01 Nov 2018

Early Detection of Disease using Electronic Health Records and Fisher's Wishart Discriminant Analysis

Sijia Yang

Jian Bian

Zeyi Sun
Missouri University of Science and Technology, sunze@mst.edu

Licheng Wang

et. al. For a complete list of authors, see https://scholarsmine.mst.edu/engman_syseng_facwork/697

Recommended Citation

S. Yang et al., "Early Detection of Disease using Electronic Health Records and Fisher's Wishart Discriminant Analysis," *Procedia Computer Science*, vol. 140, pp. 393-402, Elsevier B.V., Nov 2018.
The definitive version is available at https://doi.org/10.1016/j.procs.2018.10.299

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 140 (2018) 393–402

Procedia Computer Science

www.elsevier.com/locate/procedia

Complex Adaptive Systems Conference with Theme: Cyber Physical Systems and Deep Learning, CAS 2018, 5 November – 7 November 2018, Chicago, Illinois, USA

Early Detection of Disease Using Electronic Health Records and Fisher's Wishart Discriminant Analysis

Sijia Yang[a], Jian Bian[b], Zeyi Sun[c], Licheng Wang[a*], Haojin Zhu[d], Haoyi Xiong[b*], Yu Li[c]

[a]*State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications,Beijing, 100086, China*
[b]*Department of Computer Science, Missouri University of Science and Technology, Rolla, MO, 65409, US*
[c]*Department of Engineering Management and Systems Engineering, Missouri University of Science and Technology, Rolla, MO, 65409, US*
[d]*Department of Computer Science and Engineering, Shanghai Jiaotong University, Shanghai, 200040, US*

Abstract

Linear Discriminant Analysis (LDA) is a simple and effective technique for pattern classification, while it is also widely-used for early detection of diseases using Electronic Health Records (EHR) data. However, the performance of LDA for EHR data classification is frequently affected by two main factors: ill-posed estimation of LDA parameters (e.g., covariance matrix), and "linear inseparability" of the EHR data for classification. To handle these two issues, in this paper, we propose a novel classifier FWDA --- Fisher's Wishart Discriminant Analysis, which is developed as a faster and robust nonlinear classifier. Specifically, FWDA first surrogates the distribution of "potential" inverse covariance matrix estimates using a Wishart distribution estimated from the training data. Then, FWDA samples a group of inverse covariance matrices from the Wishart distribution, predicts using LDA classifiers based on the sampled inverse covariance matrices, and "weighted-averages" the prediction results via Bayesian Voting scheme. The weights for voting are optimally updated to adapt each new input data, so as to enable the nonlinear classification.

Keywords: Early Detection, Electronic Health Records, Linear Discriminant Analysis, Fisher's Wishart Discriminant Analysis

1. Introduction

The ubiquity of Electronic Health Records (EHR) systems [1, 2] in healthcare settings provides a unique opportunity for early detection of patients' potential diseases using their historical health records.

10.1016/j.procs.2018.10.299

To predict the patients' future disease using EHR data, existing work proposed to first extract useful features, such as diagnosis-frequencies [1- 3], pairwise diagnosis transition [4, 5], and graphs of diagnosis sequences [6], to represent each patient's EHR data using the representation learning techniques. Then, a series of supervised learning techniques have been adopted to train predictive models, such as Support Vector Machine (SVM), Random Forest (RF), Bayesian Network, Linear Discriminant Analysis (LDA) [1-3, 5, 7], using well represented EHR data with the labels of the target disease.

Among these methods, LDA with diagnosis-frequency vectors is frequently used as one of the common performance benchmarks [5, 7], because of LDA's provable Bayesian optimality [8]. However, recent studies demonstrate the limitation of LDA under high dimension low sample size (HDLSS) settings [9, 10], such as the EHR records [11]. Because it is difficult to recover the "true" parameters, e.g., covariance matrix, from a relatively small number of training samples. When the number of dimensions of EHR data is larger than the number of samples, the sample covariance estimation used in classical LDA, is singular and not invertible. Thus LDA cannot produce any valid prediction in this case. Even when the sample size is larger than the number of dimensions, the sample (inverse) covariance estimation could be quite different with the "true" (inverse) covariance matrix, with inconsistent largest eigenvalues and almost-orthogonal eigenvectors [12]. Such ill-posed estimation significantly disgraces LDA performance. Furthermore, EHR data is usually not simply linear. In terms of prediction, linear classifiers including LDA might not be the best algorithm to handle such data.

To address the ill-posed problems and/or non-linearity issues of data, several regularization-based methods have been proposed to accurately estimate the (inverse) covariance matrix [13-15] or linear coefficients [16, 17] under high dimension and low sample size (HDLSS) settings [18]. Further, to handle the non-linearity, some kernel-based or nonparameteric LDA classifiers [19-22] have been proposed. In summary, these methods intend to improve LDA classification through optimizing the parameters of LDA, such as (inverse) covariance matrices, linear projection metrics, or kernel settings, in a so-called optimal model selection manner [23].

