Research Article

Coordinate Descent Based Hierarchical Interactive Lasso Penalized Logistic Regression and Its Application to Classification Problems

Jin-Jia Wang and Yang Lu

School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China

Correspondence should be addressed to Jin-Jia Wang; wjj@ysu.edu.cn

Received 20 August 2014; Revised 1 December 2014; Accepted 1 December 2014; Published 16 December 2014

Academic Editor: Wei-Chiang Hong

Copyright © 2014 J.-J. Wang and Y. Lu. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We present the hierarchical interactive lasso penalized logistic regression using the coordinate descent algorithm based on the hierarchy theory and variables interactions. We define the interaction model based on the geometric algebra and hierarchical constraint conditions and then use the coordinate descent algorithm to solve for the coefficients of the hierarchical interactive lasso model. We provide the results of some experiments based on UCI datasets, Madelon datasets from NIPS2003, and daily activities of the elder. The experimental results show that the variable interactions and hierarchy contribute significantly to the classification. The hierarchical interactive lasso has the advantages of the lasso and interactive lasso.

1. Introduction

Sparse linear models (such as the lasso) are a remarkable success of the regression analysis of high-dimensional data [1]. The lasso is a least squares regression with the L1 penalty function. It can also be extended to the generalized linear model [2], for example, the logistic regression with L1 penalty used for classification [3]. In the lasso model, the response variable is assumed to be a linear weighted sum of the predictor variables, and the optimization problem used to find the weighting coefficients can be solved by the coordinate descent algorithm [4]. If, in the analysis of high-dimensional data, the response variable cannot be explained by a linear weighted sum of predictor variables, a higher-order model and quadratic model need to be used. In most cases, this suggests the presence of variable interactions [5]. The presence of such interactions is considered important, as, for example, the interaction between single nucleotide polymorphisms (SNPs) plays an important role in the diagnosis of cancer and other diseases [6]. While the linear model has some advantages, such as good interpretability and simple calculations, the variable interaction models are considered to be a focus of the modern research [7].

There are three types of methods used in the hierarchical interaction models. The first one is a multistep method. This

method is based on removing or adding the best predictor variables or interaction variables in each iteration. Once the predictor variables corresponding to the interaction variables are in the model, the interaction variables must be in the model as well [8]. Alternatively, we can consider the variable selection before the interaction selection [9]. Usually, the modified LARS algorithm is used in such models to solve the interaction model [10]. The second type is the Bayes model method. This approach improves the random search variable selection method for the hierarchical interaction model [11]. The third type is based on optimization. The sparse interaction model is formulated as a nonconvex optimization problem [12] and further expressed as a convex optimization problem, such as all-pair lasso [13] or interaction group lasso [14].

In the literature on the sparse structures [15], composite absolute penalties (CAP) can also obtain the sparseness of the group and interaction, but the interaction coefficient is penalized twice [16]. To solve for the hierarchical sparseness in the nonlinear interaction problem, the existing literature [17] has introduced the VANISH method. The logic regression method considers the binary variable high-level interaction [18]. The existing literature [19] uses a simple recursive approach to select the interaction variables from high-dimensional data. The literature [20] proposed a genetic

FIGURE 1: The diagram of the subspaces in geometric algebra.

algorithm using selection to choose interaction variables in high-dimensional data.

The literature [13] presents a hierarchical interactive lasso method for regression and provides a method of model coefficients estimation using KKT conditions and the Lagrange multiplier method. Based on the literature [13] and our past work, the authors propose the concept of geometric algebra interaction and coordinate descent algorithm for the hierarchical interactive lasso penalized logistic regression. We used experimental data including 4 kinds of datasets from the UCI machine learning database, one Madelon datasets from NIPS2003, and one daily life activity recognition datasets. The experimental results reveal the outstanding advantages of the hierarchical interactive lasso method compared to the lasso and interactive lasso methods. The innovations include the following. (1) We use geometric algebra to explain variable interaction; (2) we derive an improved coordinate descent algorithm to solve the hierarchical interactive lasso penalized logistic regression; (3) we use the hierarchical interactive lasso for the classification problem.

2. The Variable Interaction Theory of Geometric Algebra

Definition 1. If the function $f(x, y)$ cannot be represented as a sum of independent functions, $f_1(x) + f_2(y)$, then x, y in the f function are said to have interaction.

A popular explanation of Definition 1 is that if a response variable cannot be represented as a linear weighted sum of the prediction variables, it is probably because there are interactions between the variables.

Interactions between variables can be easily explained by the geometric algebra theory. Figure 1 is a diagram showing all subspace in geometric algebra. The 1-vectors, namely, order-1 main variables, can represent a p dimensional subspace of the original data. That is, p dimensional base of the original data is projected on the 1-vectors. The 2-vectors show the interaction between two variables. The simplest 2-vectors coefficient can be the product of two 1-vectors. In the literature [13] our proposed area feature is considered as one of the interactions. In the literature [20] our proposed orthocenter feature is considered as one of the interactions. Higher-order interactions are represented by k -vectors. In this paper, we only study area interactions between 1-vectors.

This method can also be extended to the nonlinear complex function interactions or higher order.

3. The Binary Logistic Regression Based on Interaction and Hierarchy

The outcome variable in the binary logistic model is denoted by Y ; the income variables are predictors X . $x_1, \dots, x_j, \dots, x_p$ are order-1 main variables, and the pairwise $x_j x_k$ are interactions variables between order-1 main variables. The binary logistic model has the form

$$\log it (P(Y = 1 | X)) = \sum_{j=0}^p \beta_j x_j + \frac{1}{2} \sum_{j \neq k} \Theta_{jk} x_j x_k + \varepsilon, \quad (1)$$

where $\Theta_{jj} = 0$, the main variables coefficients are $\beta \in R^{p+1}$, the interaction variables coefficients are $\Theta \in R^{p \times p}$, x_0 is 1, and ε satisfies $N(0, \sigma^2)$.

