Direct Sparsity Optimization Based Feature Selection for Multi-Class Classification

Hanyang Peng[1] and Yong Fan[2]

[1]National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, China
[2]Department of Radiology, Perelman School of Medicine,
University of Pennsylvania,

July 13, 2016

Outline

Introduction

Why Feature Selection?

- To remove redundant or noisy features
- To improve the generalized performance
- To reduce the computational burden
- To enhance the interpretability of intrinsic characteristics of data

Introduction

Fundamental Model for Feature Selection

Solving the following l_0-Minimization problem, subject to data fitting constraints, $\boldsymbol{Xw} = \boldsymbol{y}$, and then utilize the non-elements of solution to select useful features, i.e.,

$$\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_0 \ , \ s.t., \boldsymbol{Xw} = \boldsymbol{y} \qquad (1)$$

Basis Pursuit

By satisfying some assumptions(Restricted Isometry Property, RIP), the solution of Problem (1) can be obtained by solving (2), i.e.,

$$\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_1 \ , \ s.t., \ \boldsymbol{Xw} = \boldsymbol{y} \qquad (2)$$

It is not so robust when the data $\boldsymbol{X}$ or class label $\boldsymbol{y}$ is corrupted by noise.

Introduction

Sparse SVM

Many Sparse SVM methods with discriminant margin are proposed to improve the robustness and enhance performance , such as l_1-SVM, i.e.,

$$\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_1 \ , \ s.t., \boldsymbol{y} \odot \boldsymbol{X}\boldsymbol{w} \succcurlyeq \mathbf{1} \tag{3}$$

The optimization algorithm is special design for binary–class problem, hence the multi-class problem do not have compact form.

Sparsity Regularization Based Methods

Many Sparsity Regularization Based Methods have been proposed with different sparsity regularization terms, such as Lasso

$$\min_{\boldsymbol{W}} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{w}\|_1 \tag{4}$$

A trade-off between a data-fitting loss function term and a sparsity term should be took, and it is sensitive to the parameter λ

Introduction

Sparsity Regularization Based Methods

In recent years, To learn sparse representations shared across multiple tasks or multiple classes, $l_{2,1}$ −norm based regularized method are proposed, and the class label is rearranged as one-versus-rest model, where $\boldsymbol{Y} = \{\boldsymbol{f}^i\}_{i=1}$ and $\boldsymbol{f}^i = [-1, \dots, 1, \dots, -1]$, such as Robust Feature selection (RFS),

$$\min_{\boldsymbol{W}} \|\boldsymbol{XW} - \boldsymbol{Y}\|_{2,1} + \lambda \|\boldsymbol{w}\|_{2,1} \tag{5}$$

Direct $L_{2,p}$-Minimization for feature selection

The Proposed Original Model

$$\min_{\boldsymbol{W}} \|\boldsymbol{W}\|_{2,p}, \quad s.t., \boldsymbol{Y} \odot \boldsymbol{X}\boldsymbol{W} \succcurlyeq \boldsymbol{1} \tag{6}$$

Advantages

- $\ell_{2,p}$-norm $(0 < p < 1)$ can give rise to more sparse solutions
- No regularization term , do not need to make a compromise between residual of data-fitting and sparsity
- Enlarging discriminant margin between classes can boost generalization performance

It is difficult to solve directly.

Direct $L_{2,p}$-Minimization for feature selection

Equivalent Model

The optimization problem of (6) can be reformulated by introducing a slack variable $\boldsymbol{E}$ whose elements have the same sign as the corresponding elements of $\boldsymbol{Y}$, *i.e.*,

$$\min_{\boldsymbol{W},\boldsymbol{E}} \|\boldsymbol{W}\|_{2,p}, \quad s.t., \boldsymbol{XW} = \boldsymbol{Y} + \boldsymbol{E}, \boldsymbol{Y} \odot \boldsymbol{E} \succcurlyeq \mathbf{0} \qquad (7)$$

Direct Optimization

- Step 1: solve the linear equation $\boldsymbol{XW} = \boldsymbol{Y} + \boldsymbol{E}$ to obtain the solution space of $\boldsymbol{W}$ with variable $\boldsymbol{E}$
- Step 2: directly search the solution space to find a solution to minimize $\|\boldsymbol{W}\|_{2,p}$