Instead of "bidding" the optimal parameter among all possible (but uncertain) parameters, in this work, we intend to improve LDA, through averaging the classification results of LDA among all possible parameters [24] while adapting to the new input data. Specifically, given the training set and an input data, we try to first sample a set of all possible (inverse) covariance matrices from the both training and the new input data, then "average" the LDA classification results over the all generated (inverse) covariance matrices via Bayesian Voting Scheme [25]. Theoretical studies show that such Bayesian voting scheme can secure a wider margin and guarantee the good classification performance with a lower generalization error bound [25]. With lower risk, it on average can outperform the LDA classifier based on the any single (inverse) covariance matrix estimator [26]. More important, the set of sampled (inverse) covariance matrices for LDA averaging are updated by each new input data. In this way, the proposed classier enables nonlinear classification by leveraging local information of the input data.

However, to compute the aforementioned Input-Adaptive Bayesian Voting Scheme is not computationally easy. As the set of sampled (inverse) covariance matrices is assumed to be updated to adapt each new input data for classification, the sampling complexity of prediction is very high. Especially, when the number of dimensions (data) is high, it is quite time-consuming to sample the (inverse) covariance matrices from the data, while ensuring each sampled matrix is positive-semidefinite. Thus, we propose a novel method FWDA -- Fast Wishart Discriminant Analysis, which can approximate the optimal prediction results with minimal sampling efforts.

Specifically, FWDA first surrogates the distribution of inverse covariance matrices using a Wishart distribution estimated from the training data, then "weighted-averages" the classification results of LDA classifiers based on the sampled inverse covariance matrices. The "weights" are updated by each new input data for classification optimally in Bayes manner. In this way, FWDA can approximate to the aforementioned input-adaptive Bayesian voting design, with proven convergence rate. Our theoretical analysis further proves that (1) the error of approximation could quickly converge with increasing number of sampled inverse covariance matrices m in speed $O(m^{0.5})$; and (2) the error is not sensitive to the dimensions of the data, that means the performance of high dimension data classification could well-guaranteed.

In the rest of this paper, we first introduce the backgrounds, then formulate the problem of our research and elaborate the technical challenges in Section 2. In Section 3, we present the proposed algorithm FWDA, with the theoretical analysis on the approximation performance. In Section 4, we evaluate FWDA with other baseline algorithms for early detection of diseases using large-scale real-world EHR data.

2. Background and Problem Formulation

2.1. Binary Classification for Early Detection of Diseases using EHR data

First of all, we introduce the EHR data representation using diagnosis-frequency vectors, and present settings of disease detection through binary classification of diagnosis-frequency vectors. Later, we briefly discuss the solution based on the typical LDA classifier.

EHR Data Representation using Diagnosis-Frequency Vectors - There are many existing approaches to represent EHR data including the use of diagnosis-frequencies [1-3], pairwise diagnosis transition [4, 5], and graph representations of diagnosis sequences [6]. Among these approaches, the diagnosis-frequency is a common way to represent EHR data.

Given each patient's EHR data, this method first retrieves the diagnosis codes [27] recorded during each visit. Next, the frequency of each diagnosis appearing in all past visits are counted, followed by further transformation on the frequency of each diagnosis into a vector of frequencies. For example, <1, 0, …, 3>, where 0 means the second diagnosis does not exist in all past visits, In this paper, we denote the dimension of diagnosis-frequency vectors as p. Note that the dimension p >15,000 when using ICD-9 codes, p >250 even when using clustered ICD-9 codes [28], while the number of samples for training m is significantly smaller than p.

Early Detection by Binary Classification - Given m training samples (i.e., EHR frequency vectors) along with corresponding labels i.e., $(x_0,l_0)\ldots(x_{m-1}, l_{m-1})$ where l_i belongs to {-1,+1} refers to whether the patient i is diagnosed with the target disease or not (i.e., positive sample or negative sample), the early disease detection task is to determine if a new patient's data vector x would develop into the target disease by classifying the vector x to +1 (positive) or -1 (negative).

2.2. Linear Discriminant Analysis

To solve the binary classification problem aforementioned, we consider a simple LDA classifier $f(x) \in \{\pm 1\}$ based on the given p-dimension data vector x and labeled samples $x_1, x_2, ..., x_n$

$$f(x,\hat{\Sigma}) = sign((x-\overline{x})^T \hat{\Sigma}^{-1}(\overline{x}_{+1} - \overline{x}_{-1})) \tag{1}$$

where $\overline{x}$ refers to the mean vectors of all samples $x_1, x_2, ..., x_n$; $\overline{x}_{+1}$, $\overline{x}_{-1}$ refer to the mean vectors of the positive samples and negative samples receptively. The $\hat{\Sigma}$ is the covariance matrix estimated from data $x_1, x_2, ..., x_n$. The most common estimation of $\hat{\Sigma}$ is the sample estimation:

$$\overline{\Sigma} = \frac{1}{n-1} \sum_{1 \le j \le n} (x_j - \overline{x})^T (x_j - \overline{x}) \tag{2}$$

Thus, we write $f(x,\hat{\Sigma})$ as the classical Fisher's Linear Discriminant Analysis.

