Pattern Recognition Letters
journal homepage: www.elsevier.com

High-order Covariate Interacted Lasso for Feature Selection

Zhihong Zhang[a], Yiyang Tian[b], Lu Bai[c,**], Jianbing Xiahou[a], Edwin Hancock[d]

[a]*Software School, Xiamen University, Xiamen, China*
[b]*School of Information Science and Technology, Xiamen University, Xiamen, China*
[c]*School of Information,Central University of Finance and Economics, Beijing, China*
[d]*Department of Computer Science, University of York, York, YO10 5GH, UK*

ABSTRACT

Lasso-type feature selection has been demonstrated to be effective in handling high dimensional data. Most existing Lasso-type models over emphasize the sparsity and overlook the interactions among covariates. Here on the other hand, we devise a new regularization term in the Lasso regression model to impose high order interactions between covariates and responses. Specifically, we first construct a feature hypergraph to model the high-order relations among covariates, in which each node corresponds to a covariate and each hyperedge has a weight corresponding to the interaction information among covariates connected by that hyperedge. For the hyperedge weight, we use multidimensional interaction information (MII) to measure the significance of different covariate combinations with respect to response. Secondly, we use the feature hypergraph as a regularizer on the covariate coefficients which can automatically adjust the relevance measure between a covariate and the response by the interaction weights obtained from hypergraph. Finally, an efficient alternating direction method of multipliers (ADMM) is presented to solve the resulting sparse optimization problem. Extensive experiments on different data sets show that although our proposed model is not a convex problem, it outperforms both its approximately convex counterparts and a number of state-of-the-art feature selection methods.

1. Introduction

Feature interaction presents a challenge to feature selection for classification. In many classification problems, a feature that is completely useless by itself sometimes can provide a significant performance improvement when taken in combination with others. If we only consider relevance and redundancy, but ignore interaction in feature selection, some salient features may be missed (Jakulin and Bratko, 2003). Therefore, identifying discriminative high-order feature interactions is important in machine learning, data mining and data visualization. High order feature interactions often convey essential information about the structures of the problem under consideration and also reveal characteristic features of the datasets under study. For example, genes and proteins seldom perform their functions independently, so many human diseases are often manifested as the dysfunction of some pathways or functional gene modules. As a result, the disrupted patterns due to diseases are often more obvious at a pathway or module level. Identifying these disrupted gene interactions for different diseases such as cancer will help us understand the underlying mechanisms of the diseases and develop effective drugs to cure them.

Recently, linear regression with a sparsity inducing regularizer has been demonstrated to be effective in handling high dimensional data. Sparsity indicates that a regression function can be efficiently represented by a linear combination of active atoms selected from the entire set of variables, and the cardinality of the selected atoms is significantly smaller than the former number of variables. It enables simultaneous parameter estimation and variable selection. For instance, the effects of the explanatory variables $\mathbf{X} = \{x_1, \cdots, x_d\}$ on the response variable Y can be estimated by the corresponding coefficients when fitting the data to the model

$$Y = \beta_1 x_1 + \cdots + \beta_d x_d + \epsilon$$

To improve the prediction accuracy and interpretability of ordinary least squares (OLS), Lasso (Least Absolute Shrink-

**Corresponding author:
e-mail: bailucs@cufe.edu.cn (Lu Bai)

age and Selection Operator) (Tibshirani, 1996) adds an ℓ_1-norm penalty to OLS so as to continuously shrink some coefficients to zero and automatically select a subset of variables. Lasso assumes that the input variables are nearly independent, i.e., they are not highly correlated, while in most real-world data sources, variables are often correlated. Furthermore, in the presence of highly correlated features lasso tends to only select one of these features resulting in suboptimal performance (Zou and Hastie, 2005). For this reason, the Elastic Net (De Mol et al., 2009) uses an additional ℓ_2-regularization term to promote a grouping effect. This method permits groups of correlated features to be selected when the groups are not known in advance. While promising, these methods do not incorporate prior knowledge into the regression/classification process, which is critical in many applications.

Given feature grouping information, the group Lasso (Yuan and Lin, 2006) is a refinement in which variables are organized into groups and each group of variables is penalized based on a combination of the ℓ_1-norm and the ℓ_2-norm. If there is a group of variables in which the pairwise correlations are relatively high, the Lasso tends to select only one variable from the group and is not sensitive to the feature selected. By contrast, the group Lasso considers this group as a whole and determines whether it is important to the problem at hand. If this is the case, each variable in the group is selected, otherwise none are selected. However, the requirement of a nonoverlapping group structure in group Lasso limits its practical applicability. For example, in microarray gene expression data analysis, genes may form overlapping groups since each gene may participate in multiple pathways (Jacob et al., 2009). A further extension of the group Lasso, namely sparse group Lasso, yields sparsity at both the group and individual feature levels. By contrast, it not only determines which groups are selected, but also further selects some of the most important feature variables from each selected group. The coefficients are sparse not only between groups, but also within each group (Zhao et al., 2009).

From the above review of the literature, it is clear that traditional Lasso-type models assume conditional independence among the variables, and their aim is to conduct regression individually for each response vector rather than jointly for all the response vectors. Therefore, they consider data approximation and representation only, without explicitly incorporating correlation information between the response vectors and variables (referred to as relevant information) as well as the variable correlation (referred to as redundant information) in feature selection. Some recent works have been proposed to solve the correlation problem. Chen et al. (2013) proposed an uncorrelated Lasso (unLasso) for variable selection, where variable de-correlation is considered simultaneously with variable selection. Therefore, the selected variables are uncorrelated as much as possible, resulting in little redundancy. Jiang et al. (2014) proposed a covariate-correlated Lasso (ccLasso) that selects the covariates that are correlated more strongly with the response variable. Therefore, the selected covariates are highly relevant to the response, resulting in high relevance.

Although much improvement has been achieve in the works (Chen et al., 2013; Jiang et al., 2014) mentioned above, the selected variables might not be optimal. This is because they only consider relevance and redundancy but ignore high-order variable interactions in the feature selection. As a result, some salient variables may be missing. A variable by itself may have little correlation with the response, but when it is combined with additional variables, it can be strongly influence the response. Unintentional removal of such variables can result in poor classification performance. Therefore, to detect the effects of a variable on the response, it may be necessary to consider it jointly with others.