Assume that the training samples are $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_i, y_i), \dots, (\mathbf{x}_N, y_N)$, $\mathbf{x}_i \in R^p$,

$$y_i = \begin{cases} 1 & Y_i = 1 \\ 0 & Y_i \neq 1. \end{cases} \quad (2)$$

Our goal is to select a feature subset from the order-1 main variables (dimension p) and order-2 interaction variables (dimension $p(p-1)/2$). We then estimate the coefficient values for nonzero model parameters. We can obtain the probability of two classes as follows:

$$\begin{aligned} \Pr(Y = 1 | \mathbf{x}_i) &= \frac{1}{1 + \exp(-\sum_j \beta_j x_{ij} - (1/2) \sum_{j \neq k} \Theta_{jk} x_{ij} x_{ik})} \\ &= p(\mathbf{x}_i), \\ P(Y = 2 | \mathbf{x}_i) &= \frac{1}{1 + \exp(\sum_j \beta_j x_{ij} + (1/2) \sum_{j \neq k} \Theta_{jk} x_{ij} x_{ik})} \\ &= 1 - p(\mathbf{x}_i). \end{aligned} \quad (3)$$

The maximum likelihood estimation is used to estimate the unknown model parameters, which make the likelihood function of N independent observations the largest. We define

$$L(\beta, \Theta) = \prod_{i=1}^N p(\mathbf{x}_i)^{y_i} [1 - p(\mathbf{x}_i)]^{(1-y_i)}. \quad (4)$$

Then, the logarithmic likelihood function of (4) is

$$\begin{aligned} \frac{1}{N} \ln [L(\beta, \Theta)] &= \frac{1}{N} \sum_{i=1}^N \left[y_i \left(\mathbf{x}_i^T \beta + \frac{1}{2} \mathbf{x}_i^T \Theta \mathbf{x}_i \right) + \ln (1 - p(\mathbf{x}_i)) \right]. \end{aligned} \quad (5)$$

We use the second-order Taylor expansion at the current estimated value $(\tilde{\beta}, \tilde{\Theta})$ for (5) and obtain the subproblem as follows:

$$l_Q(\beta, \Theta) = \frac{1}{2} \omega_i \left(z_i - \sum_j \tilde{\beta}_j x_{ij} - \frac{1}{2} \sum_{j \neq k} \tilde{\Theta}_{jk} x_{ij} x_{ik} \right)^2 + \frac{(y_i - \tilde{p}(\mathbf{x}_i))^2}{2 * \tilde{p}^2(\mathbf{x}_i) [1 - \tilde{p}(\mathbf{x}_i)]^2} + \left[y_i \left(\sum_j \tilde{\beta}_j x_{ij} + \sum_{j \neq k} \tilde{\Theta}_{jk} x_{ij} x_{ik} \right) + \ln(1 - \tilde{p}(\mathbf{x}_i)) \right], \quad (6)$$

where $\omega_i = \tilde{p}(\mathbf{x}_i)[1 - \tilde{p}(\mathbf{x}_i)]$, $z_i = \sum_j \beta_j x_{ij} + (1/2) \sum_{j \neq k} \Theta_{jk} x_{ij} x_{ik} + (y_i - \tilde{p}(\mathbf{x}_i))/\tilde{p}(\mathbf{x}_i)[1 - \tilde{p}(\mathbf{x}_i)]$.

The proof that (5) implies (6) is presented in the Appendix A.

In order to obtain the sparse solution for the main variable coefficients and interaction coefficients, the penalty function is used to enhance the stability of the interactive model:

$$(\beta, \Theta) = \arg \min_{\beta \in R^{p+1}, \Theta \in R^{p \times p}} -l_Q(\beta, \Theta) + \lambda_1 \|\beta\|_1 + \lambda_2 \|\Theta\|_1. \quad (7)$$

We focus on those interactions that have large main variable coefficients. Such restrictions are known as “hierarchy.” The mathematical expression for them is $\tilde{\Theta}_{jk} \neq 0 \Rightarrow \tilde{\beta}_j \neq 0$ or $\tilde{\beta}_k \neq 0$. So, we add the constraints enforcing the hierarchy into (7) as follows:

$$\begin{aligned} &(\beta, \Theta) \\ &= \min_{\beta \in R^{p+1}, \Theta \in R^{p \times p}} -l_Q(\beta, \Theta) + \lambda_1 \|\beta\|_1 + \lambda_2 \|\Theta\|_1 \quad (8) \\ &\text{s.t. } \|\Theta_j\|_1 \leq |\beta_j| \quad \text{for } j = 1, \dots, p, \end{aligned}$$

where Θ_j is the j th column of Θ . If $\tilde{\Theta}_{jk} \neq 0$, then $\|\tilde{\Theta}_j\|_1 > 0$, $\|\tilde{\Theta}_k\|_1 > 0$, so $\tilde{\beta}_j \neq 0$, and $\tilde{\beta}_k \neq 0$. The new constraint guarantees the hierarchy, but we cannot obtain a convex solution because (8) is not convex. So, instead of β we use β^+, β^- . And the corresponding convex relaxation of (8) is as follows:

$$\begin{aligned} &(\beta^+, \Theta) = \min_{\beta^+ \in R^p, \beta^- \in R^p, \Theta \in R^{p \times p}} -l_Q(\beta^+ - \beta^-, \Theta) \\ &\quad + \lambda_1 1^T (\beta^+ + \beta^-) + \lambda_2 \sum_j \|\Theta_j\|_1 \quad (9) \\ &\text{s.t. } \left\{ \begin{array}{l} \|\Theta_j\|_1 \leq \beta_j^+ + \beta_j^- \\ \beta_j^+ \geq 0, \quad \beta_j^- \geq 0 \end{array} \right\} \quad \text{for } j = 1, \dots, p, \end{aligned}$$

where $\beta = \beta^+ - \beta^-$, $\beta^+ = \max\{\pm\beta, 0\}$, $\beta^+, \beta^- \in R^{p+1}$, and $\|\beta\|_1 = \beta^+ + \beta^-$.

4. Coordinate Descent Algorithm and KKT Conditions

The basic idea of the coordinate descent algorithm is to convert multivariate problems into multiple single variable subproblems. It allows optimizing only one-dimensional variables at a time. The solution can be updated in a cycle. We solve (9) using the coordinate descent algorithm.

