Clinical Risk Prediction by Exploring High-Order Feature Correlations

Fei Wang, Ping Zhang, Xiang Wang, Jianying Hu
IBM T. J. Watson Research Center, Yorktown Heights, NY
{fwang,ping,wangxi,jyhu}@us.ibm.com

Abstract

Clinical risk prediction is one important problem in medical informatics, and logistic regression is one of the most widely used approaches for clinical risk prediction. In many cases, the number of potential risk factors is fairly large and the actual set of factors that contribute to the risk is small. Therefore sparse logistic regression is proposed, which can not only predict the clinical risk but also identify the set of relevant risk factors. The inputs of logistic regression and sparse logistic regression are required to be in vector form. This limits the applicability of these models in the problems when the data cannot be naturally represented vectors (e.g., medical images are two-dimensional matrices). To handle the cases when the data are in the form of multi-dimensional arrays, we propose *HOSLR*: High-Order Sparse Logistic Regression, which can be viewed as a high order extension of sparse logistic regression. Instead of solving one classification vector as in conventional logistic regression, we solve for K classification vectors in *HOSLR* (K is the number of modes in the data). A block proximal descent approach is proposed to solve the problem and its convergence is guaranteed. Finally we validate the effectiveness of *HOSLR* on predicting the onset risk of patients with Alzheimer's disease and heart failure.

1 Introduction

Predictive modeling of clinical risk, such as disease onset [1] or hospitalization [2], is an important problem in medical informatics. Effective risk prediction can be very helpful for the physician to make proper decision and provide the right service at point-of-care.

Typically we need three steps to perform patient clinical risk prediction:

1. Collecting all potential risk factors from patient historical data and utilizing them to properly represent each patient (e.g., as a vector [1][3]).
2. Identifying important risk factors from the risk factor pool collected in the first step, such that the value change of the selected risk factors could generate big impact on the predicted risk.
3. Training a proper predictive model based on the patients represented with the selected risk factors from the second step. Such model will be used to score the clinical risk of new testing patients.

One representative clinical risk prediction work that follows those three steps is the work by Sun *et al.* [1], where the goal is to predict the onset risk for potential heart failure patients. The authors first collect all potential risk factors from the two year patient electronic health records, and then designed an scalable orthogonal regression method to identify important risk factors, which will be used to train a logistic regression model for risk prediction at last. The authors showed that they can achieve the state-of-the-art performance as well as identify clinically meaningful risk factors for heart failure.

Note that in practice, depending on the concrete risk factor identification method and predictive model, step 2 and 3 could be combined into one step, i.e., a unified model can be constructed for both prediction and risk factor identification (e.g., LASSO [4]). This will make the constructed model more integrative and interpretable. Sparse Logistic Regression [5] is one such model. As is known to all that logistic regression is a popular model for clinical risk prediction [6] [1][3]. However, the pool of potential risk factors is usually very large and noisy, which would affect the efficiency and performance of predictive modeling. The main difference between sparse and convectional logistic regression is it adds an one norm regularizer on the model coefficients to encourage the model sparsity, so that only

those *important* risk factors will contribute to the final predictions. In recent years people have also been doing research on constructing different regularization terms to enforce different sparsity structures on the model coefficients, such as the ℓ_p norm [7], group sparsity [8] and elastic net regularization [9].

One limitation of the existing sparse logistic regression type of approaches is that they assume vector based inputs, which means that we need to have a vector based representation for each patient before we can use those methods to evaluate the patient's clinical risk. However, we are in the era of *big data* with *variety* as one representative characteristic, so does medical data, i.e., there are many medical data are not naturally in vector form. For example, typical medical images (e.g., X-Ray and MRI) are two dimensional matrices, with some more advanced medical imaging technologies can even generate three-dimensional image sequences (e.g., functional Magnetic Resonance Imaging (fMRI)). In a recent paper, Ho *et al*. [10] proposed a *tensor* (which can be viewed as high order generalization of matrix) based representation of patient Electronic Health Records (EHRs) to capture the interactions between different *modes* in patient EHRs. For example, medication order information for every patient could be captured by a 2nd order tensor with 2 modes, where each mode is an aspect of a tensor: a) medication and b) diagnosis. With such a representation we can take into consideration the correlation between diagnosis and drugs when predicting the patient risk. If there are more inter-correlated modes in the data then we will need to represent the patient in higher order tensors. In these cases, if we still want to apply logistic regression one straightforward way is to stretch those matrices and tensors into vectors as people did in image processing, but this will lose the correlation information among different dimensions. Moreover, after stretching the dimensionality of the data objects will become very high, which will make traditional sparse logistic regression inefficient.