In order to solve the aforementioned problem with existing Lasso-type variable selection methods in this paper, we propose a high-order covariate interacted Lasso (referred to as interactedLasso). This not only discovers the correlations between the variables and the response, but also discriminates arbitrarily order variable interactions with the response. This distinguishes it from most of the existing feature selection work, which only consider feature relevance and redundancy but ignore high-order feature interactions. Specifically, we first construct a feature hypergraph to model the high-order relations among features, in which each node corresponds to a feature and each hyperedge has a weight corresponding to the interaction information among features connected by that hyperedge. For the hyperedge weight, we use multidimensional interaction information (MII) (Zhang and Hancock, 2012) to measure the significance of different feature combinations with respect to the class. The advantage of MII is that it can go beyond pairwise and consider third or higher order features interaction, which can convey information concerning whether a feature is redundant or interactive. As a result, we can evaluate the significance of candidate features by considering their neighborhood dependency, and thus avoid missing some valuable features arising in individual feature combinations. Secondly, we use the feature hypergraph as a regularizer on the feature coefficients which can automatically adjust the relevance measure between a feature and the class using the interaction weights of the hypergraph. Finally, an efficient alternating augmented Lagrangian method (ADMM) is presented to solve the proposed interactedLasso optimization problem. Promising experimental results show the benefits of the proposed interactedLasso model.

2. Related Work

Feature interaction is an increasingly important research problem. Zhao and Liu (2009) propose to search for interacting features using a consistency criteria to measure feature relevance. Wu et al. (2009) identified discriminative interacting features using regularization techniques. The algorithm heuristically adds some possible high-order interactions into the input feature set in a greedy way based on Lasso penalized logistic regression. Recently, Min et al. (2014) proposed an efficient way to identify combinatorial interactions among interactive genes in complex diseases by using overlapping group lasso and screening.

The work mentioned above has demonstrated the existence and effectiveness of feature interactions, and they are to some extent able to deal with feature interaction. Here these methods

require the order of feature interaction to be specified in advance, and the process of enumerating all possible interaction orders is usually time consuming. In real-world applications, it is hard to estimate the order of feature interactions and different features may have a different optimal interaction order. Enumerating all orders of feature interaction, a large set of feature combinations are generated with redundancy. For example, in (Zhao and Liu, 2009), if we consider i-order feature interactions ($1 \leq i \leq \delta_m$) in a straightforward manner, we have to evaluate C_d^i candidate feature combinations, wherein δ_m is the predefined maximum order of feature interaction. That is to say, to enumerate δ_m relation orders, we have to evaluate $\sum_{i=1}^{\delta_m} C_d^i$ candidate feature combinations, which is computational intractable if δ_m is large. Thus, there is a need to develop new efficient techniques to automatically capture the relevant order of feature interactions in regression models. This problem is the focus of this paper. To do this we make use of a feature hypergraph representation.

Hypergraph representations allow vertices to be multiply connected by hyperedges and can hence capture multiple or higher order relationships between features. Due to their effectiveness in representing multiple relationships, hypergraph based methods have been applied to various practical problems, such as partitioning circuit netlists, clustering (Zhou et al., 2006), clustering categorial data, and image segmentation. For multi-label classification, Sun et al. (2008) construct a hypergraph to exploit the correlation information contained in different labels. In this hypergraph, instances correspond to the vertices and each hyperedge includes all instances annotated with a common label. With this hypergraph representation, the higher-order relations among multiple instances sharing the same label can be explored. Following the theory of spectral graph embedding (Chung, 1997), they transform the data into a lower-dimensional space through a linear transformation, which preserves the instance-label relations captured by the hypergraph. The projection is guided by the label information encoded in the hypergraph and a linear Support Vector Machine (SVM) is used to handle the multi-label classification problem. Huang et al. (2011) used a hypergraph cut algorithm (Zhou et al., 2006) to solve the unsupervised image categorization problem, where a hypergraph is used to represent the complex relationships among unlabeled images based on shape and appearance features. Specifically, they first extract the region of interest (ROI) of each image, and then construct hyperedges among images based on shape and appearance features in their ROIs. Hyperedges are defined as either a) a group formed by each vertex (image) or b) its k-nearest neighbors (based on shape or appearance descriptors). The weight of each hyperedge is computed as the sum of the pairwise affinities within the hyperedge. In this way, the task of image categorization is transferred into a hypergraph partition problem which can be solved using the hypergraph cut algorithm (Wagner and Klimmek, 1996).

One common feature of these existing hypergraph representations is that they exploit domain specific and goal directed representations. Specifically, most of them are confined to uniform hypergraphs where all hyperedges have the same cardinality and do not lend themselves to generalization. The reason for this lies in the difficulty in formulating a nonuniform hypergraph in a mathematically neat way for computation. There has yet to be a widely accepted and consistent way for representing and characterizing nonuniform hypergraphs, and this remains an open problem when exploiting hypergraphs for feature selection.

To address these shortcomings, an effective method for hypergraph construction is needed, such that the ambiguities of relational order can be overcome. Inspired by the recent work (Hu et al., 2008) which utilized the neighborhood dependency to evaluate the significance of a feature, in this paper, we attempt to build a hyperedge connecting a feature and its corresponding neighbors. Instead of generating a hyperedge for each feature, we generate a group of hyperedges by varying the neighborhood size in a specified range. This makes our approach significantly more robust than previous hypergraph methods. Moreover, it can capture the important order of feature interactions, because we do not need to tune the neighborhood size.

Recall that finding high-order feature interactions entails exhaustive search of all feature subsets. In this paper, we attempt to analyze high order feature interaction in the framework of feature hypergraph. Therefore, our search of feature interaction is accelerated since fewer candidate feature combinations are evaluated. Moreover, to judge whether there exists interaction or redundancy between features, we measure the weight of each hyperedge by using multidimensional interaction information (MII) (Zhang and Hancock, 2012). Since redundant features produce negative influence and interaction features produce positive influence according to MII, the hyperedge weight can be used to measure the redundancy and interaction of candidate features. Thus, we can adjust the relevance measure between a feature and the class using its corresponding hyperedge weight.

In summary, our method offers three advantages: (1) We develop an nonuniform hypergraph (i.e. the hyperedge cardinality varies) construction approach by varying the size of correlated features. This makes our approach more robust than that in (Zhang and Hancock, 2012), because we do not need to turn the size of correlated features as a parameter and enumerate all possible orders of feature interaction; (2) For the hyperedge weight, we use multidimensional interaction information (MII) to measure the significance of different feature combinations with respect to the class. The advantage of MII is that it can go beyond pairwise order and capture third or higher order feature interactions, which can reflect whether a feature is redundant or relevant at higher order. As a result, we can evaluate the significance of candidate features by considering their neighborhood dependency, and thus avoid overlooking some valuable features arising in individual feature combinations; (3) We use the feature hypergraph as a regularizer on the feature coefficients which can automatically adjust the relevance measure between a feature and the class using the interaction weights of hypergraph. Therefore, the final selected feature subset is jointly informative with the class.