The Lagrange function corresponding to (9) is as follows:

$$\begin{aligned} L(\beta^+, \beta^-, \Theta) = &-l_Q(\beta^+ - \beta^-, \Theta) \\ &+ (\lambda_1 1 - \hat{\alpha} - \gamma^+ 1)^T \beta^+ \\ &+ (\lambda_1 1 - \hat{\alpha} - \gamma^- 1)^T \beta^- \\ &+ \langle \text{diag}(\lambda_2 1 + \hat{\alpha}) U, \Theta \rangle, \end{aligned} \quad (10)$$

where

$$\hat{U}_{jk} = \begin{cases} \in [-1, 1] & \hat{\Theta}_{kj} = 0 \\ \text{sign}(\hat{\Theta}_{jk}) & \hat{\Theta}_{jk} \neq 0, \end{cases} \quad (11)$$

and $\hat{\alpha}$ and $\gamma^\pm$ are the dual variables corresponding to the hierarchical constraint and the nonnegative constraints. Formula (10) can be decomposed into p subproblems:

$$\begin{aligned} L(\beta_j^+, \beta_j^-, \Theta_j) = &-l_Q(\beta_j^+ - \beta_j^-, \Theta_j) \\ &+ (\lambda_1 - \hat{\alpha}_j - \gamma^+)^T \beta_j^+ \\ &+ (\lambda_1 - \hat{\alpha}_j - \gamma^-)^T \beta_j^- \\ &+ \langle (\lambda_2 + \alpha) U_j, \Theta_j \rangle. \end{aligned} \quad (12)$$

The solution of (12) as a convex problem can be obtained by a set of optimality conditions, known as the KKT (Karush-Kuhn-Tucker) conditions. This is the key advantage of our approach.

The stationary conditions of (12) according to KKT are $\partial L / \partial \hat{\beta}_j^\pm = 0$, $\partial L / \partial \hat{\Theta}_{jk} = 0$. The complementary conditions are $\|\hat{\Theta}_j\|_1 \leq \hat{\beta}_j^+ + \hat{\beta}_j^-$, $\gamma_j^\pm \hat{\beta}_j^\pm = 0$, $\hat{\beta}^\pm \geq 0$, $\hat{\gamma}^\pm \geq 0$, $\hat{\alpha} \geq 0$, $\hat{\alpha}_j (\|\hat{\Theta}_j\|_1 - \hat{\beta}_j^+ - \hat{\beta}_j^-) = 0$. We assume that $(1/N) \sum_{i=1}^N x_{ij}^2 = 1$. For our problem, the KKT conditions can be written as follows:

$$\begin{aligned} \hat{\beta}_j^+ - \hat{\beta}_j^- &= S \left[x_{ij} (z_i - A_i) + \hat{\beta}_j^+ - \hat{\beta}_j^-, \frac{\lambda_1 + \hat{\alpha}_j}{\omega_i} \right], \\ \hat{\Theta}_{jk} &= \frac{S \left[(x_{ij} x_{ik})^T (z_i - A_i + \hat{\Theta}_{jk} x_{ij} x_{ik}), 2(\lambda_2 + \hat{\alpha}_j) / \omega_i \right]}{\|x_{ij} x_{ik}\|_2}, \end{aligned} \quad (13)$$

where $A_i = \sum_j \tilde{\beta}_j x_{ij} + (1/2) \sum_{j \neq k} \tilde{\Theta}_{jk} x_{ij} x_{ik}$, z_i, ω_i are from (6), and S denotes the soft-threshold operator defined by $S(c, \lambda) = \text{sign}(c)(|c| - \lambda)_+$.

TABLE 1: Datasets information.

	Datasets	Samples	Variable dimension	Interactive dimension	Class
1	Breast-cancer-Wisconsin	683	9	36	2
2	Ionosphere	351	33	528	2
3	Liver_disorders	345	6	15	2
4	Sonar	208	60	1770	2

TABLE 2: The experimental results of our method.

Datasets	Coefficients	Error rate (%)	SD	Time (/s)	λ
Breast-cancer-Wisconsin	22	3.0	0.01	9.49	3.14
Ionosphere	101	28.0	0.02	124.63	2.82
Liver_disorders	25	26.0	0.02	12.11	2.37
Sonar	117	14.0	0.02	150.69	0.73

The proof of both expressions (13) can be found in Appendix B.

Now, we define $f(\hat{\alpha}_j) = \|\hat{\Theta}_j\|_1 - \hat{\beta}_j^+ - \hat{\beta}_j^-$. Then

$$\begin{aligned}
 f(\hat{\alpha}_j) &= \left\| \sum_{k=1}^P S \left[(x_{ij}x_{ik})^T (z_i - A_i + \hat{\Theta}_{jk}x_{ij}x_{ik}), \right. \right. \\
 &\quad \left. \left. \frac{2(\lambda_2 + \hat{\alpha}_j)}{\omega_i} \right] \right\|_1 \\
 &\quad \cdot \left(\|x_{ij}x_{ik}\|_2^2 \right)^{-1} \\
 &\quad - \left\| S \left[x_{ij} (z_i - A_i) + \hat{\beta}_j^+ - \hat{\beta}_j^-, \frac{\lambda_1 + \hat{\alpha}_j}{\omega_i} \right] \right\|_1.
 \end{aligned} \tag{14}$$

The remaining KKT conditions only involve $\hat{\alpha} : \hat{\alpha} f(\hat{\alpha}) = 0$, $f(\hat{\alpha}) \leq 0$, $\hat{\alpha} \geq 0$. Observing that f is nonincreasing with respect to $\hat{\alpha}$ and is piecewise linear, it is easy to get the solution for $\hat{\alpha}$.

In conclusion, the overall idea of the coordinate descent algorithm is that the minimization of (9) is equivalent to the minimization of (10). Formula (10) can be decomposed into p independent formula (12). Formula (12) can be solved as (13). The final coefficient optimization iteration formula is as follows:

$$\begin{aligned}
 &\hat{\beta}_j^{+(m+1)} - \hat{\beta}_j^{-(m+1)} \\
 &= S \left[x_{ij} (z_i^m - A_i^m) + \hat{\beta}_j^{+(m)} - \hat{\beta}_j^{-(m)}, \frac{\lambda_1 + \hat{\alpha}_j^m}{\omega_i^k} \right], \\
 &\hat{\Theta}_{jk}^{m+1} \\
 &= \frac{S \left[(x_{ij}x_{ik}) (z_i^m - A_i^m + \hat{\Theta}_{jk}^m x_{ij}x_{ik}), 2(\lambda_2 + \hat{\alpha}_j^m) / \omega_i^m \right]}{\|x_{ij}x_{ik}\|_2},
 \end{aligned} \tag{15}$$

where $\hat{\beta}_j^{+(m)} - \hat{\beta}_j^{-(m)}$ is the estimated value of the j th main variable coefficient after m iterations and $\hat{\Theta}_{jk}^m$ is the estimation of the interaction coefficient between the j th variable and the k th variable after m iterations.