In recent years, there has been a lot of research on extending traditional vector based approaches to 2nd (matrix based) or higher order (tensor based) settings. Two representative examples are two-dimensional Principal Component Analysis (PCA) [11] and Linear Discriminant Analysis [12], which have been found to be more effective on computer vision tasks compared to traditional vector based PCA and LDA. Recently, Huang and Wang [13] developed a matrix variate logistic regression model and applied it in electroencephalography data analysis. Tan *et al.* [14] further extended logistic regression to tensor inputs and achieved good performance in a video classification task.

In this paper, we propose *HOSLR*, a *H*igh-*O*rder *S*parse *L*ogistic *R*egression method that can perform prediction based on matrix or tensor inputs. Our model learns a linear decision vector on every mode of the input, and we added an ℓ_1 regularization term on each decision vector to encourage sparsity. We developed a *Block Proximal Gradient* (BPG) [15] method to solve the problem iteratively. The convergence of the proposed algorithm can be guaranteed by the Kurdyka–Lojasiewicz inequality [16] (for proof details please see a more technical version of this paper [17]). Finally we validate the effectiveness of our algorithm on two real world medical scenarios on the risk prediction of patients with Alzheimer's Disease and Heart Failure.

The rest of this paper is organized as follows. Section 2 reviews some related works. The details along with the convergence analysis of *HOSLR* is introduced in Section 3. Section 4 presents the experimental results, followed by the conclusions in Section 5.

2 Related Work

Logistic regression [18] is a statistical prediction method that has widely been used in medical informatics [1][6][19]. Suppose we have a training data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, where $\mathbf{x}_i \in \mathbb{R}^d$ $(1 \leqslant i \leqslant n)$ is the i-th training data vector with dimensionality d, and associated with each $\mathbf{x}_i$ we also have its corresponding label $y_i \in \{+1, -1\}$. The goal of logistic regression is to train a linear decision function $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ to discriminate the data in class +1 from the data in class -1 by minimizing the following logistic loss

$$\ell_{org}(\mathbf{w}, b) = \frac{1}{n}\sum_{i=1}^{n} \log\left[1 + \exp(-y_i(\mathbf{w}^\top \mathbf{x}_i + b))\right] \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^d$ is the decision vector and b is the bias. They can be learned with gradient descent type of approaches.

In many medical informatics applications, the data vectors $\{\mathbf{x}_i\}_{i=1}^n$ are sparse and high-dimensional (e.g., each patient could be a tens of thousands dimensional vector with bag-of-feature representation [1]). To enhance the interpretability

of the model in these scenarios, we can minimize the following ℓ_1-regularized logistic loss

$$\ell_{sp}(\mathbf{w}, b) = \frac{1}{n}\sum_{i=1}^{n} \log\left[1 + \exp(-y_i(\mathbf{w}^\top \mathbf{x}_i + b))\right] + \lambda\|\mathbf{w}\|_1 \tag{2}$$

where $\|\cdot\|_1$ is the vector $\ell 1$ norm and $\lambda > 0$ is a factor trading off the prediction accuracy and model sparsity. The resultant model is usually referred to as sparse logistic regression model [5][20]. Compared with the conventional logistic regression model obtained by minimizing $\mathcal{J}_{org}$, the $\mathbf{w}$ obtained by minimizing $\mathcal{J}_{sp}$ is sparse thanks to the ℓ_1 norm regularization. In this way, we can not only get a predictor, but also know what are the feature dimensions that are important to the prediction (which are the features with nonzero classification coefficients).

Sparse logistic regression has widely been used in health informatics because it can achieve a good balance between model accuracy and model interpretability. For example, sparse logistic regression has been used in the prediction of Leukemia [21], Alzheimer's disease [22] and cancers [23]. In recent years people also designed different regularization terms [7][8][9] to enforce more complex sparsity patterns on the learned model. However, all these works require a vector based data representations. Under this framework, if the data naturally come as tensors (like medical imaging), we need to first stretch them into vectors before we can apply sparse logistic regression. This may lose the correlation structure among different modes in the original data, while for the *HOSLR* method proposed in this paper, we directly work with data in tensor representations. Fig.1 provides a graphical illustration on the difference of traditional vector based logistic regression and high order logistic regression when working on multi-dimensional data.