The remainder of this paper is organized as follows. We briefly review the standard Lasso and Elastic Net in Section 3

and introduce our formulation of high order covariate interacted Lasso in Section 4. Then an effective iterative algorithm is presented to solve the sparse optimization problem in Section 5. Experimental results and performance comparisons with competing methods are presented in Section 6. We conclude this paper by summarizing the proposed method in Section 7.

3. Brief Review of Sparse Learning Based Feature Selection

According to the structure of the norm, sparsity can be obtained from the following two types of regularization terms for feature selection: a) Flat sparsity, where the sparsity is often achieved by the ℓ_1-norm or ℓ_0-norm regularizer to select individual features; b) Structural sparsity, where the $\ell_{2,1}$-norm or $\ell_{2,0}$-norm are imposed to select group features.

Typically we have a set of training data $(x_1, y_1), \ldots, (x_n, y_n)$ from which to estimate the parameters β. Each $x_i = \{f_1^i, f_2^i, \ldots, f_d^i\}^T \in \Re^{d\times 1}$ is a predictive vector of feature measurements for the i-th case. The most popular estimation methods is least squares, in which we select the coefficients $\beta = \{\beta_1, \ldots, \beta_d\}^T$ to minimize the residual sum of squares

$$\min_{\beta} \sum_{i=1}^{n} ||y_i - \sum_{j=1}^{d} \beta_j f_j^i||_2^2 = \min_{\beta} ||\mathbf{y}^T - \beta^T \mathbf{X}||_2^2$$
$$s.t. \quad \sum_{j=1}^{d} ||\beta||_0 = k \tag{1}$$

where $\mathbf{y} \in \Re^{n\times 1}$ is the label vector, $\mathbf{X} \in \Re^{d\times n}$ is the training data, and k is the number of features selected. Solving Eq.1 directly has been proved NP-hard, very difficult even by optimization. In many practical situations it is convenient to allow for a certain degree of error, and we can relax the optimization constraint using the following formulation

$$\min_{\beta} ||\mathbf{y}^T - \beta^T \mathbf{X}||_2^2 + \lambda ||\beta||_0 \tag{2}$$

where $\lambda \geq 0$ is the regularization parameter. Unfortunately Eq.2 is still challenging, and for practical purposes an alternative formulation using ℓ_1-norm regularization instead of ℓ_0-norm has been proposed

$$\min_{\beta} ||\mathbf{y}^T - \beta^T \mathbf{X}||_2^2 + \lambda ||\beta||_1 \tag{3}$$

where $||\beta||_1$ is ℓ_1-norm of vector β (sum of absolute elements), $||\beta||_1 = \sum_{j=1}^{d} |\beta_j|$. The tuning parameter $\lambda \geqslant 0$ controls the amount of regularization applied to the estimate. The larger λ, the larger the number of zeros in β. The nonzero components give the selected variables. After we obtain the optimal value of β, we choose the feature indices corresponding to the top k largest values of the summation of the absolute values along each column. In statistics, Eq.3 is referred to as the regularized counterpart of the Lasso problem (Tibshirani, 1996). This has been widely studied (e.g. (Efron et al., 2004; Osborne et al., 2000a,b)) and proved to have a closed form solution. However, one of the main limitations of ℓ_1-norm feature selection is that it focuses on estimating the response vector for each variable individually without considering relations with the remaining variables. Moreover, the ℓ_1-minimization algorithm is not stable when compared with ℓ_2-minimization (Xu et al., 2012). Therefore, if the goal is to select features across all the classes, some structural sparsity is preferred. In multi-task learning, the $\ell_{2,1}$-norm square regularization term to couple feature selection across tasks. A concrete example is the Elastic Net (Zou and Hastie, 2005).

The Elastic Net (Zou and Hastie, 2005) adds an ℓ_2-minimization term into the Lasso objective function, which can then be formulated as

$$\min_{\beta\in\Re^d} ||\mathbf{y}^T - \beta^T \mathbf{X}||_2^2 + \lambda_1 ||\beta||_1 + \lambda_2 ||\beta||_2^2, \tag{4}$$

where $\lambda_1, \lambda_2 \geqslant 0$ are tuning parameters. Apart from enjoying a similar sparsity of representation to Lasso, the Elastic Net encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together (Zou and Hastie, 2005).

Predictors with high correlation contain similar properties, and contain some overlapped information. In some cases, especially when the number of selected predictors is very limited, more information needs to be contained in the selected predictors. Strongly correlated predictors should not participate in the model together. When strongly correlated predictors are present, then only one is selected. As a result the limited selected predictors will contain more information.

4. Interacted Lasso

In this section, we attempt to analyze feature relevance, redundancy and interaction in the framework of the feature hypergraph. Moreover, we use the feature hypergraph as a regularizer on the feature coefficients which can automatically adjust the relevance measure between a feature and the class through the interaction weights of hypergraph. As a result, the final selected feature subset are jointly informative with the class.

4.1. Hypergraph Construction via Multiple Feature Neighborhoods

For our hypergraph construction, we regard each feature in the data set as a vertex on hypergraph $H = (V, E, \mathbf{W})$, where $V = \{v_1, v_2, \ldots, v_d\}$ is the vertex set, $E = \{e_1, e_2, \ldots, e_m\}$ is set of non-empty subsets of V or hyperedges and $\mathbf{w}(\mathbf{e})$ is a weight function which associates a real value with each hyperedge. Assume there are d-dimensional features in the data set, and thus, the generated hypergraph contains d vertices. In our method, a hyperedge is constructed from a feature and its k nearest neighbors. Instead of generating a hyperedge for each feature, we generate a group of hyperedges by varying the neighborhood size k in a specified range. Specifically, in our experiment, we vary k value from 2 to 7 with an incremental step of 1. This makes our approach much more robust than previous hypergraph methods, because we do not need to tune the neighborhood size.