2.3. Nonlinear Extension with Adaptive Weighting

Given a binary linear classifier $h_\omega(x) \in \{\pm 1\}$, which is parameterized by ω, one can extend such classifier to an nonlinear classifier via input-adaptive weighting, such as:

$$sign(\int_\omega h_\omega(x) p(\omega; x) d_\omega) \tag{3}$$

where the signal function $sign(\cdot)$ maps to the non-negative input to +1 and the negative input to -1, and is the prior probability of the parameter ω with the input x given. As a binary classifier, the above classifier in Eq. 3 can be viewed as a Bayesian Voting Classifier [25], which outputs the label with the highest weighted vote.

2.4. Problem Formulation

To handle the uncertainty of (inverse-) covariance matrix estimates for LDA, through combining Bayesian Voting and LDA, we can consider a new nonlinear classifier as:

$$sign(\int_{\hat{\Sigma}\geq 0} f(\mathbf{x}, \hat{\Sigma}) P(\hat{\Sigma} | x_1, x_2, ...x_n, \mathbf{x}) d\hat{\Sigma}), \tag{4}$$

where $P(\hat{\Sigma} | x_1, x_2, ..., x_n, \mathbf{x})$ is the {probability of the covariance matrix $\hat{\Sigma}$, given the n number of training samples x_1, x_2, ..., x_n as well as the new sample for prediction $\mathbf{x}$. In our research, we named this pattern as Input Adaptive Bayesian Voting. Note that we take the new input vector x into account for generating the "hypothesis" $\hat{\Sigma}$ of Bayesian inference.

With all above backgrounds and settings in mind, the problem of this research is to compute Eq. 4. However, there exists at least two major technical challenges:

Challenge I: Fast Computation and Lazy Sampling To compute the integral in Eq. 4, a common solution is to leverage a Monte-Carlo Integration algorithm [29] that first randomly samples a group of positive-semidefinite matrices e.g., $\Sigma_1, \Sigma_2, ..., \Sigma_m$ from the distribution with probability density function $P(\hat{\Sigma} | x_1, x_2, ..., x_n, \mathbf{x})$, then averages $f(\mathbf{x}, \hat{\Sigma})$ over the sampled positive-semidefinite matrices as $1/m \Sigma_{i=1}^{m} f(x, \Sigma_i)$. This method can give an approximate result of Eq. 4. However, the density function of the sampled positive-semidefinite matrices $P(\hat{\Sigma} | x_1, x_2, ..., x_n, \mathbf{x})$ depends on the input $\mathbf{x}$. That means, for each new testing sample $\mathbf{x}$, we have to build a new probability distribution based on $P(\hat{\Sigma} | x_1, x_2, ..., x_m, \mathbf{x})$, then sample a new group of positive-semidefinite matrices and run the Monte-Carlo Integration accordingly. Obviously, the computational cost to re-sample a new group of positive-semidefinite matrices for each new input $\mathbf{x}$ is high. Thus, we need a "Lazy Sampling" mechanism, which only samples a group of positive-semidefinite matrices once, then uses the same group of matrices for arbitrary input $\mathbf{x}$.

Challenge II: Approximation and Convergence. The accuracy of classification highly depends on whether the proposed algorithm can approximate to the Eq. 4 as well as the convergence rate. For the high-dimensional numeric integration [30], the approximation is usually bottle-necked by the number of dimensions (e.g., the dimensionality of positive-semidefinite matrices $p \times p$) and the sampling complexity (e.g., the number of sampled positive-semidefinite matrices m). Intuitively, the convergence of algorithms can be improved, with increasing sampling complexity and lower dimensionality. However, we aim at proposing algorithm to approximate Eq. 4 with a low computational/sampling complexity while ensuring a fast convergence rate. Especially we require a convergence rate that is not sensitive to the dimensionality of the data p, so as to enable the high dimensional data classification.

In the rest of this paper, we present a novel nonlinear classifier, Fast Wishart Discriminant Analysis -- FWDA, which tackle the two research challenges, with low computational/sampling complexity and proven dimensionality-insensitive convergence rate.

3. FWDA: Algorithms and Analysis

In this section, we introduce our solution to compute Eq. 4, as follow: We first re-formulate Eq. 4. Then, we introduce the algorithms of FWDA to compute the reformulation of Eq. 4. Finally, we analyze FWDA.

3.1. Problem Reformulation

We first define $P(x | \Sigma)$ as the probability of input vector x given the covariance matrix Σ, and $P(\Sigma | x_1, x_2, ..., x_n)$ as the probability of the covariance matrix Σ, given the training samples x_1, x_2, ..., x_n. Then, we define a function:

$$g(x) = \int_{\Theta\geq 0} f(x, \Sigma) P(x | \Sigma) P(\Sigma | x_1, x_2 ... x_n) d\Sigma \tag{5}$$

Theorem 1. *Eq. 4 is equivalent to the classification result of sign*($g(x)$).