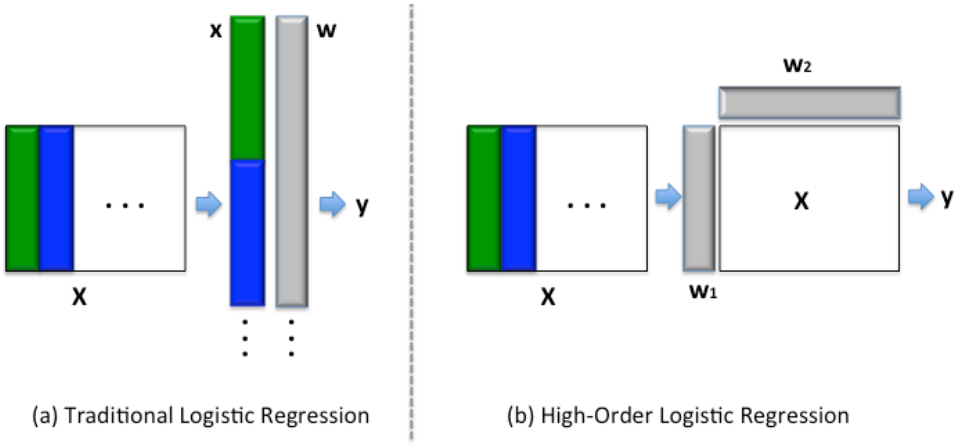

Figure 1: Traditional vector based logistic regression and high-order logistic regression work on multi-dimensional data.

3 Methodology

We introduce the details of *HOSLR* in this section. First we will formally define the problem.

3.1 Problem Statement

Without the loss of generality, we assume each observation is a tensor $\mathcal{X}^i \in \mathbb{R}^{d_1 \times d_2 \times \cdots d_K}$, suppose its corresponding response is $y^i \in \{0, 1\}$, then *HOSLR* assumes

$$y^i \leftarrow \mathcal{X}^i \times_1 \mathbf{w}^1 \times_2 \mathbf{w}^2 \cdots \times_K \mathbf{w}^K + b \tag{3}$$

where $\times_k$ is the mode-k product, and $\mathbf{w}^k \in \mathbb{R}^{d_k \times 1}$ is the prediction coefficients on the k-th dimension. Then

$$\mathcal{X}^i \times_1 \mathbf{w}^1 \times_2 \mathbf{w}^2 \cdots \times_K \mathbf{w}^K = \sum_{i_1=1}^{d_1}\sum_{i_2=1}^{d_2}\cdots\sum_{i_K=1}^{d_K} w^1_{i_1} w^2_{i_2} \cdots w^K_{i_K} X^i_{i_1 i_2 \cdots i_K} \tag{4}$$

Let $\mathcal{W} = \{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_K\}$ be the set of prediction coefficient vectors. The loss we want to minimize is

$$\begin{aligned} \ell(\mathcal{W}, b) &= \frac{1}{n}\sum_{i=1}^{n} \ell(\mathcal{X}_i, y_i, \mathcal{W}) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \mathcal{X}^i \times_1 \mathbf{w}^1 \times_2 \mathbf{w}^2 \cdots \times_K \mathbf{w}^K + b) \\ &= \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f_{(\mathcal{W},b)}(\mathcal{X}^i)) \end{aligned}$$

where for notational convince, we denote

$$f_{(\mathcal{W},b)}(\mathcal{X}^i) = \mathcal{X}^i \times_1 \mathbf{w}^1 \times_2 \mathbf{w}^2 \cdots \times_K \mathbf{w}^K + b \tag{5}$$

The loss function we considered in this paper is *Logistic Loss*:

$$\ell_l(\mathcal{W}, b) = \log[1 + \exp(-y_i f_{(\mathcal{W},b)}(\mathcal{X}^i))] \tag{6}$$

We also introduce the regularization term

$$\mathcal{R}(\mathcal{W}) = \mathcal{R}_1(\mathcal{W}) + \mathcal{R}_2(\mathcal{W}) = \sum_{k=1}^{K} \lambda_k \|\mathbf{w}^k\|_1 + \frac{1}{2}\sum_{k=1}^{K} \mu_k \|\mathbf{w}^k\|_2^2 \tag{7}$$

which is usually referred to as *elastic net* regularization [24]. This regularizer is a combination of ℓ_1 and ℓ_2 norm regularizations, thus it can achieve better numerical stability and reliability [24]. Then the optimization problem we want to solve is

$$\min_{\mathcal{W}} \mathcal{J}(\mathcal{W}, b) = \ell(\mathcal{W}, b) + \mathcal{R}(\mathcal{W}) \tag{8}$$

We adopt a *Block Coordinate Descent* (BCD) procedure to solve the problem. Starting from some initialization $(\mathcal{W}_{(0)}, b_{(0)})$, at the i-th step of the t-th round of updates, we update $(\mathbf{w}^k_{(t)}, b_{(t)})$ by

$$(\mathbf{w}^k_{(t)}, b_{(t)}) = \arg\min_{(\mathbf{w},b)} \left[\ell(\mathcal{W}^{1\sim(k-1)}_{(t)}, \mathbf{w}, \mathcal{W}^{(k+1)\sim K}_{(t-1)}, b) + \lambda_k \|\mathbf{w}\|_1 + \frac{\mu_k}{2}\|\mathbf{w}\|_2^2\right]$$

where $\mathcal{W}^{1\sim(k-1)}_{(t)} = \left\{\mathbf{w}^1_{(t)}, \mathbf{w}^2_{(t)}, \cdots, \mathbf{w}^{k-1}_{(t)}\right\}$ and $\mathcal{W}^{(k+1)\sim K}_{(t-1)} = \left\{\mathbf{w}^{k+1}_{(t-1)}, \mathbf{w}^{k+2}_{(t-1)}, \cdots, \mathbf{w}^{K}_{(t-1)}\right\}$.