5. The Experimental Results and Analysis

5.1. The Experimental Results and Analysis of Four UCI Datasets. There are four UCI study database datasets, which include the breast-cancer-Wisconsin datasets, Ionosphere datasets, Liver_disorders datasets, and Sonar datasets, as shown in Table 1.

We do the 10-fold cross-validation (10-CV) experiments in the paper for 20 times using R, where $\lambda_1 = 2\lambda_2 = \lambda$. Besides we complete an experiment employing the interactive hierarchical lasso logistic regression method. The results include the number of nonzero variable coefficients, average error rate of the 10-CV, standard deviation (SD), CPU time, and the value of lambda (λ) estimated. The results are shown in Table 2.

The results of the 10-CV based on the four datasets are presented in Figures 2 to 5 using the proposed method. In the figures, the horizontal axis represents the logarithmic value of λ and the vertical axis is the error rate of the 10-CV. Besides, the horizontal axis at the top of each figure represents the number of nonzero variable coefficients corresponding to the λ value.

The results for the breast-cancer-Wisconsin datasets are shown in Figure 2. The minimum error rate is 0.03 and the number of selected variables is more than 11. The results for the Ionosphere experimental datasets are shown in Figure 3. When the number of selected variables is 101, the lowest error rate is 0.28 with the smaller standard deviation. The number of the selected variables is larger than the original dimension, so the interactions provide the classified information. The results for the Liver_disorders datasets are presented in Figure 4. If the number of selected variables is 25, the lowest error rate can reach 0.26, while the standard deviation is 0.02. Finally, the results for the Sonar datasets are presented in Figure 5. When more than 80 variables are selected, the minimum error rate is 0.14.

FIGURE 2: The results of breast-cancer-Wisconsin.

FIGURE 4: The results of Liver_disorders.

FIGURE 3: The results of Ionosphere.

FIGURE 5: The results of Sonar.

In what follows, we compare our method to the existing literature [13]. The classification results and training time of our method are better than those shown in the literature [13]. The experimental results of the lasso, all-pair lasso, and conventional pattern recognition methods with 10-fold cross-validation of 20 times in the four UCI datasets are listed, respectively, in Tables 3, 4, and 5. Conventional pattern recognition methods include support vector machine (SVM), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), K -nearest neighborhood (K -NN), and decision tree (DT) methods. The lasso is a method that considers the main variables without the interaction variables. All-pair lasso is a method that considers the main variables and interaction variables but without the hierarchy. The experimental

TABLE 3: The experimental results of the lasso penalized logistic regression model.

Datasets	Error rate (%)	SD	Time (/s)
Breast-cancer-Wisconsin	3.22	0.0001	0.3744
Ionosphere	31.15	0.0075	0.4863
Liver_disorders	30.77	0.0077	0.0908
Sonar	22.13	0.0121	15.6806

results show that our model is better in classification results and more stable. This highlights the advantage of the variable interactions and hierarchy.

5.2. The Experimental Results for High-Dimensional Small Sample Data. The Madelon datasets from NIPS2003 was

TABLE 4: The experimental results of the all-pair lasso penalized logistic regression model.

Datasets	Error rate (%)	SD	Time (/s)
Breast-cancer-Wisconsin	2.93	0.0001	12.1213
Ionosphere	27.00	0.0089	3.1734
Liver_disorders	26.36	0.0058	4.9190
Sonar	20.63	0.0115	4.1041

used to evaluate our method. The sample numbers of the training, validation, and testing of the data were, respectively, 2000, 600, and 1800. The class number is 2. The variable dimension is 500. So the interactive dimension is 124750. You can find more information about the datasets following the link <http://www.nipsfsc.ecs.soton.ac.uk/>; you can also download the datasets and see the results of challenges, balance error rates, and the area under the curve. The model is trained by using a training set. The model parameters are selected by using a validation set. Also, the prediction results of the final model using the test set are uploaded online. The classification score of the final model is obtained. Our results are shown in Table 6. The results show that our method is slightly better than the lasso and all-pair lasso. This implies that the interactions may also be important in the Madelon datasets.

5.3. Activity Recognition (AR) Using Inertial Sensors of Smartphones. Anguita et al. collected sensor data of smartphones [10]. They used the support vector machine (SVM) method to solve the classification problem of the daily life activity recognition. These results play an extremely significant role in disability and elderly care. Datasets can be downloaded following the literature [10]. 30 volunteers aged 19–48 years participated in the study. Each person performed six activities wearing the smartphone on the waist. To obtain the data class label, experiments were conducted using video recording. The smartphone used in the experiments had built-in accelerometer and gyroscope for measuring 3D linear acceleration and angular acceleration. The sampling frequency was 50 Hz, which is more than enough for capturing human movements.

We use the datasets to evaluate our method. We use the upstairs and downstairs movements as two active classes. The training sets have 986 samples and 1,073 samples, respectively. The test sets have 420 samples and 471 samples, respectively. The variable dimension is 561, which includes the time and frequency from sensor signals.

Experimental results of the three lasso methods and some pattern recognition methods are shown in Table 7. The results show that our method is better than the pattern recognition methods, since it takes the variable selection and interaction into account. Our method achieves the best classification results with less training and testing time.

5.4. The Numerical Simulation Results and Discussion. Now, suppose that the number and dimension of the samples are

$n = 200$, $p = 20$. We take interactions into consideration and provide the following three kinds of simulation based on formula (1).

- (1) The real model is hierarchical: $\Theta_{jk} \neq 0 \Rightarrow \beta_j \neq 0$ or $\beta_k \neq 0$, $j, k = 1, \dots, p$. There are 10 nonzero elements in β and 20 nonzero elements in Θ .
- (2) The real model only includes interactive variables: $\beta_j = 0$, $j = 1, \dots, p$. There are 20 nonzero elements in Θ .
- (3) The real model only includes main variables: $\Theta_{jk} = 0$, $j, k = 1, \dots, p$. There are 10 nonzero elements in β .

The SNR of the main variables is 1.5, and SNR of the interaction variables is 1. The experiment results of 100 times are shown in Figure 6.