4.2. Computing Hyperedge Weight by High-order Features Correlation

Given a set of features $f_{i_1}, f_{i_2}, \ldots, f_{i_K}$, the interaction information among them can be measured by joint entropy (MacKay, 2003):

$$H(f_{i_1}, f_{i_2}, \ldots, f_{i_K}) = -\sum_{f_{i_1}, f_{i_2}, \ldots, f_{i_K}} P(f_{i_1}, f_{i_2}, \ldots, f_{i_K}) \log_2 P(f_{i_1}, f_{i_2}, \ldots, f_{i_K}) \quad (5)$$

where $P(f_{i_1}, f_{i_2}, \ldots, f_{i_K})$ is the probability of features $f_{i_1}, f_{i_2}, \ldots, f_{i_K}$ occurring together. According to the above definition, we can see that joint entropy is always positive which measuring the amount of information contained in the correlated features. Based on the joint entropy, a new measure called multidimensional interaction information (MII) (Zhang and Hancock, 2012) is defined to measure the high-order correlation among features, i.e.

$$I(f_{i_1}, f_{i_2}, \ldots, f_{i_K}) = \sum_{k=1}^{K} (-1)^{k-1} \sum_{F \subset f_{i_1}, f_{i_2}, \ldots, f_{i_K}, |F|=k} H(F) . \quad (6)$$

In Equation 6, F is a subset of features $\{f_{i_1}, f_{i_2}, \ldots, f_{i_K}\}$, and $H(F)$ represents the joint entropy of a discrete random variable with possible values and probability mass function. It is clear that the greater the value of $I(f_{i_1}, f_{i_2}, \ldots, f_{i_K})$ is, the more relevant the K features are. On the contrary, if $I(f_{i_1}, f_{i_2}, \ldots, f_{i_K}) = 0$, the features are unrelated.

Therefore, for the constructed hypergraph $G = (V, E, \mathbf{W})$, we determine the weight of each hyperedge using an normalized MII which can measure the relevance degree contained in the features of each hyperedge with respect to class label C:

$$\mathbf{W}(f_{i_1}, f_{i_2}, \ldots, f_{i_K}; C) = K \frac{I(f_{i_1}, f_{i_2}, \ldots, f_{i_K}; C)}{H(f_{i_1}) + H(f_{i_2}) + \cdots + H(f_{i_K})} . \quad (7)$$

Therefore, a large value of $\mathbf{W}(f_{i_1}, f_{i_2}, \ldots, f_{i_K}; C)$ means $\{f_{i_1}, f_{i_2}, \ldots, f_{i_K}\}$ are strongly relevant with respect to the class label C.

4.3. Informative feature Matrix Construction

For the constructed hypergraph, the vertex-edge incident matrix $\mathbf{H} \in \Re^{|V| \times |E|}$ can be defined as:

$$H(v, e) = \begin{cases} 1 & \text{if } v \in e \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

Let $\mathbf{W}$ be the diagonal matrix containing the weight of hyperedges, and the adjacent matrix $\mathbf{S}$ is

$$\mathbf{S} = \mathbf{H}\mathbf{W}\mathbf{H}^{\mathbf{T}} \quad (9)$$

where $\mathbf{H}^{\mathbf{T}}$ is the transpose of $\mathbf{H}$.

Given the hypergtaph adjacency matrix $\mathbf{S}$ and d-dimensional feature indicator vector β with β_i representing the i-th element, we can locate the informative feature subset by finding the solutions of the following maximization problem:

$$\max f(\beta) = \sum_{i=1}^{d} \sum_{j=1}^{d} \beta_i \beta_j s_{i,j} \Rightarrow \max_{\beta \in \Re^d} \beta^T \mathbf{S} \beta \quad (10)$$

subject to $\beta \in \triangle$, where the multidimensional solution vector β fall on the simplex $\triangle = \{\beta \in \Re^d : \beta \geq 0\}$ and $s_{ii} = 0$, i.e., all diagonal entries of $\mathbf{S}$ are set to zero. Our idea is motivated by graph-based clustering which groups the most dominant vertices into a cluster. On the other hand, in our work, the feature subset $\{f_i | 1 \leq i \leq d, \beta_i > 0\}$ is the most coherent subset of the initial feature set, with maximum internal homogeneity of the feature relevance (7). According to the value of β, all features F fall into two disjoint subsets, $A_1(\beta) = \{f_i | \beta_i = 0\}$ and $A_2(\beta) = \{f_i | \beta_i > 0\}$. We refer to the set of nonzero variables $A_2(\beta)$ as the informative feature subset, because the objective function (10) selects RFS by maximizing features' average relevance.

4.4. Interacted Lasso for Feature Selection

Our discriminative feature subset selection is motivated by the desire to encourage the selected features to jointly correlate more with the response while giving less redundancy among them. Therefore, we unify Eq.3 and Eq.10, and propose the so called interactedLasso for representation and variable selection, which is formulated as

$$\min_{\beta \in \Re^d} \frac{1}{2} \|\mathbf{y}^T - \beta^T \mathbf{X}\|_2^2 + \lambda_1 \|\beta\|_1 - \lambda_2 \beta^T \mathbf{S} \beta, \quad (11)$$

where $\lambda_1, \lambda_2 \geq 0$ are tuning parameters. Note that $\beta^T \mathbf{S} \beta$ is a nonconvex contrain.

It is worth noting that unlike the previous Lasso-type feature selection methods using convex optimization methods, which may be suboptimal in terms of the accuracy of feature selection and parameter estimation. Here, the proposed method imposes more strict nonconvex constraints, i.e., 'high order variable response interactions', in finding the optimal regression β. Once the solution β^* of Eq.11 is obtained, we can easily recover the number of the selected features and index of the selected feature: a feature f_i is selected if and only if $\beta_i^* > 0$. Consequently, the number of selected features is determined by the number of positive coordinated of β^*.

5. Optimization Algorithm

We propose to solve the non-convex problem 11 by using the alternating direction method of multipliers (ADMM) (Boyd et al., 2011). The basic idea of the ADMM approach is to decompose a hard problem into a set of simpler ones. ADMM attempts to combine the benefits of augmented Lagrangian methods and the dual decomposition for constrained optimization problem (Boyd et al., 2011). By introducing an auxiliary variable γ into the objective function Eq.11, the problem solved by ADMM takes the following form:

$$\min_{\beta \in \Re^d} \frac{1}{2} \|\mathbf{y}^T - \beta^T \mathbf{X}\|_2^2 - \lambda_2 \beta^T \mathbf{S} \beta + \lambda_1 \|\gamma\|_1, \quad s.t. \quad \beta - \gamma = 0 \quad (12)$$

which is clearly equivalent to the problem in Eq.11. We can regard γ as a proxy for β. The augmented Lagrangian associated

with the constrained problem 12 given by

$$L(\beta,\gamma,z)=\frac{1}{2}\|\mathbf{y}^T-\beta^T\mathbf{X}\|_2^2-\lambda_2\beta^T\mathbf{S}\beta+\lambda_1\|\gamma\|_1 +\langle\beta-\gamma,z\rangle+\frac{\rho}{2}\|\beta-\gamma\|_2^2 \quad (13)$$

Here ρ is a positive penalty parameter (or dual update length) and z is a dual variable (i.e. the Lagrange multiplier) corresponding to the equality constraint $\beta=\gamma$. By introducing an additional variable γ and an additional constraint $\beta-\gamma=0$, we have simplified the problem as 11 by decoupling the objective function into two parts that depend on two different variables.