Algorithm 1 Block Coordinate Descent Procedure

Require: Data set $\{\mathcal{X}_i, y_i\}_{i=1}^n$, Regularization parameters $\{\lambda_k, \mu_k\}_{k=1}^K$
1: **Initialization**: $(\mathcal{W}_{(0)}, b_{(0)})$, $t = 0$
2: **while** Not Converge **do**
3: **for** $k = 1 : K$ **do**
4: Update $(\mathbf{w}^k_{(t)}, b_{(t,k)})$ by solving problem (9)
5: $t = t + 1$
6: **end for**
7: **end while**

3.2 Proximal Gradient Descent

Algorithm 2 summarized the whole algorithmic flow of our algorithm, where $\alpha^k_{(t)} = \lambda_k/(\tau^k_{(t)} + \mu_k)$ and $\mathcal{S}_{\alpha^k_{(t)}}$ is the component-wise shrinkage operator defined as

$$\left(\mathcal{S}_{\alpha^k_{(t)}}(\mathbf{v})\right)_i = \begin{cases} v_i - \alpha^k_{(t)}, & \text{if } v_i > \alpha^k_{(t)} \\ v_i + \alpha^k_{(t)}, & \text{if } v_i < -\alpha^k_{(t)} \\ 0, & \text{if } |v_i| \leqslant |\alpha^k_{(t)}| \end{cases} \tag{9}$$

At each iteration the most time consuming part is evaluating the gradient, which takes $O(n\prod_{i=1}^K d_i)$ time, that is linear with respect to data set size and data dimension. The detailed algorithm derivation can be referred to [17].

Algorithm 2 Block Proximal Gradient Descent for Multilinear Sparse Logistic Regression

Require: Data set $\{\mathcal{X}_i, y_i\}_{i=1}^n$, Regularization parameters $\{\lambda_k, \mu_k\}_{k=1}^K$, $r_0 = 1$, $\delta_\omega < 1$

Initialization: $(\mathcal{W}_{(0)}, b_{(0)})$, $t = 1$
while Not Converge **do**
  for $k = 1 : K$ **do**
    Compute $\tau_{(t)}^k$ with $\tau_{(t)}^k = \frac{\sqrt{2}}{n} \sum_{i=1}^n \left(\left\| \nabla_{\mathbf{w}^k}^{(t,k)} f_{(\mathcal{W},b)}(\mathcal{X}^i) \right\|_2 + 1 \right)^2$
    Compute $\omega_{(t)}^k$ with $\omega_{(t)}^k = \min\left(\omega_{(t)}, \delta_\omega \sqrt{\frac{\tau_{(t-1)}^k}{\tau_{(t)}^k}}\right)$
    Compute $\widetilde{\mathbf{w}}_{(t)}^k$ with $\widetilde{\mathbf{w}}_{(t)}^k = \mathbf{w}_{(t-1)}^k + \omega_{(t)}^k(\mathbf{w}_{(t-1)}^k - \mathbf{w}_{(t-2)}^k)$
    Update $\mathbf{w}_{(t)}^k$ by $\mathbf{w}_{(t)}^k = \mathcal{S}_{\alpha_{(t)}^k}\left(\frac{\tau_{(t)}^k \widetilde{\mathbf{w}}_{(t)}^k - \nabla_{\mathbf{w}^k} \ell_{(t)}^k(\widetilde{\mathbf{w}}_{(t)}^k, b_{(t,k-1)})}{\tau_{(t)}^k + \mu_k}\right)$
    Compute $\widetilde{b}_{(t,k)}$ with $\widetilde{b}_{(t,k)} = b_{(t,k-1)} + \omega_{(t)}^k(b_{(t,k-1)} - b_{(t,k-2)})$
    Update $b_{(t,k)}$ by $b_{(t,k)} = \widetilde{b}_{t,k} - \frac{1}{\tau_{(t)}^k} \nabla_b \ell_{(t)}^k(\mathbf{w}_{(t)}^k, \widetilde{b}_{(t,k)})$
  end for
  if $\ell(\mathcal{W}_{(t-1)}, b_{(t-1,K)}) \leqslant \ell(\mathcal{W}_{(t)}, b_{(t,K)})$ **then**
    Reupdate $\mathbf{w}_{(t)}^k$ and $b_{(t,k)}$ using $\mathbf{w}_{(t)}^k = \mathcal{S}_{\alpha_{(t)}^k}\left(\frac{\tau_{(t)}^k \widetilde{\mathbf{w}}_{(t)}^k - \nabla_{\mathbf{w}^k} \ell_{(t)}^k(\widetilde{\mathbf{w}}_{(t)}^k, b_{(t,k-1)})}{\tau_{(t)}^k + \mu_k}\right)$ and $b_{(t,k)} = \widetilde{b}_{t,k} - \frac{1}{\tau_{(t)}^k} \nabla_b \ell_{(t)}^k(\mathbf{w}_{(t)}^k, \widetilde{b}_{(t,k)})$, with $\widetilde{\mathbf{w}}_{(t)}^k = \mathbf{w}_{(t-1)}^k$ and $\widetilde{b}_{(t,k)} = b_{(t,k-1)}$
  end if
  $t = t + 1$
end while

