. Li, M. Georgiopoulos, and G. C. Anagnostopoulos, *Reduced-rank local distance metric learning*, in Proc. of ECML PKDD, 2013.

[9] Z. Huang, W. Dong, H. Duan, and H. Li, *Similarity measure between patient traces for clinical pathway analysis: problem, method, and applications*, IEEE journal of biomedical and health informatics, 18(2014).

[10] S. Klenk, J. Dippon, P. Fritz, and G. Heidemann, *Determining patient similarity in medical social networks*, in Proc. of MedEX, 2010.

[11] D. Kong, R. Fujimaki, J. Liu, F. Nie, and C. Ding, *Exclusive Feature Learning on Arbitrary Structures via L1,2 norm*, in Proc. of NIPS, 2014.

[12] D. Kong, Ji. Liu, B. Liu, and Xuan. Bao, *Uncorrelated Group LASSO*, in Proc. of AAAI, 2016.

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

[14] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, L. O. Ramig, and others, *Suitability of dysphonia measurements for telemonitoring of Parkinson's disease*, IEEE transactions on biomedical engineering, 56(2009), pp. 1015–1022.

[15] W. Liu, S. Ma, D. Tao, J. Liu, and P. Liu, *Semi-supervised sparse metric learning using alternating linearization optimization*, in Proc. of SIGKDD, 2010.

[16] W. Liu, C. Mu, R. Ji, S. Ma, J. R. Smith, and S. Chang, *Low-Rank Similarity Metric Learning in High Dimensions*, in Proc. of AAAI, 2015.

[17] G. Qi, J. Tang, Z. Zha, T. Chua, and H. Zhang, *An efficient sparse metric learning in high-dimensional space via l1-penalized log-determinant regularization*, in Proc. of ICML, 2009.

[18] J. Sun, D. Sow, J. Hu, and S. Ebadollahi, *Localized supervised metric learning on temporal physiological data*, in Proc. of ICPR, 2010.

[19] J. Sun, D. Sow, J. Hu, and S. Ebadollahi, *A system for mining temporal physiological data streams for advanced prognostic decision support*, ICDM, 2010.

[20] J. Sun, F. Wang, J. Hu, and S. Edabollahi, *Supervised patient similarity measure of heterogeneous patient records*, ACM SIGKDD Explorations Newsletter, 14(2012), pp. 16–24.

[21] Q. Suo, F. Ma, Y. Yuan, M. Huai, W. Zhong, A. Zhang, and J. Gao, *Personalized Disease Prediction Using a CNN-Based Similarity Learning Method*, in Proc. of BIBM, 2017.

[22] F. Wang, J. Hu, and J. Sun, *Medical prognosis based on patient similarity and expert feedback*, in Proc. of ICPR, 2012.

[23] F. Wang, and J. Sun, *PSF: a unified patient similarity evaluation framework through metric learning with weak supervision*, Journal of biomedical and health informatics, 19(2015), pp. 1053–1060.

[24] F. Wang, J. Sun, and S. Ebadollahi, *Integrating distance metrics learned from multiple experts and its application in patient similarity assessment*, SDM, 2011.

[25] K. Q. Weinberger, J. Blitzer, and L. K. Saul, *Distance metric learning for large margin nearest neighbor classification*, in Proc. of NIPS, 2006.

[26] Y. Ying, K. Huang, and C. Campbell, *Sparse metric learning via smooth optimization*, NIPS, 2009.

[27] P. Zadeh, R. Hosseini, and S. Sra, *Geometric mean metric learning*, in Proc. of ICML, 2016.

[28] M. Zhan, S. Cao, B. Qian, S. Chang, and J. Wei, *Low-Rank Sparse Feature Selection for Patient Similarity Learning*, in Proc. of ICDM, 2016.

[29] P. Zhang, F. Wang, J. Hu, and R. Sorrentino, *Towards personalized medicine: leveraging patient similarity and drug similarity analytics*, in Proc. of AMIA, 2014.

[30] 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 Proc. of ICDM, 2016.