PROOF: Assuming all x_1, x_2, ..., x_n, $\mathbf{x}$ are *i.i.d* drawn from an unknown distribution, according to the Bayesian theorem, we decompose $P(\Sigma | x_1, x_2, ..., x_n, x)$ as

$$P(\Sigma^{-1} \mid x_1,...x_n,x) = \frac{P(x \mid \Sigma^{-1})P(x_1,...x_n \mid \Sigma^{-1})P(\Sigma^{-1})}{P(x)P(x_1,...x_n)} \\ = P(x \mid \Sigma^{-1})P(\Sigma^{-1} \mid x_1,x_2...x_n) \cdot P(x)^{-1} \tag{6}$$

Thus, Eq. 4 can be re-written as $sign(p(x)^{-1}g(x))$. As $p(x)^{-1}$ is a positive for $\forall x$. Thus, we can conclude $sign(g(x)) = sign(p(x)^{-1}g(x))$ should be consistently equivalent to the Eq. 4.

Thus, the key of proposed research is to compute Eq. 5. We propose a straightforward method (FWDA): the algorithm consists of a probabilistic model that can generate *m* sampled (inverse) covariance matrices according to the density function $P(\Sigma \mid x_1, x_2, ..., x_n)$, then calculates Eq. 4 through Monte-Carlo Integration using the sampled (inverse) covariance matrices. The design of FWDA is described in following.

3.2. Wishart Distribution Model based on De-sparsified Graphical Lasso

To sample (inverse) covariance matrices according to $P(\Sigma \mid x_1, x_2, ..., x_n)$, FWDA leverages a Wishart Distribution [31] namely $W(\hat{T},\upsilon)$, where $\hat{T}$ refers to the "mean" positive-definite matrix for the Wishart distribution and υ is the degree of freedom.

Given any $p \times p$ positive definite matrix Θ (as the inverse of potential covariance matrix), we estimate the probability density of Θ, based on $W(\hat{T},\upsilon)$, as:

$$P_w(\Theta \mid \hat{T},\upsilon) = \frac{1}{2^{\upsilon p/2} \mid \hat{T} \mid^{\upsilon/2} \Gamma_p(\frac{\upsilon}{2})} \mid \Theta \mid^{(\upsilon-p-1)/2_e-(1/2)\mathrm{tr}(\hat{T}^{-1}\Theta)} \tag{7}$$

where $\mid \cdot \mid$ refers to the Determinant and the multivariate gamma function is defined as:

$$\Gamma_p(\frac{\upsilon}{2}) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma(\frac{\upsilon}{2} + \frac{1-j}{2})$$

Specifically, in our research, we set the degree of freedom υ as $\upsilon = n\text{-}1$, and further estimate $\hat{T}$ using De-sparsified Graphical Lasso [32]:

$$\hat{T} = 2\hat{\Theta} - \hat{\Theta}\overline{\hat{\Sigma}}\hat{\Theta} \tag{8}$$

where $\hat{\Theta}$ refers to the Graphical Lasso estimator

$$\hat{\Theta} = \arg\min_{\Theta \geq 0} \left(\mathrm{tr}(\hat{\Sigma}\Theta) - \log \mid \Theta \mid + \lambda \sum_{j \neq k} \mid \Theta_{jk} \mid \right) \tag{9}$$

where $\hat{\Sigma}$ refers to the sample covariance matrix based on the samples $x_1, x_2, ..., x_n$, $\Sigma_{j \neq k} \mid \Theta_{jk} \mid$ refers to the sum of absolute value of the non-diagonal elements in matrix Θ.

3.3. Binary Classification as Bayesian Inference via Regularized Wishart Prior

Using the typical inverse-wishart sampling algorithm [33], FWDA first randomly generated *m* inverse-covariance matrices $\Theta_1, \Theta_2, \ldots, \Theta_m$ drawn from the Wishart Distribution $W(\hat{T},\upsilon)$. With the $\Theta_1, \Theta_2, \ldots, \Theta_m$, we approximate Eq. 4 as:

$$\overline{g}(x) = \frac{1}{m} \sum_{1 \leq i \leq m} \left(f(x,\Theta_i^{-1})P(x \mid \Theta_i^{-1}) \right) \tag{10}$$

where $P(x|\Theta_i^{-1})$ refers to the probability of the input vector **x** given the inverse covariance matrix Θ_i. In this paper, we characterize the probability as:

$$P(x \mid \Theta_i^{-1}) = \frac{1}{\sqrt{2\pi\left|\Theta_i^{-1}\right|}} e^{-\frac{1}{2}(\mathbf{x}-\bar{x})^T \Theta_i (\mathbf{x}-\bar{x})} \tag{11}$$

where $\bar{x} = n^{-1}\Sigma_1^n x_j$ refers to the mean vector of all training data. Thus, our algorithm FWDA uses $sign(\bar{g}(x))$ as the classification result. The performance analysis of the proposed algorithm based on Eq.10 to approximating the formulated problem expressed in Eq. 4 will be addressed in the following section.

3.4. Approximation Analysis

In this section, we present how close $\bar{g}(x)$ used in FWDA can approximate the re-formulated problem $g(x)$. First of all, considering the fast convergence rate of De-Sparsified Graphical Lasso [32] i.e., $\left\|\hat{T} - \Theta^*\right\|_\infty = O_p(\sqrt{\log p / n})$, with a fixed number of dimensions p and an increasing number of samples n, we are more confident to follow an assumption frequently made in many of previous Bayesian inference studies [34-36] :

Assumption 1. *For any positive-semidefinite matrix* Σ i.e., $\forall\Sigma \geq 0$ *and* $\Theta = \Sigma^{-1}$, *there exists* $P(\Sigma|x_1, x_2, ..., x_n) = P_w(\Theta|\hat{T}, \upsilon)$, *where* $P_w(\Theta|\hat{T}, \upsilon)$ *refers to the Wishart probability of* Θ *based on the mean positive-semidefinite matrix* $\hat{T}$ *and* $\upsilon = n\text{-}1$. $\hat{T}$ *is an estimate of inverse covariance matrix on samples* $x_1, x_2, ..., x_n$.