Solution Space of W

Gaussian Elimination

$$[\boldsymbol{X} \vdots (\boldsymbol{Y}+\boldsymbol{E})\,] = [\boldsymbol{X}_1\, \boldsymbol{X}_2 \vdots (\boldsymbol{Y}+\boldsymbol{E})\,] \xrightarrow{\text{left-multiply } \boldsymbol{L}} \begin{bmatrix} \boldsymbol{I} & \boldsymbol{M} & \vdots & \boldsymbol{N}+\boldsymbol{LE} \\ \boldsymbol{0} & \boldsymbol{0} & \vdots & \boldsymbol{0} \end{bmatrix} \quad (8)$$

The solution space of $\boldsymbol{W}$ is

$$\boldsymbol{W} = \mathbf{P}\boldsymbol{U} + \boldsymbol{Q} + \boldsymbol{F} = \begin{bmatrix} \boldsymbol{M} \\ \boldsymbol{I} \end{bmatrix} \boldsymbol{U} + \begin{bmatrix} \boldsymbol{N} \\ \boldsymbol{0} \end{bmatrix} + \begin{bmatrix} \boldsymbol{LE} \\ \boldsymbol{0} \end{bmatrix} \quad (9)$$

The problem (10) can be reformulated as

$$\min_{\boldsymbol{U},\boldsymbol{E}} \left\| \boldsymbol{PU} + \boldsymbol{Q} + \begin{bmatrix} \boldsymbol{LE} \\ \boldsymbol{0} \end{bmatrix} \right\|_{2,p}, s.t., \boldsymbol{Y} \odot \boldsymbol{E} \geqslant \boldsymbol{0}, \quad (10)$$

$\ell_{2,\mathrm{p}}$-norm $(0 < p \leq 1)$ is non-smooth and non-convex when $0 < p < 1$

An Optimization Algorithm for the Model

Iterative Optimization Algorithm

- we alternately optimize variables $\boldsymbol{U}$ and $\boldsymbol{E}$ for optimization problem (14).

- Adopting Iteratively Reweighted Least Square (IRLS) straregy, $\ell_{2,\mathrm{p}}$-minimization problem can be reformulated as a least square minimization problem.

At each iterative step, the objective function of the sub-problem can become convex and smooth.

An Optimization Algorithm for the Model

Optimizing Variable U

At k-th step

$$\boldsymbol{W}^k = \mathbf{P}\boldsymbol{U}^k + \boldsymbol{Q} + \begin{bmatrix} \boldsymbol{LE}^k \\ \mathbf{0} \end{bmatrix}$$

$$\boldsymbol{G}^k = \boldsymbol{Q} + \begin{bmatrix} \boldsymbol{LE}^k \\ \mathbf{0} \end{bmatrix}$$

$\boldsymbol{U}^{k+1} = \underset{\boldsymbol{U}}{\operatorname{argmin}} \left\| \boldsymbol{\Sigma}^k (\boldsymbol{PU} + \boldsymbol{G}^k) \right\|_F^2$, where the i-th diagonal element of $\boldsymbol{\Sigma}^k$ is $1/\|\boldsymbol{w}_i^k\|_2^{1-p/2}$

Optimizing Variable E

At k-th step

$$\boldsymbol{V}^k = -\boldsymbol{M}\boldsymbol{U}^{k+1} + \boldsymbol{N} + \boldsymbol{LE}^k$$

$$\boldsymbol{H} = -\boldsymbol{M}\boldsymbol{U}^{k+1} + \boldsymbol{N}$$

$\boldsymbol{E}^{k+1} = \underset{\boldsymbol{E}}{\operatorname{argmin}} \left\| \boldsymbol{\Lambda}^k (\boldsymbol{LE} + \boldsymbol{H}) \right\|_F^2$, s.t. $\boldsymbol{Y} \odot \boldsymbol{E} \succcurlyeq \mathbf{0}$, where the i-th diagonal element of $\boldsymbol{\Lambda}^k$ is $1/\|\boldsymbol{v}_i^k\|_2^{1-p/2}$

Proof of Convergence

Lemma 1.

Given any two vectors $\boldsymbol{a}$ *and* $\boldsymbol{b}$*, we have*

$$(1-\theta)\|\boldsymbol{a}\|_2^2 + \theta\,\|\boldsymbol{b}\|_2^2 \geq \|\boldsymbol{a}\|_2^{2-2\theta}\|\boldsymbol{b}\|_2^{2\theta}$$

where $0 < \theta < 1$ *and the equality holds if and only if* $a = b$.

Lemma 2.