The alternating direction method of multipliers (ADMM) that solves our original problem in 11 seeks for a saddle point of the augmented Lagrangian by iteratively minimizing $L(\beta,\gamma,z)$ over β,γ, and updating z according to the following update rule:

1) β-minimization: $\beta^{k+1}=\arg\min_{\beta\in\Re^d} L(\beta,\gamma^k,z^k)$

2) γ-minimization: $\gamma^{k+1}=\arg\min_{\gamma\in\Re^d} L(\beta^{k+1},\gamma,z^k)$

3) z-update: $z^{k+1}=z^k+\rho(\beta^{k+1}-\gamma^{k+1})$

All the challenges of the algorithm now reside essentially in the resolution of these problems until some stopping criterion is satisfied. Applying ADMM, we carry out the following steps at each iteration:

Update β: In the (k+1)-th iteration, β^{k+1} is computed by minimizing $L(\beta,\gamma,z)$ with γ^k and z^k fixed. Then we need to solve the following subproblem:

$$\min_{\beta\in\Re^d}\frac{1}{2}\|\mathbf{y}^T-\beta^T\mathbf{X}\|_2^2-\lambda_2\beta^T\mathbf{S}\beta+\langle\beta-\gamma^k,z^k\rangle+\frac{\rho}{2}\|\beta-\gamma^k\|_2^2 \quad (14)$$

Taking derivatives with respect to β and set it to zero, we have

$$\frac{\partial}{\partial\beta}\Big[\frac{1}{2}\|\mathbf{y}^T-\beta^T\mathbf{X}\|_2^2-\lambda_2\beta^T\mathbf{S}\beta+\langle\beta-\gamma^k,z^k\rangle+\frac{\rho}{2}\|\beta-\gamma^k\|_2^2\Big]=0$$

$$\Rightarrow\begin{cases}\frac{\partial}{\partial\beta}\frac{1}{2}\|\mathbf{y}^T-\beta^T\mathbf{X}\|_2^2=-\mathbf{X}\mathbf{y}+\mathbf{X}\mathbf{X}^T\beta,\\ \frac{\partial}{\partial\beta}(-\lambda_2\beta^T\mathbf{S}\beta)=-2\lambda_2\mathbf{S}\beta,\\ \frac{\partial}{\partial\beta}\langle\beta-\gamma^k,z^k\rangle=z^k,\\ \frac{\partial}{\partial\beta}(\frac{\rho}{2}\|\beta-\gamma^k\|_2^2)=\rho(\beta-\gamma^k).\end{cases} \quad (15)$$

$$\Rightarrow\beta^{k+1}=(\rho\mathbf{I}+\mathbf{X}\mathbf{X}^T-2\lambda_2\mathbf{S})^{-1}[\mathbf{X}\mathbf{y}-z^k+\rho\gamma^k]$$

Update γ: Now supposing that β_i^{k+1} and the Lagrangian multipliers z_i^k, $i=1,\cdots,d$ are fixed in the Lagrangian, the optimization problem related to $\gamma_i^{k+1}, i=1,\cdots,d$ boils down to be:

$$\min_{\gamma_i}\lambda_1\sum_{i=1}^d\|\gamma_i\|_1-\sum_{i=1}^d\langle\gamma_i,z_i^k\rangle+\frac{\rho}{2}\sum_{i=1}^d\|\beta_i^{k+1}-\gamma_i\|_2^2 \quad (16)$$

Taking the derivative with respect to γ_i and setting it to zero, we have

$$\frac{\partial}{\partial\gamma_i}\Big[\lambda_1\sum_{i=1}^d\|\gamma_i\|_1-\sum_{i=1}^d\langle\gamma_i,z_i^k\rangle+\frac{\rho}{2}\sum_{i=1}^d\|\beta_i^{k+1}-\gamma_i\|_2^2\Big]=0$$

$$\Rightarrow\frac{\partial(\lambda_1|\gamma_i|)}{\partial\gamma_i}=z_i^k-\rho(\gamma_i-\beta_i^{k+1})$$

$$\Rightarrow\gamma_i^{k+1}=\begin{cases}\frac{1}{\rho}(z_i^k+\rho\beta_i^{k+1}-\lambda_1), & \text{if } z_i^k+\rho\beta_i^{k+1}>\lambda_1\\ \frac{1}{\rho}(z_i^k+\rho\beta_i^{k+1}+\lambda_1), & \text{if } z_i^k+\rho\beta_i^{k+1}<-\lambda_1\\ 0 & \text{if } z_i^k+\rho\beta_i^{k+1}\in[-\lambda_1,\lambda_1].\end{cases} \quad (17)$$

Update z: Update $z_i^{k+1}, i=1,\cdots,d$:

$$z_i^{k+1}=z_i^k+\rho(\beta_i^{k+1}-\gamma_i^{k+1}). \quad (18)$$

A summary of the proposed method is shown in Algorithm 1 below

Algorithm 1: The proposed ADMM algorithm for interactedLasso

Input: X,y,β^0,z^0, λ_1,λ_2 and ρ
Output: β

```
while not converge do
    Update β^{k+1} according to Eq.15;
    Update γ_i^{k+1}, i = 1,···,d according to Eq.17;
    Update z_i^{k+1}, i = 1,···,d according to Eq.18.
end while
```

The algorithm stops when the primal and dual residuals (Boyd et al., 2011) satisfy a certain stopping criterion. The stopping criterion can be specified by two thresholds: absolute tolerance ε_{abs} and relative tolerance ε_{rel} (see Boyd et al. (2011) for more details). The penalty parameter ρ affects the primal and dual residuals, and hence affects the termination of the algorithm. A large ρ tends to produce small primal residuals, but increases the dual residuals (Boyd et al., 2011). A fixed ρ (say 10) is commonly used. However there are some alternative schemes of varying the penalty parameter to achieve better convergence.

6. Convergence and Complexity Analysis

In this section, we will analyze the properties of the InteractedLasso algorithm according to three criteria. We first provide a convergence analysis and then discuss its computational complexity and the parameter determination problems.

6.1. Convergence Proof

On the convergence of Algorithm 1, we have the following result.