4 Experiments

In this section we will present the experimental results on applying *HOSLR* to predict the onset risk of potential Alzheimer's Disease patients from their fMRI images, and the onset risk of potential heart failure patients from their EHR data.

4.1 Experiments on Predicting the Onset Risk of Alzheimer's Disease

Alzheimer's disease (AD) is the most common form of dementia. It worsens as it progresses and eventually leads to death. There is no cure for the disease. AD is usually diagnosed in elder people (typically over 65 years of age), although the less-prevalent early-onset Alzheimer's can occur much earlier. There are currently more than 5 million Americans living with Alzheimer's disease and that number is poised to grow to as many as 16 million by 2050. The care for has been the country's most expensive condition, which costs the nation $203 billion annually with projections to reach $1.2 trillion by 2050 [25].

Early detection of AD is of key importance for its effective intervention and treatment, where *functional magnetic resonance imaging or functional MRI* (fMRI) [26] is an effective approach to investigate alterations in brain function related to the earliest symptoms of Alzheimer's disease, possibly before development of significant irreversible structural damage.

In this set of experiment, we adopted a set of fMRI scans collected from real clinic cases of 1,005 patients [27], whose cognitive function scores (semantic, episodic, executive and spatial - ranges between -2.8258 and 2.5123) were also acquired at the same time using a cognitive function test. There are three types of MRI scans that were collected from the subjects: (1) FA, the fractional anisotropy MRI gives information about the shape of the diffusion tensor at each voxel, which reflects the differences between an isotropic diffusion and a linear diffusion; (2) FLAIR, Fluid attenuated inversion recovery is a pulse sequence used in MRI, which uncovers the white matter hyperintensity of the brain; (3)

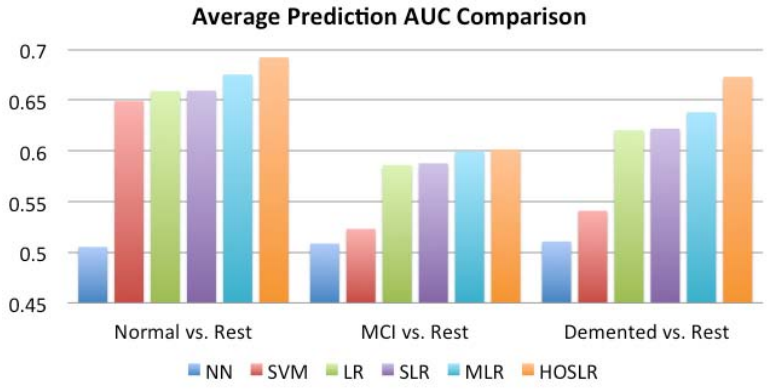

Figure 2: Average prediction AUC over 5-fold cross validation comparison for different methods.

GRAY, gray MRI images revealing the gray matter of the brain. In the raw scans, each voxel has a value from 0 to 1, where 1 indicates that the structural integrity of the axon tracts at that location is perfect, while 0 implies either there are no axon tracts or they are shot (not working). The raw scans are preprocessed (including normalization, denoising and alignment) and then restructured to 3D tensors with a size of $134 \times 102 \times 134$. Associated with each sample we have a label, which could be either *normal*, *Mild Cognitive Impairment* (MCI) or *demented*.