When the real model is hierarchical, our method is the best, and the lasso is the worst. This is shown in Figure 6(a). When the real model only includes interaction variances, the interactive lasso is the best, and our method takes the second place, while the lasso is still the worst, as shown in Figure 6(b). The reason for this result is that when our method fits the model, the interaction variables are considered to be main variables. When the real model only includes main variables, the lasso is the best, and our method still takes the second place, and the all-pair lasso is the worst, as shown in Figure 6(c).

We believe that many actual classification problems could be hierarchical and interactive. They contain both main variables and interaction variables. Our method fits in this kind of situation.

6. Conclusion

Taking into consideration the interaction between variables, the hierarchical interactive lasso penalized logistic regression using the coordinate descent algorithm is derived. We provide the model definition, constraint condition, and the convex relaxation condition for the model. We obtain a solution for the coefficients of the proposed model based on the convex optimization and coordinate descent algorithm. We further provide experimental results based on four UCI datasets, NIPS2003 feature selection challenge datasets, and true daily life activities identification datasets. The results show that the interaction widely exists in the classification models. They also demonstrate that the variable interaction contributes to the response. The classification performance of our method is superior to the lasso, all-pair lasso, and some pattern recognition methods. It turns out that the variable interaction and hierarchy are two important factors. Our further research is planned as follows: other convex optimization methods including the generalized gradient descent method or alternating direction multiplier method, the hierarchical interactive lasso penalized multiclass logistic regression method, the elastic net method, or the hierarchical group lasso method. The application of the multisensor interaction in the daily life activities of the elderly is a new way of using of our method.

TABLE 5: The experimental results of the traditional pattern recognition methods.

Datasets	Classifier	Error rate (%)	SD	Time (/s)
Breast-cancer-Wisconsin	SVM	3.42	0.0019	1.3092
	LDA	3.96	0.0007	0.2429
	QDA	4.92	0.0014	0.2309
	K-NN	4.61	0.0030	0.6636
	DT	4.99	0.0041	1.6201
Ionosphere	SVM	35.83	0.0063	2.9149
	LDA	33.10	0.0090	0.6168
	QDA	32.01	0.0076	0.4516
	K-NN	28.30	0.0090	0.6363
	DT	38.23	0.0408	3.0865
Liver_disorders	SVM	30.80	0.0100	1.1401
	LDA	31.31	0.0089	0.2474
	QDA	40.19	0.0119	0.1958
	K-NN	36.89	0.0124	0.6089
	DT	36.94	0.0187	1.9921
Sonar	SVM	25.70	0.0197	1.2181
	LDA	25.13	0.0145	0.5897
	QDA	23.80	0.0205	0.4594
	K-NN	13.09	0.0101	0.5345
	DT	35.91	0.0303	0.7153

TABLE 6: The experimental results for the Madelon datasets.

Methods	Balance error rate (%)			Area under the curve (%)		
	Training	Validation	Testing	Training	Validation	Testing
Lasso	38.08	37.21	36.61	64.78	62.79	63.40
All-pair lasso	34.98	37.01	36.57	65.02	62.99	63.43
Our method	30.00	36.35	35.12	70.00	68.08	67.76

Appendices

A. Proofs from (5) to (6)

For notational convenience, we have written p_i instead of $p(x_i)$. A logarithmic likelihood function of (4) is as follows:

$$\begin{aligned}
 L_Q &= \frac{1}{N} \ln [L(\beta, \Theta)] \\
 &= \frac{1}{N} \sum_{i=1}^N \left[y_i \left(x_i^T \beta + \frac{1}{2} x_i^T \Theta x_i \right) + \ln (1 - p_i) \right]. \quad (\text{A.1})
 \end{aligned}$$

First, we give the first- and second-order partial derivative and mixed partial derivative of (4) with respect to β and Θ :

$$\begin{aligned}
 \frac{\partial L_Q}{\partial \beta} &= \frac{\partial (1/N) \sum_{i=1}^N \left[y_i \left(x_i^T \beta + (1/2) x_i^T \Theta x_i \right) + \ln (1 - p_i) \right]}{\partial \beta} \\
 &= \frac{1}{N} \sum_{i=1}^N \left[\left(x_i^T y_i \right)^T - \frac{\exp \left(x_i^T \beta + (1/2) x_i^T \Theta x_i \right)}{1 + \exp \left(x_i^T \beta + (1/2) x_i^T \Theta x_i \right)} \cdot x_i \right]
 \end{aligned}$$

$$= \frac{1}{N} \sum_{i=1}^N (y_i^T x_i - p_i^T x_i) = \frac{1}{N} \sum_{i=1}^N x_i (y_i - p_i),$$

$$\begin{aligned}
 \frac{\partial^2 L_Q}{\partial \beta^2} &= \frac{\partial}{\partial \beta} \frac{1}{N} \sum_{i=1}^N x_i^T (y_i - p_i) = -\frac{1}{N} \sum_{i=1}^N x_i^T \frac{\partial p_i}{\partial \beta} \\
 &= \frac{1}{N} \sum_{i=1}^N x_i^T \frac{\exp \left(-x_i^T \beta - (1/2) x_i^T \Theta x_i \right)}{[1 + \exp \left(-x_i^T \beta - (1/2) x_i^T \Theta x_i \right)]^2} \cdot (-x_i) \\
 &= -\frac{1}{N} \sum_{i=1}^N x_i \frac{\exp \left(x_i^T \beta - (1/2) x_i^T \Theta x_i \right)}{1 + \exp \left(-x_i^T \beta - (1/2) x_i^T \Theta x_i \right)} \\
 &\quad \cdot \frac{1}{1 + \exp \left(-x_i^T \beta - (1/2) x_i^T \Theta x_i \right)} \cdot x_i^T \\
 &= -\frac{1}{N} \sum_{i=1}^N (1 - p_i) \cdot p_i \cdot x_i^T \cdot x_i,
 \end{aligned}$$

$$\frac{\partial L_Q}{\partial \Theta}$$

TABLE 7: The experimental results of the three lasso methods and five traditional pattern recognition methods.