With Assumption 1., we can substitute $P(\Sigma|x_1, x_2, ..., x_n)$ with $P_w(\Theta|\hat{T}, \upsilon)$, i.e., the conjugate prior of inverse covariance matrix based on Wishart Distribution, to enable the Bayesian inference.

Theorem 2 *Under Assumption 1, when the number of sampled inverse covariance matrices* $m \to \infty$, our algorithm $\bar{g}(x)$ converges to $g(x)$ with convergence rate $O_p(\sqrt{-\log(\eta/2)/2m})$ under at least probability $1-\eta$.

PROOF: Sampled inverse covariance matrices Θ_1 , Θ_2 ,…, Θ_m are all drawn from the Wishart Distribution $W(\hat{T}, \upsilon)$ with the probability density function $P_w(\Theta|\hat{T}, \upsilon)$, thus there exists

$$\begin{aligned}
&\bar{g}(x) = \sum_{1\leq i\leq m}\left(f(x, \Theta_i^{-1}) \frac{P(x \mid \Theta_i) \cdot P_w(\Theta_i \mid \hat{T}, \upsilon)}{P_w(\Theta_i \mid \hat{T}, \upsilon)} \right) \\
&\text{Consider Central Limits Theorem } [37]. \\
&\underline{m \to +\infty} \int_{\Theta\geq 0} f(x, \Theta^{-1}) P(x, \Theta^{-1}) P_w(\Theta \mid \hat{T}, \upsilon) d\Theta \\
&\text{Consider the Assumption 1, and } \Theta = \Sigma^{-1} \\
&= \int_{\Sigma\geq 0} f(x, \Sigma) P(x \mid \Sigma) P(\Sigma \mid x_1, x_2, ...x_n) d\Sigma \\
&= g(x).
\end{aligned} \tag{12}$$

Further, based on Hoeffding's inequality [38], we can conclude that, with at least probability $1-\eta$

$$\left|g(x) - \bar{g}(x)\right| \leq \sqrt{-\frac{1}{2m} \cdot \log\frac{\eta}{2}}.$$

Based on **Theorem. 2**, we can conclude that the classification result of $sign(\bar{g}(x))$ should be equivalent to Eq. 4, when the number of sampled inverse covariance matrices m is large. Our later experiments show that, with more than 50 sampled inverse covariance matrices $m \geq 50$, FWDA can deliver decent performance and consistently outperform baseline algorithms, including SVM, Kernel SVM, Random Forest and AdaBoost.

4. Experimental Results

In this section, we introduce the experimental design of our evaluation. Then we present the experimental results, including the performance comparison between the FWDA framework, existing LDA baselines and other predictive models. Later a comparison between inverse covariance matrix supports our theoretical analysis of FWDA.

4.1. Experiment Setups

In this study, to evaluate FWDA, we used the de-identified EHR data from the College Health Surveillance Network (CHSN), which contains over 1 million patients and 6 million visits from 31 student health centers across the United States [39]. In the experiments, we use the EHR data from 10 participating schools. The available information includes ICD-9 diagnostic codes, CPT procedural codes, and limited demographic information. There are over 200,000 enrolled students in those 10 schools representing all geographic regions of the US. The demography of enrolled students (sex, race/ethnicity, age, undergraduate/graduate status) closely matched the demography for the population of US universities.

Among all diseases recorded in CHSN, we choose mental health disorders, including anxiety disorders, mood disorders, depression disorders, and other related disorders, as the targeted disease for early detection. We represent each patient using his/her diagnosis-frequency vector based on the clustered code set, where four clustered codes (i.e., 651, 657, 658, 662) are considered to represent the diagnoses of mental health disorders. Specifically, if a patient has any of these four codes in his/her EHR, we say that he/she has been diagnosed with mental health disorders as ground truth.

Note that in our research, we do not predict these four types of mental disorders separately, as these four disorders are usually correlated and heavily overlapped in clinical practices [40]. Further, patients with less than two visits were excluded from the analysis.

Until now, the diagnosis-frequency vectors used as predictors in our experiment only include the diagnosis frequency of physical health disorders and all mental health related information has been removed. In this case, our experiment is equivalent to predicting whether a patient would develop mental health disorders according to his/her past diagnoses of physical disorders.

To demonstrate the effectiveness of our method, we compared our method with baseline algorithms in terms of the following metrics: Accuracy and F1-Score. Specifically, the Accuracy metric characterizes the proportion of patients who are accurately classified in the early detection of mental disorders. The F1-Score measures both correctness and completeness of the early detection.