Given an optimization problem:

$$\min_{\boldsymbol{Z}} \|\, \boldsymbol{S}\,\boldsymbol{\Phi}(\boldsymbol{Z})\,\|_{\boldsymbol{F}}^2\,, s.t.\ \boldsymbol{Z} \in \boldsymbol{\mathcal{F}}$$

where $\boldsymbol{\Phi}(\boldsymbol{Z})$ *is a function of* $\boldsymbol{Z}$*,* $\boldsymbol{\mathcal{F}}$ *is the feasible region, and* $\boldsymbol{S}$ *is a diagonal matrix whose* i*-th diagonal element is* $1/\|\boldsymbol{\Phi}(\boldsymbol{Z}_0)_i\|_2^{1-p/2}$ *(*$\boldsymbol{Z}_0$ *could be any object in* $\boldsymbol{\mathcal{F}}$*,* $\boldsymbol{\Phi}(\boldsymbol{Z}_0)_i$ *is the* i*-th row vector of* $\boldsymbol{\Phi}(\boldsymbol{Z}_0)$ *and* $0 < p \leq 2$ *), we have*

$$\|\boldsymbol{\Phi}(\boldsymbol{Z}^*)\|_{2,p} \leq \|\boldsymbol{\Phi}(\boldsymbol{Z}_{\boldsymbol{0}})\|_{2,p}$$

where $\boldsymbol{Z}^*$ *is the optimal solution of Eqn. (19) and the equality holds if and only if* $\boldsymbol{\Phi}(\boldsymbol{Z}^*) = \boldsymbol{\Phi}(\boldsymbol{Z}_0)$

Proof of Convergence

Theorem 1.

The sequence $\{\boldsymbol{W}^k\}$ *produced via the Algorithm has the following properties:* $\|\boldsymbol{W}^k\|_{2,p}$ *is non-increasing at* successive *iteration steps and* $\left\{\|\boldsymbol{W}^k\|_{2,p}\right\}$ *converges to a limited value.*

Theorem 2.

If sequences $\{\boldsymbol{W}^k\}$ *and* $\{\boldsymbol{E}^k\}$ *produced in The Algorithm have limit points, the limit points satisfy the Karush–Kuhn–Tucker (KKT) conditions of Eqn. (6). When* $p \geq 1$, *the limited points are globally optimal.*

Experiments

Effect of parameter p

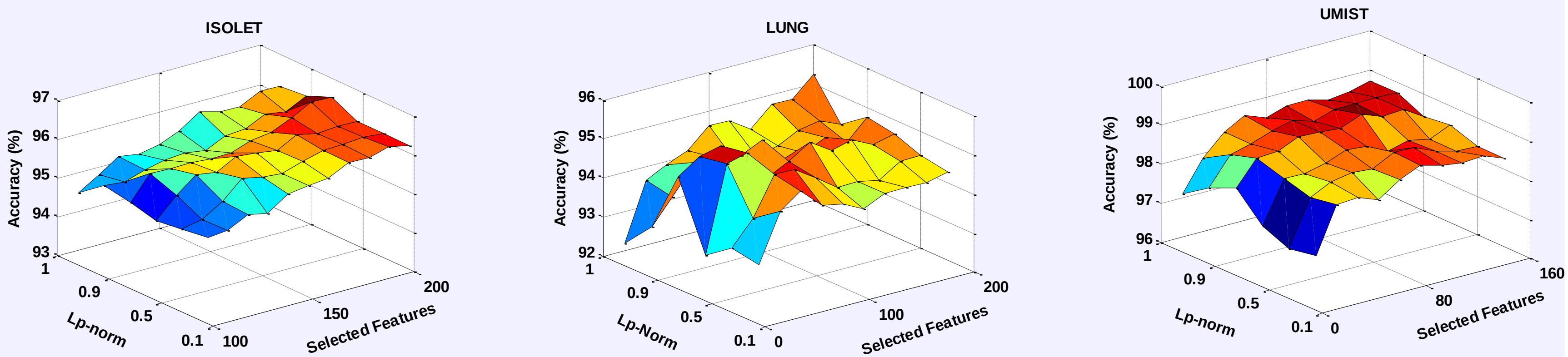

Figure 1: Classification accuracy with different numbers of features selected with different values of p. The results shown were obtained based on datasets: (a) ISOLET, (b) LUNG, and (c) UMIST