We constructed three binary classification problems to test the effectiveness of our *HOSLR* method, i.e., *Normal* vs. *Rest* (MCI and Demented), *MCI* vs. *Rest* (Normal and Demented), *Demented* vs. *Rest* (Normal and MCI). For *HOSLR*, because the input fMRI images are three dimensional tensors, we set the ℓ_1 term regularization parameters on all three dimensions equal, i.e., $\lambda_1 = \lambda_2 = \lambda_3$ and tune it from the grid $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$ with five fold cross validation. The ℓ_2 term regularization parameters are set to $\mu_1 = \mu_2 = \mu_3 = 10^{-4}$. For comparison purpose, we also implemented the following baseline algorithms:

- **Nearest Neighbor** (NN). This is the one nearest neighbor classifier with standard Euclidean distance.
- **Support Vector Machine** (SVM). This is the regular vector based SVM method.
- **Logistic Regression** (LR). This is the traditional vector based logistic regression method.
- **Sparse Logistic Regression** (SLR). This is the vector based sparse logistic regression.
- **Multilinear Logistic Regression** (MLR). This is equivalent to *HOSLR* with all $\ell 1$ regularization parameters setting to 0.

We use *LIBLINEAR* [28] for the implementation of LR and SLR, and *LIBSVM* [29] for the implementation of SVM. Note that in order to test those vector based approaches, we need to stretch those fMRI tensors into very long vectors (with dimensionality 1,831,512). Fig.2 summarized the average performance over 5-fold cross validation in terms of Areas Under the receiver operating characteristics Curve (AUC) values. The data we used are the FLAIR images. From the figure we can observe that *HOSLR* beats all other competitors in all three tasks. This is because *HOSLR* can not only take into consideration the spatial correlation between three different dimensions in those fMRI images, but also exploring their joint sparsity structures (the FLAIR images are sparse in nature).

4.2 Experiments on Predicting the Onset Risk of Congestive Heart Failure Patients

Congestive heart failure (CHF), occurs when the heart is unable to pump sufficiently to maintain blood flow to meet the needs of the body, is a major chronic illness in the U.S. affecting more than five million patients. It is estimated CHF costs the nation an estimated $32 billion each year [30]. Effective prediction of the onset risk of potential CHF

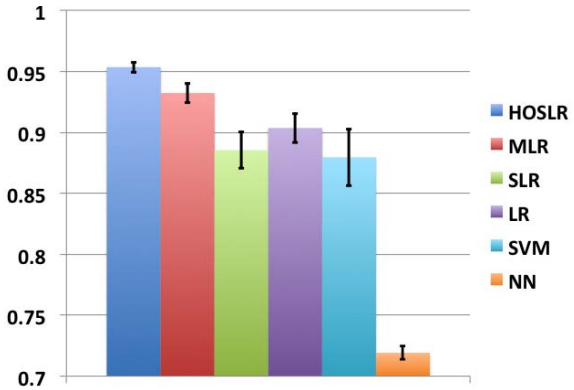

Figure 3: Prediction performance for different methods on the CHF onset prediction task in terms of averaged AUC value with 5-fold cross validation along with their standard deviations.

patients would help identify the patient at risk in time, and thus the decision makers can provide the proper treatment. This can also help save huge amount of unnecessary costs.

The data set we use in this set of experiments is from a real world Electronic Health Record (EHR) data warehouse including the longitudinal EHR of 319,650 patients over 4 years. On this data set, we identified 1,000 CHF case patients according to the diagnostic criteria in [3]. Then we obtained 2,000 group matched controls according to patient demographics, comorbidities and primary care physicians similar as in [3]. We use the medication orders of those patients within two years from their operational criteria date (for case patients, their operational criteria dates are just their CHF confirmation date; for control patients that date is just the date of their last records in the database). On each medication order we use the corresponding pharmacy class according to the United States Pharmacopeial (USP) convention[1] and the primary diagnosis in terms of Hierarchical Condition Category (HCC) codes [31] for the medication prescription. In total there are 92 unique pharmacy classes and 195 distinct HCC codes appeared in those medication orders. Therefore each patient can be represented as a 92×195 matrix, where the (i, j)-th entry indicates the frequency that the i-th drug was prescribed during the two years with the j-th diagnosis code as primary diagnosis.

The parameters for *HOSLR* are set in the same manner as the experiments in last subsection. For comparison purpose, we also implemented NN, SVM, LR, SLR, MLR and reported the averaged AUC value over 5-fold cross validation along with their standard deviations on Fig.3. From the figure we can get similar observations as we saw in Fig.2.