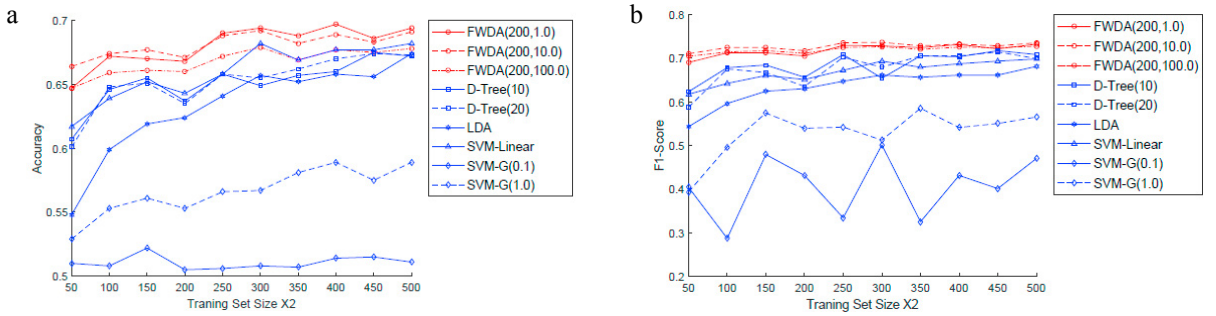

Figure 1: Overall Performance Comparison with Downstream Classifiers and Regularized LDA (a) Accuracy;(b) F1-Score

Baseline Algorithms To understand the performance impact of FWDA beyond classic LDA, we first propose LDA baseline approach to compare against FWDA, then three discriminative learning models are prepared for the comparison and two ensemble learning algorithms:

LDA-- This algorithm is based on the common implementation of generalized Fishier's discriminant analysis listed in Eq. 1. Specifically, LDA uses the sample covariance estimation, and inverts the covariance matrix using pseudo-inverse [41] when the matrix inverse is not available.

Linear Support Vector Machine (SVM-Linear), SVM with Gaussian Kernels (SVM-G) and Decision Tree-- SVM (Linear and Kernel) is a class of simple and efficient discriminative classifier formally defined by a separating hyperplane [42]. Decision Tree [43] is also a popular algorithm widely used in classification applications. We use the well-trained these baseline classifiers on the target datasets.

Random Forest and AdaBoost-- These are ensemble learning algorithms [44] which construct a set of classifiers to vote for a classification result given the new input data.

With above algorithms, we perform experiments with following settings: To build the training sets, we randomly selected 50, 100, 150, 200, and 250 patients with mental health disorders as the positive training samples, and randomly selected the same number of patients having not been diagnosed with any mental health disorders as negative training samples. Thus the training set of the two classes of patients is balanced. To build the test sets, we randomly selected 200 patients (not included in the training set) from both positive/negative groups as the testing set. Thus the testing set is also balanced. For each setting, we execute the seven algorithms and repeat 30 times.

4.2 Experiment Results

4.2.1 Overall Comparison

Fig. 1 presents the performance of our method and baselines on 200 testing samples. FWDA(200, 10.0) represents that FWD classifier using 200 sampled inverse covariance matrices by De-Sparsified Graphical Lasso with $\lambda = 10.0$. As can be seen from the experiment results, FWDA clearly outperforms the baseline algorithms in terms of overall accuracy, and F1-score. Specifically, FWDA achieves 3.6%--5.3% increase in accuracy and 5.9%--16% increase in F1-score compared to LDA; FWDA achieves 3.3%--4.9% increase in accuracy and 2.5%--4.6% increase in F1-score compared to Decision Tree (average for two regularized D-Tree). Compared to Linear SVM, the accuracy and F1-score of FWDA in most parameter settings are 0.3%--4.9% higher and 2.9%--4.8% higher, respectively. For other two Gaussian Kernel SVM, FWDA significantly outperforms them with an obvious gap. Thus, we can conclude that FWDA overall outperforms the baseline algorithms in all experimental settings. Please note that, though FWDA outperforms D-Tree and Linear SVM marginally, FWDA enjoys a more tight upper bound of expected error rate.

4.2.2. Comparison with Ensemble learners

As FWDA ensembles the classification results from multiple classifiers, we also compared FWDA to the existing ensemble learning algorithms, such as Random Forest and AdaBoost. To compare with ensemble learners with 100 and 200 basis classifiers, we use FWDA with 100 and 200 sampled inverse covariance matrices (i.e., ensemble with 100 and 200 LDA classifiers), with $\lambda = 1.0$ for Wishart mean matrix regularization. The performance comparison is illustrated in Fig. 2. It is obvious that FWDA outperforms these two algorithms in both 100-instance and 200-instance settings, while the performance of Random Forest is not quite stable. Moreover, Fig. 2 also shows the performance of FWDA classifiers with 100 and 200 sampled inverse covariance matrices are very similar. This indicates that FWDA can provide robust prediction performance, even when only a small number of inverse covariance matrices are sampled.