Methods	Error rate in training (%)	Training time (/s)	Error rate in testing (%)	Testing time (/s)
Lasso	0.00	0.2649	1.46	0.0312
All-pair lasso	0.00	0.8807	1.35	0.0953
Our method	0.00	0.6012	1.12	0.0556
SVM	0.00	0.9516	1.23	0.0936
1-NN	0.00	0.0000	9.20	0.7956
3-NN	0.00	0.0000	8.98	0.7488
QDA	6.61	1.4196	4.04	0.2184
DT	0.00	0.7332	14.93	0.1092

FIGURE 6: The error rate of the three lasso methods in the simulation experiment (the Bayes error is shown by the purple dotted line in the graph).

$$\begin{aligned}
&= \frac{\partial (1/N) \sum_{i=1}^N [y_i (x_i^T \beta + (1/2) x_i^T \Theta x_i) + \ln(1 - p_i)]}{\partial \Theta} \\
&= \frac{1}{N} \sum_{i=1}^N \left[\left(\frac{1}{2} y_i x_i^T x_i \right)^T - \frac{\exp(x_i^T \beta + (1/2) x_i^T \Theta x_i)}{1 + \exp(x_i^T \beta + (1/2) x_i^T \Theta x_i)} \right. \\
&\quad \left. \cdot \frac{1}{2} \cdot x_i^T x_i \right] \\
&= \frac{1}{N} \sum_{i=1}^N \left(\frac{1}{2} y_i x_i^T x_i - \frac{1}{2} p_i x_i^T x_i \right) \\
&= \frac{1}{2N} \sum_{i=1}^N x_i^T x_i (y_i - p_i), \\
&\frac{\partial^2 L_Q}{\partial \Theta^2} \\
&= \frac{\partial}{\partial \Theta} \frac{1}{2N} \sum_{i=1}^N x_i^T x_i (y_i - p_i) = -\frac{1}{2N} \sum_{i=1}^N x_i^T x_i \frac{\partial p_i}{\partial \Theta} \\
&= \frac{1}{2N} \sum_{i=1}^N \frac{\exp(-x_i^T \beta - (1/2) x_i^T \Theta x_i)}{[1 + \exp(-x_i^T \beta - (1/2) x_i^T \Theta x_i)]^2} \\
&\quad \cdot \left(-\frac{1}{2} x_i^T x_i \right) \cdot x_i^T x_i \\
&= -\frac{1}{4N} \sum_{i=1}^N \frac{\exp(x_i^T \beta - (1/2) x_i^T \Theta x_i)}{1 + \exp(-x_i^T \beta - (1/2) x_i^T \Theta x_i)} \\
&\quad \cdot \frac{1}{1 + \exp(-x_i^T \beta - (1/2) x_i^T \Theta x_i)} \cdot \|x_i\|^2 \\
&= -\frac{1}{4N} \sum_{i=1}^N (1 - p_i) \cdot p_i \cdot \|x_i\|^2, \\
&\frac{\partial^2 L_Q}{\partial \beta \partial \Theta} \\
&= \frac{\partial}{\partial \beta} \frac{1}{N} \sum_{i=1}^N x_i^T (y_i - p_i) = -\frac{1}{N} \sum_{i=1}^N x_i^T \frac{\partial p_i}{\partial \Theta} \\
&= \frac{1}{N} \sum_{i=1}^N x_i^T \frac{\exp(-x_i^T \beta - (1/2) x_i^T \Theta x_i)}{[1 + \exp(-x_i^T \beta - (1/2) x_i^T \Theta x_i)]^2} \\
&\quad \cdot \left(-\frac{1}{2} x_i^T x_i \right) \\
&= -\frac{1}{N} \sum_{i=1}^N x_i^T \frac{\exp(x_i^T \beta - (1/2) x_i^T \Theta x_i)}{1 + \exp(-x_i^T \beta - (1/2) x_i^T \Theta x_i)} \\
&\quad \cdot \frac{1}{1 + \exp(-x_i^T \beta - (1/2) x_i^T \Theta x_i)} \\
&= -\frac{1}{2N} \sum_{i=1}^N (1 - p_i) \cdot p_i \cdot x_i^T \cdot x_i \cdot x_i^T.
\end{aligned}$$

(A.2)

Then (A.1) is expanded by using Taylor series with respect to the expended point $(\tilde{\beta}, \tilde{\Theta})$:

$$\begin{aligned}
&l_Q(\beta, \Theta) \\
&= L_Q + (\beta - \tilde{\beta}) \cdot \frac{\partial L_Q}{\partial \tilde{\beta}} + (\Theta - \tilde{\Theta}) \cdot \frac{\partial L_Q}{\partial \tilde{\Theta}} \\
&\quad + \frac{1}{2} \left[(\beta - \tilde{\beta}) \cdot \frac{\partial L_Q}{\partial \tilde{\beta}^2} + 2(\beta - \tilde{\beta})^T (\Theta - \tilde{\Theta}) \frac{\partial^2 L_Q}{\partial \tilde{\beta} \partial \tilde{\Theta}} \right. \\
&\quad \left. + (\Theta - \tilde{\Theta}) \cdot \frac{\partial L_Q}{\partial \tilde{\Theta}^2} \right] \\
&= \frac{1}{N} \sum_{i=1}^N \left[y_i \left(x_i^T \tilde{\beta} + \frac{1}{2} x_i^T \tilde{\Theta} x_i \right) + \ln(1 - \tilde{p}_i) \right] \\
&\quad + (\beta - \tilde{\beta}) \cdot \frac{1}{N} \sum_{i=1}^N x_i (y_i - \tilde{p}_i) \\
&\quad + (\Theta - \tilde{\Theta}) \cdot \frac{1}{2N} \sum_{i=1}^N x_i^T x_i (y_i - \tilde{p}_i) \\
&\quad - \frac{1}{2} (\beta - \tilde{\beta}) \cdot \frac{1}{N} \sum_{i=1}^N (1 - \tilde{p}_i) \cdot \tilde{p}_i \cdot x_i^T \cdot x_i \\
&\quad - 2(\beta - \tilde{\beta})^T (\Theta - \tilde{\Theta}) \\
&\quad \cdot \frac{1}{2N} \sum_{i=1}^N (1 - \tilde{p}_i) \cdot \tilde{p}_i \cdot x_i^T \cdot x_i \cdot x_i^T \\
&\quad - (\Theta - \tilde{\Theta}) \cdot \frac{1}{4N} \sum_{i=1}^N (1 - \tilde{p}_i) \cdot \tilde{p}_i \cdot \|x_i\|^2 \\
&= -\frac{1}{2N} \sum_{i=1}^N (1 - \tilde{p}_i) \\
&\quad \cdot \tilde{p}_i \left\{ [x_i^T (\beta - \tilde{\beta})]^2 \right. \\
&\quad \left. + (\beta - \tilde{\beta})^T (\Theta - \tilde{\Theta}) x_i \right. \\
&\quad \left. \cdot x_i^T x_i + \frac{1}{4} [x_i^T (\Theta - \tilde{\Theta}) x_i]^2 \right\} \\
&\quad + \frac{1}{N} \sum_{i=1}^N \left[y_i \left(x_i^T \tilde{\beta} + \frac{1}{2} x_i^T \tilde{\Theta} x_i \right) + \ln(1 - \tilde{p}_i) \right] \\
&\quad + \frac{1}{N} \sum_{i=1}^N x_i^T (\beta - \tilde{\beta}) (y_i - \tilde{p}_i) \\
&\quad + \frac{1}{2N} \sum_{i=1}^N x_i^T (\Theta - \tilde{\Theta}) x_i (y_i - \tilde{p}_i) \\
&= -\frac{1}{2N} \sum_{i=1}^N (1 - \tilde{p}_i)
\end{aligned}$$