Another interesting thing to check is which medications and diagnosis play key roles during the decision. Because in this set of experiments we have two feature modes: medications and diagnosis, we will get two decision vectors $\mathbf{w}_{\text{med}}$ and $\mathbf{w}_{\text{diag}}$, one on each mode. The bilinear decision function in this case can be written as

$$f(\mathbf{X}) = \mathbf{w}_{\text{med}}^\top \mathbf{X} \mathbf{w}_{\text{diag}} = \mathbf{1}^\top \left((\mathbf{w}_{\text{med}} \mathbf{w}_{\text{diag}}^\top) \odot \mathbf{X} \right) \mathbf{1} \tag{10}$$

where $\mathbf{1}$ denotes all-one vector of appropriate dimension, $\odot$ is element-wise matrix product. The importance of the (i, j)-th feature X_{ij} to the decision can be evaluated as $w_{\text{med}}(i) w_{\text{diag}}(j)$. Therefore if both the magnitudes of $w_{\text{med}}(i)$ and $w_{\text{diag}}(j)$ are large, then the feature pair (medication i, diagnosis j) will definitely be important. We list in Table 1 the top diagnoses and medications according to their coefficient magnitudes in $\mathbf{w}_{\text{diag}}$ and $\mathbf{w}_{\text{med}}$. From the table we can see that the diagnoses are mainly hypertension, heart disease and some common comorbidities of heart failure including chronic lung disease (e.g., *Chronic Obstructive Pulmonary Disease* (COPD) [32]) and chronic kidney disease [33]. The top medications include drugs for treating heart disease such as Beta blockers and calcium blockers, and medicine for treating lung disease such as *Corticosteroids*. There are also medicine for treating heart failure related symptom, such as Gout, which is a well-known Framingham symptom [34]. Vaccine is also an important treatment for reducing the stress on heart [35].

[1] http://www.usp.org/

Diagnosis	
Heart Disease	Congestive Heart Failure
	Acute Myocardial Infarction
	Specified Heart Arrhythmias
	Ischemic or Unspecified Stroke
Hypertension	Hypertension
	Hypertensive Heart Disease
Lung Disease	Fibrosis of Lung and Other Chronic Lung Disorders
	Asthma
	Chronic Obstructive Pulmonary Disease (COPD)
Kidney Disease	Chronic Kidney Disease, Very Severe (Stage 5)
	Chronic Kidney Disease, Mild or Unspecified (Stage 1-2 or Unspecified)

Medication
Antihyperlipidemic
Antihypertensive
Beta Blockers
Calcium Blockers
Cardiotonics
Cardiovascular
Corticosteriods
Diuretics
General Anesthetics
Gout
Vaccines

Table 1: Top diagnosis and medications according to the magnitude of their corresponding decision coefficient

5 Conclusions

We propose a high order sparse logistic regression method called *HOSLR* in this paper, which can directly take data matrices or tensors as inputs and do prediction on that. *HOSLR* is formulated as an optimization problem and we propose an effective BCD strategy to solve it. We validate the effectiveness of *HOSLR* on two real world medical scenarios on predicting the onset risk of Alzheimer's disease and heart failure. We demonstrate that *HOSLR* can not only achieve good performance, but also discover interesting predictive patterns.

References

[1] Jimeng Sun, Jianying Hu, Dijun Luo, Marianthi Markatou, Fei Wang, Shahram Edabollahi, Steven E Steinhubl, Zahra Daar, and Walter F Stewart. Combining knowledge and data driven insights for identifying risk factors using electronic health records. In *AMIA Annual Symposium Proceedings*, volume 2012, page 901. American Medical Informatics Association, 2012.

[2] Edward F Philbin and Thomas G DiSalvo. Prediction of hospital readmission for heart failure: development of a simple risk score based on administrative data. *Journal of the American College of Cardiology*, 33(6):1560–1566, 1999.

[3] Jionglin Wu, Jason Roy, and Walter F Stewart. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. *Medical care*, 48(6):S106–S113, 2010.

[4] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[5] Shirish Krishnaj Shevade and S Sathiya Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246–2253, 2003.

[6] Marc Miravitlles, Tina Guerrero, Cristina Mayordomo, Leopoldo Sánchez-Agudo, Felip Nicolau, and José Luis Segú. Factors associated with increased risk of exacerbation and hospital admission in a cohort of ambulatory COPD patients: a multiple logistic regression analysis. *Respiration*, 67(5):495–501, 2000.

[7] Zhenqiu Liu, Feng Jiang, Guoliang Tian, Suna Wang, Fumiaki Sato, Stephen J Meltzer, and Ming Tan. Sparse logistic regression with Lp penalty for biomarker identification. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007.

[8] Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group LASSO for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[9] Srikanth Ryali, Kaustubh Supekar, Daniel A Abrams, and Vinod Menon. Sparse logistic regression for whole-brain classification of fMRI data. *NeuroImage*, 51(2):752–764, 2010.

[10] Joyce C Ho, Joydeep Ghosh, Steve Steinhubl, Walter Stewart, Joshua C Denny, Bradley A Malin, and Jimeng Sun. Limestone: High-throughput candidate phenotype generation via tensor factorization. *Journal of Biomedical Informatics*, 2014.