Theorem 1. *Let $\{\beta^k\}, \{\gamma^k\}, \{z^k\}$ be the iterative sequences generated by Algorithm 1. Suppose that the sequence $\{z^k\}$ converges to a point, i.e., $\lim_{k\to\infty} z^k = \bar{z}$ for some $\bar{z}$. Then every limit point $(\bar{\beta}, \bar{\gamma})$ of the sequence $\{(\beta^k, \gamma^k)\}$, together with $\bar{z}$, satisfy the necessary first order conditions of the problem 12: 1) Primal feasibility: $\bar{\beta} - \bar{\gamma} = 0$. 2) Dual feasibility: $\nabla f(\bar{\beta}) + \bar{z} = 0$ and $0 \in \partial g(\bar{\gamma}) - \bar{z}$, where ∂ denotes the sub-differential operator (see (Rockafellar, 1970)).*

One can easily prove Theorem 1 by following a proof similar to that of Proposition 3 in (Magnússon et al., 2014). We observe from Theorem 1 that, in general, Algorithm 1 converges to a local solution to problem 12.

The algorithm stops when the primal and dual residuals (Boyd et al., 2011) satisfy a stopping criterion. The stopping criterion can be specified by two thresholds namely a) the absolute tolerance ε_{abs} and b) the relative tolerance ε_{rel} (see Boyd et al. (2011) for more details). The penalty parameter ρ affects the primal and dual residuals, and hence in turn affects the termination of the algorithm. A large ρ tends to produce small primal residuals, but increases the dual residuals (Boyd et al., 2011). A fixed ρ (say 10) is commonly used. But there are some alternative schemes for varying the penalty parameter which achieve better convergence (Yang et al., 2013).

6.2. Complexity Analysis

At each iteration, the time complexity for updating β according to Eq.15 is $O(d^2 * n)$, where d is the dimension of input data and n is the number of data points. The computational costs of updating γ in Eq.17 and z in Eq.18 are $O(d)$. Thus, the overall computational complexity of **Algorithm 1** is $\max\left\{O(k * d^2 * n), O(k * d)\right\}$ where k the required number of iterations to converge.

6.3. Parameter Determination

A parallel issue to optimizing the interactedLasso algorithm is selecting optimal values of the parameters λ_1 and λ_2. The parameter λ_1 is a regularization parameter controlling the sparsity of β, and the parameter λ_2 is used to trade off the importance of data linear regression and high order covariate interactions. In order to assign an appropriate value of λ_2, we employ a cross-validation procedure for λ_2 estimation. In addition, λ_1 is empirically determined by grid search.

7. Experiments and Comparisons

In this section, we discuss the merits and limitations of the proposed feature selection approach. A comprehensive experimental study on 8 data sets is conducted in order to compare our feature selection approach with 6 state-of-the-art methods.

7.1. Experimental Setting

To demonstrate the effectiveness of the proposed approach, we conduct experiments on 8 benchmark data sets, i.e., the USPS handwritten digit data set (Hull, 1994), Isolet speech data set and Pie data set from the UCI Machine Learning Repository (Frank and Asuncion, 2010), YaleB face data set (Georghiades et al., 2001), malignant glioma (GLIOMA) data set (Nutt et al., 2003), ALLAML (Nie et al., 2010), Leukemia and Lymphoma datasets (Vinh et al., 2016). Table. 1 summarizes the extents and properties of the 8 data-sets.

Table 1. Summary of 8 benchmark data sets

Data-set	Sample	Features	Classes
Isolet1	1560	617	26
USPS	9298	256	10
YaleB	2414	1024	38
Pie	11554	1024	68
Leukemia	73	7129	2
Lymphoma	96	4026	9
GLIOMA	50	4434	4
AMLLML	72	7129	2

7.2. Experiment setup

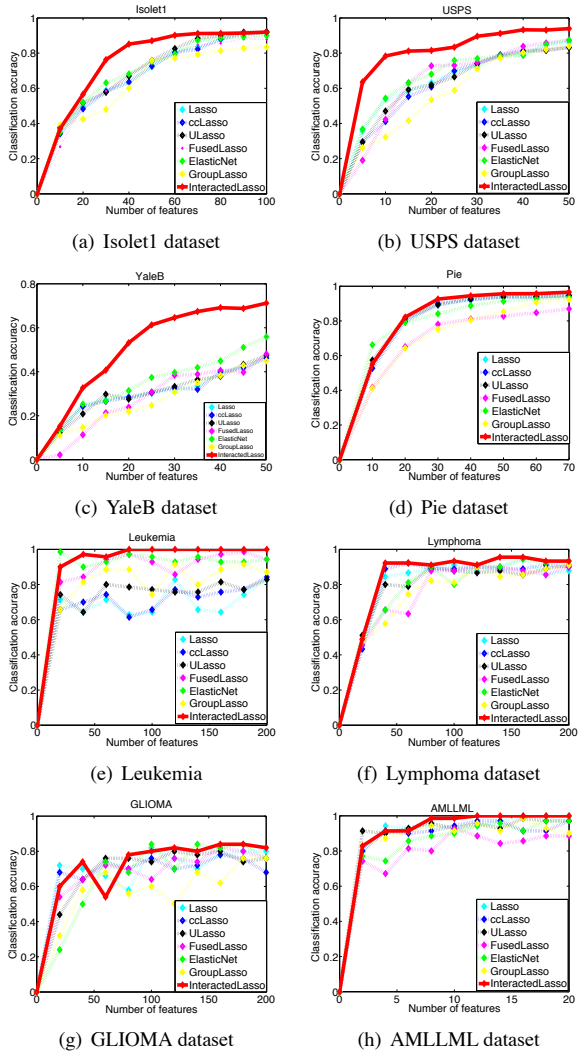
In order to explore the discriminative capabilities of the information captured by our method, we use the selected features for further classification. We compare the classification results from our proposed method (InteractedLasso) with six representative Lasso-type feature selection algorithms. These methods are the Lasso (Tibshirani, 1996), unLasso (Chen et al., 2013), ccLasso (Jiang et al., 2014), Fused Lasso (Tibshirani et al., 2005), Elastic Net (Zou and Hastie, 2005) and group Lasso (Ma et al., 2007). We will briefly introduce these methods one by one.