$$\begin{aligned}
& \cdot \tilde{p}_i \left\{ x_i^T (\beta - \tilde{\beta}) \right. \\
& \quad \left. + \frac{1}{2} x_i^T (\Theta - \tilde{\Theta}) x_i \right\}^2 \\
& + \frac{1}{N} \sum_{i=1}^N x_i^T (\beta - \tilde{\beta}) (y_i - \tilde{p}_i) \\
& + \frac{1}{2N} \sum_{i=1}^N x_i^T (\Theta - \tilde{\Theta}) x_i (y_i - \tilde{p}_i) \\
& + \frac{1}{N} \sum_{i=1}^N \left[y_i \left(x_i^T \tilde{\beta} + \frac{1}{2} x_i^T \tilde{\Theta} x_i \right) + \ln(1 - \tilde{p}_i) \right] \\
& = -\frac{1}{2N} \sum_{i=1}^N (1 - \tilde{p}_i) \\
& \quad \cdot \tilde{p}_i \left\{ x_i^T (\beta - \tilde{\beta}) + \frac{1}{2} x_i^T (\Theta - \tilde{\Theta}) x_i \right. \\
& \quad \left. + \frac{y_i - \tilde{p}_i}{\tilde{p}_i [1 - \tilde{p}_i]} \right\}^2 \\
& + \frac{1}{N} \sum_{i=1}^N \left[y_i \left(x_i^T \tilde{\beta} + \frac{1}{2} x_i^T \tilde{\Theta} x_i \right) + \ln(1 - \tilde{p}_i) \right] \\
& - \frac{1}{2N} \sum_{i=1}^N \left[\frac{y_i - \tilde{p}_i}{\tilde{p}_i [1 - \tilde{p}_i]} \right]^2 \\
& = -\frac{1}{2N} \sum_{i=1}^N \omega_i \left(z_i - \sum_j \tilde{\beta}_j x_{ij} - \frac{1}{2} \sum_{j \neq k} \tilde{\Theta}_{jk} x_{ij} x_{ik} \right)^2 \\
& + \frac{1}{N} \sum_{i=1}^N \left[y_i \left(x_i^T \tilde{\beta} + \frac{1}{2} x_i^T \tilde{\Theta} x_i \right) + \ln(1 - \tilde{p}_i) \right] \\
& - \frac{1}{2N} \sum_{i=1}^N \left[\frac{y_i - \tilde{p}_i}{\tilde{p}_i [1 - \tilde{p}_i]} \right]^2,
\end{aligned} \tag{A.3}$$

where

$$\begin{aligned}
\omega_i &= \tilde{p}(x_i) [1 - \tilde{p}(x_i)], \\
z_i &= \sum_j \beta_j x_{ij} + \frac{1}{2} \sum_{j \neq k} \Theta_{jk} x_{ij} x_{ik} \\
&+ \frac{y_i - \tilde{p}(x_i)}{\tilde{p}(x_i) [1 - \tilde{p}(x_i)]}.
\end{aligned} \tag{A.4}$$

B. Proofs from (12) to (13)

First, there are three cases in calculating $\hat{\beta}_j^+ - \hat{\beta}_j^-$:

$$(1) \hat{\beta}_j^+ \geq 0, \hat{\beta}_j^- = 0:$$

$$\hat{\beta}_j^+ - \hat{\beta}_j^- = x_j^T (z_i - A) + \hat{\beta}_j^+ - \hat{\beta}_j^- - \frac{\lambda_1 + \hat{\alpha}_j}{\omega_i} \tag{B.1}$$

$$(2) \hat{\beta}_j^+ \geq 0, \hat{\beta}_j^- = 0:$$

$$\hat{\beta}_j^+ - \hat{\beta}_j^- = x_j^T (z_i - A) + \hat{\beta}_j^+ - \hat{\beta}_j^- + \frac{\lambda_1 + \hat{\alpha}_j}{\omega_i} \tag{B.2}$$

$$(3) \hat{\beta}_j^+ \geq 0, \hat{\beta}_j^- = 0:$$

$$\hat{\beta}_j^+ - \hat{\beta}_j^- = x_j^T (z_i - A) + \hat{\beta}_j^+ - \hat{\beta}_j^-. \tag{B.3}$$

We derive

$$\hat{\beta}_j^+ - \hat{\beta}_j^- = S \left[x_{ij} (z_i - A_i) + \hat{\beta}_j^+ - \hat{\beta}_j^-, \frac{\lambda_1 + \hat{\alpha}_j}{\omega_i} \right]. \tag{B.4}$$

Secondly, we give the following by the stationary condition $\partial L / \partial \hat{\Theta}_{jk} = 0$:

$$\begin{aligned}
& \frac{1}{2} \omega_i \cdot 2 \left(z_i - \sum_j \tilde{\beta}_j x_{ij} - \frac{1}{2} \sum_{j \neq k} \tilde{\Theta}_{jk} x_{ij} x_{ik} \right) \\
& \cdot \left(-\frac{1}{2} x_{ij} x_{ik} \right) + (\lambda_2 + \alpha_j) U_{jk} = 0, \\
& \omega_i \cdot \left(z_i - \sum_j \tilde{\beta}_j x_{ij} - \frac{1}{2} \sum_{j \neq k} \tilde{\Theta}_{jk} x_{ij} x_{ik} \right) \\
& \cdot \left(\frac{1}{2} x_{ij} x_{ik} \right) = (\lambda_2 + \alpha_j) U_{jk}, \\
& \left(z_i - \sum_j \tilde{\beta}_j x_{ij} - \frac{1}{2} \sum_{j \neq k} \tilde{\Theta}_{jk} x_{ij} x_{ik} \right) \\
& \cdot \left(\frac{1}{2} x_{ij} x_{ik} \right) = \frac{(\lambda_2 + \alpha_j) U_{jk}}{\omega_i}.
\end{aligned} \tag{B.5}$$