[11] Jian Yang, David Zhang, Alejandro F Frangi, and Jing-yu Yang. Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(1):131–137, 2004.

[12] Jieping Ye, Ravi Janardan, Qi Li, et al. Two-dimensional linear discriminant analysis. In *Advances in Neural Information Processing Systems*, volume 17, 2004.

[13] Hung Hung and Chen-Chien Wang. Matrix variate logistic regression model with application to EEG data. *Biostatistics*, 14(1):189–202, 2013.

[14] Xu Tan, Yin Zhang, Siliang Tang, Jian Shao, Fei Wu, and Yueting Zhuang. Logistic tensor regression for classification. In *Intelligent Science and Intelligent Data Engineering*, pages 573–581. Springer, 2013.

[15] Yangyang Xu. Alternating proximal gradient method for sparse nonnegative Tucker decomposition. *Mathematical Programming Computation*, pages 1–32, 2013.

[16] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The Lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

[17] Fei Wang, Ping Zhang, Buyue Qian, Xiang Wang, and Ian Davidson. Clinical risk prediction with multilinear sparse logistic regression. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014.

[18] David W Hosmer Jr and Stanley Lemeshow. *Applied logistic regression*. John Wiley & Sons, 2004.

[19] Shuo Xiang, Lei Yuan, Wei Fan, Yalin Wang, Paul M Thompson, and Jieping Ye. Multi-source learning with block-wise missing data for Alzheimer's disease prediction. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 185–193. ACM, 2013.

[20] Jun Liu, Jianhui Chen, and Jieping Ye. Large-scale sparse logistic regression. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 547–556. ACM, 2009.

[21] Tapio Manninen, Heikki Huttunen, Pekka Ruusuvuori, and Matti Nykter. Leukemia prediction using sparse logistic regression. *PloS one*, 8(8), 2013.

[22] Anil Rao, Ying Lee, Achim Gass, and Andreas Monsch. Classification of Alzheimer's disease from structural MRI using sparse logistic regression with optional spatial regularization. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 4499–4502. IEEE, 2011.

[23] Yongdai Kim, Sunghoon Kwon, and Seuck Heun Song. Multiclass sparse logistic regression for classification of multiple cancer types using gene expression data. *Computational Statistics & Data Analysis*, 51(3):1643–1655, 2006.

[24] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[25] William Thies and Laura Bleiler. 2013 Alzheimer's disease facts and figures. *Alzheimer's & dementia: the journal of the Alzheimer's Association*, 9(2):208–245, 2013.

[26] Scott A Huettel, Allen W Song, and Gregory McCarthy. *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland, MA, 2004.

[27] Buyue Qian, Xiang Wang, Fei Wang, Hongfei Li, Jieping Ye, and Ian Davidson. Active learning from relative queries. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 1614–1620, 2013.

[28] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[29] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[30] Paul A Heidenreich, Justin G Trogdon, Olga A Khavjou, Javed Butler, Kathleen Dracup, Michael D Ezekowitz, Eric Andrew Finkelstein, Yuling Hong, S Claiborne Johnston, Amit Khera, et al. Forecasting the future of cardiovascular disease in the United States a policy statement from the American heart association. *Circulation*, 123(8):933–944, 2011.

[31] Gregory C Pope, Randall P Ellis, Arlene S Ash, JZ Ayanian, DW Bates, H Burstin, LI Iezzoni, E Marcantonio, and B Wu. Diagnostic cost group hierarchical condition category models for Medicare risk adjustment. *Health Economics Research, Inc. Waltham, MA*, 2000.

[32] Frans H Rutten, Maarten-Jan M Cramer, Jan-Willem J Lammers, Diederick E Grobbee, and Arno W Hoes. Heart failure and chronic obstructive pulmonary disease: an ignored combination? *European journal of heart failure*, 8(7):706–711, 2006.

[33] Ali Ahmed, Michael W Rich, Paul W Sanders, Gilbert J Perry, George L Bakris, Michael R Zile, Thomas E Love, Inmaculada B Aban, and Michael G Shlipak. Chronic kidney disease associated mortality in diastolic versus systolic heart failure: a propensity matched study. *The American journal of cardiology*, 99(3):393–398, 2007.

[34] Patrick A McKee, William P Castelli, Patricia M McNamara, and William B Kannel. The natural history of congestive heart failure: the Framingham study. *New England Journal of Medicine*, 285(26):1441–1446, 1971.

[35] Matthew M Davis, Kathryn Taubert, Andrea L Benin, David W Brown, George A Mensah, Larry M Baddour, Sandra Dunbar, and Harlan M Krumholz. Influenza vaccination as secondary prevention for cardiovascular disease: a science advisory from the american heart association/american college of cardiology. *Journal of the American College of Cardiology*, 48(7):1498–1502, 2006.