- Lasso (Tibshirani, 1996): The main task of Lasso is to identify the set of features whose coefficients turn out to be nonzero by ℓ_1 regularizer. However, in the presence of highly correlated features, it tends to arbitrarily select one of them.
- Fused Lasso (Tibshirani et al., 2005): The fused Lasso enforces sparsity in both the coefficients and their successive differences. It is desirable for applications with features ordered in some meaningful way.
- Elastic Net (Zou and Hastie, 2005): The Elastic Net retains the sparse property of Lasso but uses an additional ℓ_2 regularizer to encourage highly correlated features to be jointly selected.
- group Lasso (Ma et al., 2007): The group Lasso is known to enforce the sparsity on variables at an inter-group level, where variables from different groups are competing to survive.
- unLasso (Chen et al., 2013): For unLasso method, variable de-correlation is considered simultaneously with variable selection, so that the selected variables are uncorrelated as much as possible.
- ccLasso (Jiang et al., 2014): The basic idea of ccLasso is to apply prior knowledge of variable-response correlation into Lasso regularized feature selection, thus the final selected variables are strongly correlated with the responses.

A 10-fold cross-validation strategy using the C-Support Vector Machine (C-SVM) (Chang and Lin, 2011) is employed to evaluate the classification performance. Specifically, the entire sample is randomly partitioned into 10 subsets and then we choose one subset for test and use the remaining 9 for training,

and this procedure is repeated 10 times. The final accuracy is computed by averaging of the accuracies from all experiments.

7.3. Classification Comparison



(a) Isolet1 dataset (b) USPS dataset

(c) YaleB dataset (d) Pie dataset

(e) Leukemia (f) Lymphoma dataset

(g) GLIOMA dataset (h) AMLLML dataset

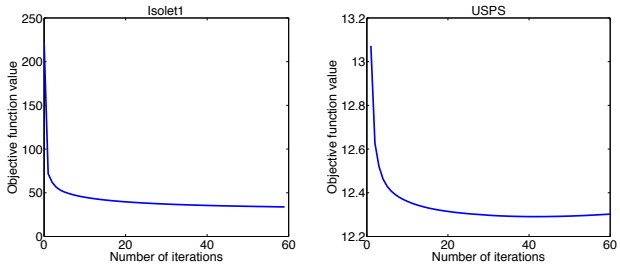
Fig. 1. Accuracy rate vs. the number of selected features on 8 benchmark datasets

The classification accuracies of different algorithms obtained with different feature subsets are shown in Fig.1. From the figure, it is clear that our proposed method dLasso is, by and large, superior to the alternative Lasso-type feature selection methods on all 8 benchmark datasets. As Fig.1 (a) and (b) shows, when the number of selected features is small, the InteractedLasso performs much better than other Lasso-type feature selection methods. The results verify that InteractedLasso can select more discriminative feature subsets than the baselines. However, we observed that the advantage of the proposed algorithm over the other 6 comparative methods tends to diminish as the selected number of features is increased. This is within our expectation, as any feature selection method will work well if we aim to select most of the features.

For clear comparison, we summarize the averaged classification accuracy of different methods when a different number of features is selected. Table. 2 reports the "aggregated " SVM classification accuracy of the different algorithms on each data set. The aggregated SVM classification accuracy is obtained by averaging the averaged accuracy achieved by SVM using the top 10,20,. . .,200 features selected by each algorithm. The boldfaced values are the highest ones. The classification accuracy (MEAN ± STD) is shown first and the number of features selected is reported in brackets. Our method InteractedLasso improved the classification accuracy by 6.3% (Isolet1), 14.44% (USPS), 17.67% (YaleB) , 1.17% (Pie), 3.6% (GLIOMA), 1.72% (ALLAML), 4% (Leukemia) and 3.56% (Lymphoma) respectively, compared to the best performances among the competing methods. In Isolet1 dataset, we can note that for small number of selected features (i.e. 10 features are selected), the highest accuracy achieved by unLasso is 37.82% which is higher than our proposed method InteractedLasso (37.05%). However, if we select more features (i.e. 30 features are selected), InteractedLasso is clearly larger than the alternative Lasso-type feature selection methods. The results verify that having too few features is not necessarily a good feature selection result. Some interactive features may be lost in the process of removing redundancy.

The bottom row of Table. 2 shows the averaged classification accuracy for all the algorithms over the 8 datasets. Our method improved the classification accuracy by 11.63% (Lasso), 10.73% (ccLasso), 9.85% (unLasso), 11.54 % (FusedLasso), 8.1% (Elastic Net), 14.96% (Group Lasso) respectively, compared to the averaged classification accuracy of all competing methods over the 8 datasets. Meanwhile, our method gives a lower standard deviation and hence more stable than the alternatives. Comparatively, group Lasso gives the worst performance. This may be explained by our observation that it is unable to handle feature redundancy and is prone to select redundant features. The reason why the proposed InteractedLasso wins over unLasso (Chen et al., 2013) and ccLasso (Jiang et al., 2014) is that InteractedLasso considers not only the relevance between a single feature and the class, but also the redundancy and interaction with other features which are expressed by the feature hypergraph. Therefore, InteractedLasso performs better when there is feature interaction in the dataset.

7.4. Convergence of InteractedLasso



(a) Isolet1 dataset (b) USPS dataset

Fig. 2. The behavior of proposed objective function value during iterations

Table 2. Aggregated SVM classification accuracy (MEAN ± STD). The last row shows the averaged classification accuracy of all the algorithms over the 8 datasets.

Dataset	Lasso	ccLasso	unLasso	FusedLasso	Elastic Net	Group Lasso	InteractedLasso
Isolet1	71.53% ± 3.70	72.02% ± 3.84	73.49% ± 3.29	70.49% ± 3.19	72.80% ± 3.93	66.97% ± 3.23	**79.79**% ± 2.95
USPS	68.17% ± 2.61	65.64% ± 1.56	66.19% ± 1.61	68.09% ± 1.40	70.47% ± 1.83	60.71% ± 1.28	**84.91**% ± 1.05
YaleB	31.62% ± 2.77	34.64% ± 2.41	34.93% ± 2.74	29.52 % ± 2.52	36.75% ± 3.90	28.44% ± 2.71	**54.42**% ± 2.67
Pie	85.11% ± 0.99	86.31% ± 0.96	86.18% ± 0.97	79.77 % ± 1.05	85.14% ± 0.95	75.67% ± 0.96	**87.48**% ± 0.87
GLIOMA	70.20% ± 2.18	71.6% ± 2.36	72.20% ± 1.61	71.20 % ± 2.17	69.60% ± 2.18	60.60% ± 2.34	**75.8**% ± 1.64
ALLAML	94.57% ±0.78	92.14% ± 0.91	94.43% ± 0.99	83.14 % ± 1.38	89.14% ± 1.15	91.57% ± 1.14	**96.29**% ± 0.88
Leukemia	70.43% ± 1.75	72.43% ± 1.71	76.71% ± 1.63	92% ± 1.26	94.29% ± 1.05	83.71% ± 1.53	**98.29**% ± 0.33
Lymphoma	84.22% ± 1.19	85.11% ± 1.27	82.78% ± 1.30	79.34 % ±1.40	82.67% ± 1.31	78.33% ± 1.49	**88.67**% ± 0.95
AVG	71.98%	72.48%	73.36%	71.67 %	75.11%	68.25%	**83.21**%

Figure 2 shows the variation of proposed objective function across the iterations in Algorithm 1. We can see that Algorithm 1 converges very quickly and the maximum number of interaction is fewer than 30, indicating the efficiency and effectiveness of the proposed InteractedLasso algorithm.