5. Conclusion

In this paper, we proposed FWDA --- Fisher's Wishart Discriminant Analysis--- a novel nonlinear discriminant analysis framework for early detection of diseases using Electronic Health Records (EHR) data. To enable the nonlinear classification, FWDA, uses the adaptive Bayesian voting scheme to weighted-average the prediction result, where the important model parameter inverse covariance matrix is sampled from the well-estimated Wishart distribution. Theoretical analysis presents that FWDA achieves a close approximation to the optimal Bayesian Voting LDA classifier. Furthermore, the experimental results on real-word EHR datasets shows that FWDA outperforms the downstream and ensemble learning classifiers.

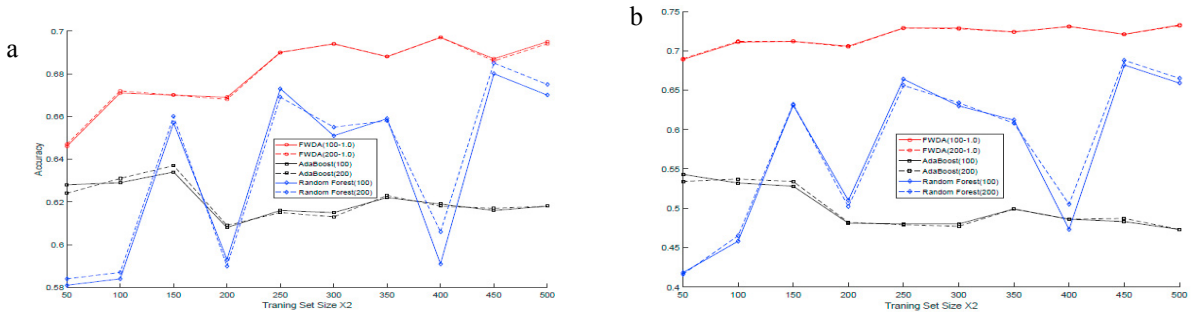

Fig.2. Performance Comparison with Ensemble Learning Classifiers (a) Accuracy; (b) F1-Score.

Reference

[1] Kenney Ng, Jimeng Sun, Jianying Hu, and Fei Wang. (2015) "Personalized Predictive Modeling and Risk Factor Identification using Patient Similarity." AMIA Summit on Clinical Research Informatics (CRI).

[2] Jimeng Sun, Fei Wang, Jianying Hu, and Shahram Edabollahi. (2012) "Supervised patient similarity measure of heterogeneous patient records." ACM SIGKDD Explorations Newsletter 14, 1, 16–24.

[3] Fei Wang and Jimeng Sun. (2015). "PSF: A Unified Patient Similarity Evaluation Framework Through Metric Learning With Weak Supervision." Biomedical and Health Informatics, IEEE Journal of 19, 3 (May 2015), 1053–1060. https://doi.org/10.1109/JBHI.2015.2425365

[4] Susan Jensen and UK SPSS. (2001) "Mining medical data for predictive and sequential patterns: PKDD 2001." In Proceedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases.

[5] Jinghe Zhang, Haoyi Xiong, Yu Huang, Hao Wu, Kevin Leach, and Laura E. Barnes. (2015) "MSEQ: Early Detection of Anxiety and Depression via Temporal Orders of Diagnoses in Electronic Health Data." In Big Data (Workshop), International Conference on. IEEE.

[6] Chuanren Liu, Fei Wang, Jianying Hu, and Hui Xiong. (2015). "Temporal Phenotyping from Longitudinal Electronic Health Records: A Graph Based Framework." In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15). ACM, New York, NY, USA, 705–714. https://doi.org/10.1145/2783258.2783352

[7] Fei Wang, Ping Zhang, Xiang Wang, and Jianying Hu. (2014) "Clinical Risk Prediction by Exploring High-Order Feature Correlations." In AMIA Annual Symposium Proceedings, Vol. American Medical Informatics Association, 1170.

[8] Onur C Hamsici and Aleix M Martinez. (2008) "Bayes optimality in linear discriminant analysis." TPAMI, 30(4):647-657.

[9] Amin Zollanvari, UlissesMBraga-Neto, and Edward R Dougherty. (2011) "Analytic study of performance of error estimators for linear discriminant analysis." IEEE Transactions on Signal Processing 59, 9 (2011), 4238–4255.

[10] Amin Zollanvari and Edward R Dougherty. (2013). "Random matrix theory in pattern classification: An application to error estimation." In 2013 Asilomar Conference on Signals, Systems and Computers.

[11] Joyce C Ho, Joydeep Ghosh, and Jimeng Sun. (2014) "Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization." In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 115--124. ACM.

[12] T Tony Cai, Zhao Ren, Harrison H Zhou, et al. (2016) "Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation." Electronic Journal of Statistics}, 10(1):1-59.

[13] Juwei Lu, Konstantinos N Plataniotis, and Anastasios N Venetsanopoulos. (2005) "Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition." Pattern Recognition Letters 26, 2, 181–191.

[14] Roger Peck and John Van Ness. (1982) "The use of shrinkage estimators in linear discriminant analysis." TPAMI 5, 530–537.

[15] Daniela M Witten and Robert Tibshirani. (2009) "Covariance-regularized regression and classification for high dimensional problems." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71, 3, 615–636.

[16] Line Clemmensen, Trevor Hastie, Daniela Witten, and Bjarne ErsbØll. (2011) "Sparse discriminant analysis." Technometrics, 53, (4).