Experiments

Effect of parameter p

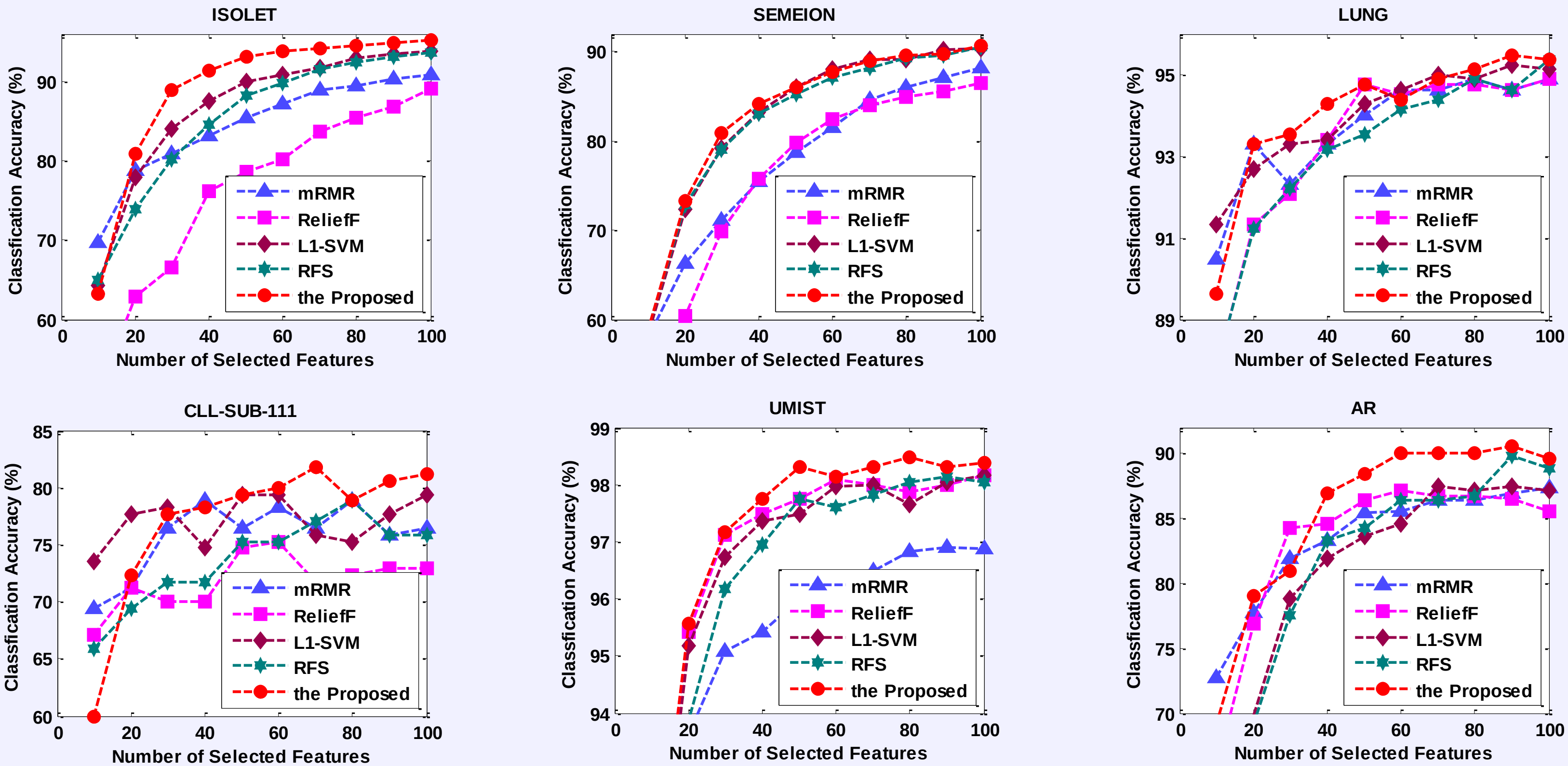

Figure 2: Average classification accuracy of 10 trials for linear—SVM built on the selected top 100 features by different algorithms. The results shown were obtained based on datasets: (a) ISOLET, (b) SEMEION, (c) LUNG , (d) CLL-SUB-111, (e) UMIST, and (f) AR

Conclusions and Discussions

Summary

- Proposed Model: $L_{2,p}$-Minimization subject to data-fitting inequality constraints
- Outstanding Features
  - $L_{2,p}$-norm boosts more sparsity
  - No regularization term free tuning the parameter
  - Enlarging margin between classes improve the robustness to noise and generalization performance
- Optimization Approach
  - Adopting Gaussian Elimination, obtaining the solution space of $\boldsymbol{W}$ with variable $\boldsymbol{E}$
  - Utilizing IRLS strategy, at each iterative step, reformulating the non-convex and non-smooth problem to a least square minimization problem

Questions

Thanks for your attention