8. Conclusion

The main goal of feature selection is to find a feature subset that is small in size but high in predictive accuracy. Feature interaction exists in many applications. It is a challenging task to find interactive feature. In this paper, we have proposed a novel Interacted Lasso regression model to identify high-order feature interactions. Our major methodological contribution is that by introducing meaningful neighborhood information constraint, we can effectively evaluate whether a feature is redundant or interactive based on a neighborhood dependency measure. We thus avoid missing some valuable features arising in individual feature combinations. Empirical experiments on real datasets show that our model outperforms several well-known techniques such as Lasso, ccLasso and unLasso.

References

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

Chang, C.C., Lin, C.J., 2011. Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2, 1–27.

Chen, S.B., Ding, C., Luo, B., Xie, Y., 2013. Uncorrelated lasso, in: Twenty-Seventh AAAI Conference on Artificial Intelligence.

Chung, F.R., 1997. Spectral graph theory. American Mathematical Soc.

De Mol, C., De Vito, E., Rosasco, L., 2009. Elastic-net regularization in learning theory. Journal of Complexity 25, 201–230.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al., 2004. Least angle regression. The Annals of statistics 32, 407–499.

Frank, A., Asuncion, A., 2010. Uci machine learning repository .

Georghiades, A.S., Belhumeur, P.N., Kriegman, D.J., 2001. From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 643–660.

Hu, Q., Yu, D., Liu, J., Wu, C., 2008. Neighborhood rough set based heterogeneous feature subset selection. Information sciences 178, 3577–3594.

Huang, Y., Liu, Q., Lv, F., Gong, Y., Metaxas, D.N., 2011. Unsupervised image categorization by hypergraph partition. IEEE Transactions on Pattern Analysis and Machine Intelligence 33, 1266–1273.

Hull, J.J., 1994. A database for handwritten text recognition research. IEEE Transactions on Pattern Analysis and Machine Intelligence 16, 550–554.

Jacob, L., Obozinski, G., Vert, J.P., 2009. Group lasso with overlap and graph lasso, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM. pp. 433–440.

Jakulin, A., Bratko, I., 2003. Analyzing attribute dependencies. 7th European Conference on Principles and Practice of Knowledge Discovery in Databases , 229–240.

Jiang, B., Ding, C., Luo, B., 2014. Covariate-correlated lasso for feature selection, in: Machine Learning and Knowledge Discovery in Databases. Springer, pp. 595–606.

Ma, S., Song, X., Huang, J., 2007. Supervised group lasso with applications to microarray data analysis. BMC Bioinformatics 8, 1–17.

MacKay, D.J., 2003. Information theory, inference and learning algorithms. Cambridge university press.

Magnússon, S., Weeraddana, P.C., Rabbat, M.G., Fischione, C., 2014. On the convergence of alternating direction lagrangian methods for nonconvex structured optimization problems. arXiv preprint arXiv:1409.8033 .

Min, M., Chowdhury, S., Qi, Y., Stewart, A., Ostroff, R., 2014. An integrated approach to blood-based cancer diagnosis and biomarker discovery., in: In Proceedings of the Pacific Symposium on Biocomputing, pp. 87–98.

Nie, F., Huang, H., Cai, X., Ding, C.H., 2010. Efficient and robust feature selection via joint 2, 1-norms minimization, in: Advances in Neural Information Processing Systems, pp. 1813–1821.

Nutt, C.L., Mani, D., Betensky, R.A., Tamayo, P., Cairncross, J.G., Ladd, C., Pohl, U., Hartmann, C., McLaughlin, M.E., Batchelor, T.T., et al., 2003. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. Cancer research 63, 1602–1607.

Osborne, M.R., Presnell, B., Turlach, B.A., 2000a. A new approach to variable selection in least squares problems. IMA journal of numerical analysis 20, 389–403.

Osborne, M.R., Presnell, B., Turlach, B.A., 2000b. On the lasso and its dual. Journal of Computational and Graphical statistics 9, 319–337.

Rockafellar, R., 1970. Convex analysis .

Sun, L., Ji, S., Ye, J., 2008. Hypergraph spectral learning for multi-label classification, in: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM. pp. 668–676.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) , 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K., 2005. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 91–108.

Vinh, N.X., Zhou, S., Chan, J., Bailey, J., 2016. Can high-order dependencies improve mutual information based feature selection? Pattern Recognition , 46–58.

Wagner, F., Klimmek, R., 1996. A Simple Hypergraph Min Cut Algorithm. Technical Report. Technical Report, b 96-02, Inst. Of Computer Science, Freie Universität Berlin.

Wu, T.T., Chen, Y.F., Hastie, T., Sobel, E., Lange, K., 2009. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25, 714–721.

Xu, H., Caramanis, C., Mannor, S., 2012. Sparse algorithms are not stable: A no-free-lunch theorem. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34, 187–193.

Yang, S., Wang, J., Fan, W., Zhang, X., Wonka, P., Ye, J., 2013. An efficient admm algorithm for multidimensional anisotropic total variation regularization problems, in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM. pp. 641–649.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68, 49–67.

Zhang, Z., Hancock, E.R., 2012. Hypergraph based information-theoretic feature selection. Pattern Recognition Letters 33, 1991–1999.

Zhao, P., Rocha, G., Yu, B., 2009. The composite absolute penalties family for grouped and hierarchical variable selection. The Annals of Statistics , 3468–3497.

Zhao, Z., Liu, H., 2009. Searching for interacting features in subset selection. Intelligent Data Analysis 13, 207–228.

Zhou, D., Huang, J., Schölkopf, B., 2006. Learning with hypergraphs: Clustering, classification, and embedding, in: Advances in neural information processing systems, pp. 1601–1608.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 301–320.

LaTeX Source Files

Click here to download LaTeX Source Files: prletter-28012014.rar