[17] Jun Shao, Yazhen Wang, Xinwei Deng, Sijian Wang, et al. (2011) "Sparse linear discriminant analysis by thresholding for high dimensional data." The Annals of statistics 39, 2, 1241–1265.

[18] Peter Buhlmann and Sara Van De Geer. (2011) "Statistics for high-dimensional data: methods, theory and applications" Springer.

[19] Seung-Jean Kim, Alessandro Magnani, and Stephen Boyd. (2006) "Optimal kernel selection in kernel fisher discriminant analysis." In ICML. ACM, 465–472.

[20] Neil D Lawrence and Bernhard Schölkopf. (2001). "Estimating a kernel Fisher discriminant in the presence of label noise." In ICML, Vol. 1. Citeseer, 306–313.

[21] Zhihua Zhang et al. (2003) “Learning metrics via discriminant kernels and multidimensional scaling: Toward expected euclidean representation.” In ICML, Vol. 2. 872–879.

[22] Zhihua Zhang, Guang Dai, Congfu Xu, and Michael I Jordan. (2010). “Regularized discriminant analysis, ridge regression and beyond.” JMLR 11, Aug, 2199–2228.

[23] Kenneth P Burnham and David R Anderson. (2003) “Model selection and multimodel inference: a practical information-theoretic approach.” Springer Science & Business Media.

[24] Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky. (1999) “Bayesian model averaging: a tutorial.” Statistical science, 382–401.

[25] Nello Cristianini and John Shawe-Taylor. (1999) “Bayesian voting schemes and large margin classifiers.” Advances in Kernel Methodsâ˘AŤ Support Vector Learning, pages 55-68.

[26] Pascal Germain, Alexandre Lacasse, Francois Laviolette, Mario Marchand, and Jean-Francis Roy. (2015) “Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm.” The Journal of Machine Learning Research, 16(1):787—860.

[27] Erik R Dubberke, Kimberly A Reske, L Clifford McDonald, and Victoria J Fraser. (2006) “Icd-9 codes and surveillance for clostridium difficile-associated disease.” Emerging infectious diseases, 12(10),1576.

[28] HCUP. (2014) “Appendix a - clinical classification software-diagnoses”
https://www.hcup-us.ahrq.gov/toolssoftware/ccs/AppendixASingleDX.txt

[29] John Hammersley. (2013) “Monte carlo methods” Springer Science & Business Media.

[30] Philip J Davis and Philip Rabinowitz. (2007) “Methods of numerical integration.” Courier Corporation.

[31] LR~Haff. (1979) “Estimation of the inverse covariance matrix: Random mixtures of the inverse wishart matrix and the identity.” The Annals of Statistics, pages 1264—1276.

[32] Jana Jankova, Sara van de Geer, et al. (2015) “Confidence intervals for high-dimensional inverse covariance estimation.” Electronic Journal of Statistics 9, 1, 1205–1229.

[33] S Sawyer. (2007) “Wishart Distributions and Inverse-Wishart Sampling.” URL: www.math. wustl. edu/sawyer/hmhandouts/Whishart. pdf

[34] Ignacio Alvarez, Jarad Niemi, and Matt Simpson. (2014) “Bayesian inference for a covariance matrix.” arXiv preprint arXiv:1408.4050.

[35] Tom Leonard and John SJ Hsu. (1992) “Bayesian inference for a covariance matrix.” The Annals of Statistics, 1669–1696.

[36] Santosh Srivastava, Maya R Gupta, and Béla A Frigyik. (2007). Bayesian quadratic discriminant analysis. Journal of Machine Learning Research 8, Jun, 1277–1305.

[37] Aad W Van der Vaart. (2000) “Asymptotic statistics.” Vol. 3. Cambridge university press.

[38] Wassily Hoeffding. (1963) “Probability inequalities for sums of bounded random variables.” Journal of the American statistical association, 58(301):13-30.

[39] James C. Turner and Adrienne Keller. (2015) “College Health Surveillance Network: Epidemiology and Health Care Utilization of College Students at U.S. 4-Year Universities.” Journal of American college health: J of ACH (June 2015), 0.
https://doi.org/10.1080/07448481.2015.1055567

[40] Kenneth S Kendler, John M Hettema, Frank Butera, Charles O Gardner, and Carol A Prescott. (2003) “Life event dimensions of loss, humiliation, entrapment, and danger in the prediction of onsets of major depression and generalized anxiety.” Archives of general psychiatry 60, 8, 789–796.

[41] Jieping Ye, Ravi Janardan, Cheong Hee Park, and Haesun Park. (2004) “An optimization criterion for generalized discriminant analysis on undersampled problems.” TPAMI 26, 8 (2004), 982–994.

[42] Ingo Steinwart and Andreas Christmann. (2008) “Support vector machines.” Springer Science & Business Media.

[43] S Rasoul Safavian and David Landgrebe. (1991) “A survey of decision tree classifier methodology.” IEEE transactions on systems, man, and cybernetics 21, 3, 660–674.

[44] Thomas G Dietterich. (2000) “Ensemble methods in machine learning.” International workshop on multiple classifier systems}, pages 1-15, Springer.