Supposing

$$\gamma^{(-jk)} = z_i - \sum_j \tilde{\beta}_j x_{ij} - \frac{1}{2} \sum_{j \neq k} \tilde{\Theta}_{jk} x_{ij} x_{ik} + \tilde{\Theta}_{jk} x_{ij} x_{ik}, \tag{B.6}$$

we have

$$\begin{aligned}
& (\gamma^{(-jk)} - \tilde{\Theta}_{jk} x_{ij} x_{ik}) \cdot (x_{ij} x_{ik}) \\
& = \frac{2(\lambda_2 + \alpha_j) U_{jk}}{\omega_i}, \\
& \tilde{\Theta}_{jk} x_{ij} x_{ik} x_{ij} x_{ik} \\
& = -\frac{2(\lambda_2 + \alpha_j) U_{jk}}{\omega_i} + \gamma^{(-jk)} \cdot x_{ij} x_{ik},
\end{aligned}$$

$$\begin{aligned}\widehat{\Theta}_{jk} &= \frac{\gamma^{(-jk)} \cdot x_{ij}x_{ik} - 2(\lambda_2 + \alpha_j)U_{jk}/\omega_i}{(x_{ij}x_{ik})^2}.\end{aligned}\quad (\text{B.7})$$

We discuss three cases about $\widehat{\Theta}_{jk}$ value:

$$(1) \widehat{\Theta}_{jk} > 0, U_{jk} = 1$$

$$\widehat{\Theta}_{jk} = \frac{\gamma^{(-jk)} \cdot x_{ij}x_{ik} - 2(\lambda_2 + \alpha_j)/\omega_i}{(x_{ij}x_{ik})^2}; \quad (\text{B.8})$$

$$(2) \widehat{\Theta}_{jk} < 0, U_{jk} = -1$$

$$\widehat{\Theta}_{jk} = \frac{\gamma^{(-jk)} \cdot x_{ij}x_{ik} + 2(\lambda_2 + \alpha_j)/\omega_i}{(x_{ij}x_{ik})^2} \quad (\text{B.9})$$

$$(3) \widehat{\Theta}_{jk} = 0.$$

We drive

$$\widehat{\Theta}_{jk} = \frac{S[\gamma^{(-jk)} \cdot x_{ij}x_{ik}, 2(\lambda_2 + \alpha_j)/\omega_i]}{(x_{ij}x_{ik})^2}. \quad (\text{B.10})$$

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (nos. 61273019, 61473339), China Postdoctoral Science Foundation (2014M561202), Hebei Postdoctoral Science Foundation Special Fund Project, and Hebei Top Young Talents Support Program.

References

- [1] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society B*, vol. 58, no. 1, pp. 267–288, 1996.
- [2] M. Y. Park and T. Hastie, " L_1 -regularization path algorithm for generalized linear models," *Journal of the Royal Statistical Society B: Statistical Methodology*, vol. 69, no. 4, pp. 659–677, 2007.
- [3] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange, "Genome-wide association analysis by lasso penalized logistic regression," *Bioinformatics*, vol. 25, no. 6, pp. 714–721, 2009.
- [4] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.
- [5] M. Lim and T. Hastie, "Learning interactions via hierarchical group-lasso regularization[EB]," <http://www.stanford.edu/~hastie/Papers/glinetnet.pdf>.
- [6] H. Schwender and K. Ickstadt, "Identification of SNP interactions using logic regression," *Biostatistics*, vol. 9, no. 1, pp. 187–198, 2008.
- [7] S. Noah and R. Tibshirani, "A Permutation Approach to Testing Interactions in Many Dimensions," <http://statweb.stanford.edu/~tibs/research.html>.
- [8] J. Wu, B. Devlin, S. Ringquist, M. Trucco, and K. Roeder, "Screen and clean: a tool for identifying interactions in genome-wide association studies," *Genetic Epidemiology*, vol. 34, no. 3, pp. 275–285, 2010.
- [9] Y. Nardi and A. Rinaldo, "The log-linear group-lasso estimator and its asymptotic properties," *Bernoulli*, vol. 18, no. 3, pp. 945–974, 2012.
- [10] M. Yuan, V. R. Joseph, and Y. Lin, "An efficient variable selection approach for analyzing designed experiments," *Technometrics*, vol. 49, no. 4, pp. 430–439, 2007.
- [11] H. Chipman, "Bayesian variable selection with related predictors," *The Canadian Journal of Statistics*, vol. 24, no. 1, pp. 17–36, 1996.
- [12] N. H. Choi, W. Li, and J. Zhu, "Variable selection with the strong heredity constraint and its oracle property," *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 354–364, 2010.
- [13] J. Bien, J. Taylor, and R. Tibshirani, "A lasso for hierarchical interactions," *The Annals of Statistics*, vol. 41, no. 3, pp. 1111–1141, 2013.
- [14] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, vol. 68, no. 1, pp. 49–67, 2006.
- [15] R. Jenatton, J.-Y. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," *Journal of Machine Learning Research*, vol. 12, no. 10, pp. 2777–2824, 2011.
- [16] P. Radchenko and G. M. James, "Variable selection using adaptive nonlinear interaction structures in high dimensions," *Journal of the American Statistical Association*, vol. 105, no. 492, pp. 1541–1553, 2010.
- [17] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Structured sparsity through convex optimization," *Statistical Science*, vol. 27, no. 4, pp. 450–468, 2012.
- [18] I. Ruczinski, C. Kooperberg, and M. LeBlanc, "Logic regression," *Journal of Computational and Graphical Statistics*, vol. 12, no. 3, pp. 475–511, 2003.
- [19] P. Hall and J.-H. Xue, "On selecting interacting features from high-dimensional data," *Computational Statistics & Data Analysis*, vol. 71, pp. 694–708, 2014.
- [20] J.-J. Wang, J. Li, T. Zhang, and W.-X. Hong, "Distinguishing visual feature extraction method using quadratic map and genetic algorithm," *Journal of System Simulation*, vol. 21, no. 16, pp. 5080–5083, 2009.