Direct Sparsity Optimization Based Feature Selection for Multi-Class Classification

Hanyang Peng\(^1\), Yong Fan\(^2\)
\(^1\)National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, 100190, Beijing, P.R. China
\(^2\)Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA
hanyang.peng@nlpr.ia.ac.cn, yong.fan@ieee.org

Abstract
A novel sparsity optimization method is proposed to select features for multi-class classification problems by directly optimizing a \(\ell_{2,p}\)-norm (\(0 < p \leq 1\)) based sparsity function subject to data-fitting inequality constraints to obtain large between-class margins. The direct sparse optimization method circumvents the empirical tuning of regularization parameters in existing feature selection methods that adopt the sparsity model as a regularization term. To solve the direct sparsity optimization problem that is non-smooth and non-convex when \(0 < p < 1\), we propose an efficient iterative algorithm with proved convergence by converting it to a convex and smooth optimization problem at every iteration step. The proposed algorithm has been evaluated based on publicly available datasets. The experiments have demonstrated that our algorithm could achieve feature selection performance competitive to state-of-the-art algorithms.

1 Introduction
Feature selection has been an important component in machine learning models (Guyon & Elisseeff, 2003). Feature selection approaches in general can be divided into three groups: filter methods (Kira & Rendell, 1992; Lewis, 1992; Peng et al., 2005), wrapper methods (Guyon et al., 2002), and embedded methods (Cawley et al., 2006; Wang et al., 2008; Xiang et al., 2012). The filter methods use proxy measures that are independent on the machine learning models to rank features according to their relevance to the learning problem. The wrapper methods search subsets of features to optimize a given learning model’s performance, and typically have higher computational cost than the filter methods. The embedded methods integrate the feature selection task into the model learning, and are able to achieve good performance with a moderate computational cost. Particularly, sparse linear model based feature selection methods have achieved promising performance. The sparse linear model based methods typically adopt \(\ell_1\)-norm regularization, and many variants have been proposed with different sparsity regularization terms, such as Lasso (Tibshirani, 1996) and sparse support vector machine (SVM) (Bradley & Mangasarian, 1998; Mangasarian, 2006).

In multi-task learning, various \(\ell_{2,1}\)-norm (Liu et al., 2009; Nie et al., 2010; Obozinski et al., 2006) and \(\ell_{\infty,1}\)-norm (Liu et al., 2009) based regularization models have been investigated for selecting features with joint sparsity across different tasks. Moreover, group Lasso based methods (Kong & Ding, 2013; Kong et al., 2014) have also been proposed in recent years. In fact, the multi-task feature selection algorithms share similarities with group lasso based methods. Since non-convex \(\ell_p\)-norm or \(\ell_{r,p}\)-norm (\(0 < p < 1\)) based regularization models can yield sparser solutions than \(\ell_1\)-norm or \(\ell_{r,1}\)-norm based models, they have gained increasing attention in recent studies (Chartrand & Staneva, 2008; Liu et al., 2007; Zhang et al., 2014).

The sparse linear model based feature selection algorithms typically take a trade-off between a data-fitting loss function term and a sparsity term, therefore there inevitably exists residual in the loss function. However, little is known about such a residual’s impact on the feature selection. Most of these algorithms are developed for classification problems. However, their data-fitting loss functions adopted are typically based on the regression of class labels, rather than surrogates of 0-1 loss that are more appropriate for classification problems (Bartlett et al., 2004). To overcome the limitations of existing sparse model based feature selection methods, we present a novel feature selection method via directly optimizing a linear model’s sparsity with \(\ell_{2,p}\)-norm \((0 < p \leq 1)\), subject to data-fitting inequality constraints, instead of adopting the sparsity as a regularization term. Our constrained optimization formulation circumvents the difficulty of regularization parameter setting, and the inequality constraint based data-fitting loss function enables large between-class margins, similar to SVMs, but capable of handling multi-class problems. We propose an efficient algorithm to solve the optimization problem associated with the direct sparsity optimization, which is non-convex and non-smooth.
when $0 < p < 1$, by transforming it to a Frobenius-norm induced problem at each iteration step, which has been proved to converge to Karush–Kuhn–Tucker (KKT) points.

The proposed algorithm has been evaluated based on 6 publicly available datasets, and extensive comparison experiments have demonstrated that our algorithm could achieve feature selection performance competitive to state-of-the-art algorithms, including Relieff (Kira & Rendell, 1992), Min-Redundancy Max-Relevance (mMRMR) (Peng, et al., 2005), $\ell_1$-norm based $\ell_1$-SVM (Bradley & Mangasarian, 1998; Mangasarian, 2006), and $\ell_{2,1}$-norm based Robust Feature Selection (RFS) (Nie, et al., 2010; Xiang, et al., 2012).

2 Direct Sparsity Optimization Based Feature Selection (DSO-FS)

Throughout this paper, matrices are written in bold uppercase, vectors are written in bold lowercase, and all the scalars are denoted by normal letters. $I$ denotes an identity matrix and $1$ denotes a vector or matrix with all the elements equal to 1. Given a matrix $A \in \mathbb{R}^{m \times n}$, the $\ell_{r,p}$-norm($r > 0, p > 0$) of $A$ is defined as:

$$
\|A\|_{r,p} = \left( \sum_{i=1}^{m} \left( \sum_{j=1}^{n} |A_{i,j}|^r \right)^{\frac{1}{r}} \right)^{\frac{1}{p}}, (1)
$$

where $\|A_{i,j}\|_r$ denotes $\ell_r$-norm of the $i$-th row vector of $A$.

Given $m$ training samples $\{x^i, y^i\}_{i=1}^m$ where $x^i \in \mathbb{R}^n$ is a data point and $y^i$ is its associated class label in $c (c \geq 2)$ classes, the multiclass classification problem can be modeled as a linear learning problem. For simplicity, the bias of the standard linear regression is absorbed into $W$ as an additional dimension with all elements equal to 1.

$$
XW = Y, \quad (2)
$$

where $X = [x^1; x^2; \ldots; x^m; 1]$, $W \in \mathbb{R}^{(n+1) \times c}$ is the weight matrix to be learned, and $Y = [f^1; f^2; \ldots; f^m] \in \mathbb{R}^{m \times c}$ is a class label matrix with labels rearranged using a one-versus-rest model, i.e., $f^i = [-1, \ldots, 1, \ldots, -1] \in \mathbb{R}^c$ (the $j$-th element is 1 and others).

For selecting a subset of features, $W$ should have sparse columns and share joint sparsity along its row direction since each row of $W$ corresponds to the same feature. Therefore, we model the feature selection problem as

$$
\min_W \|W\|_{2,0}, \quad s.t., XW = Y, \quad (3)
$$

where $\|W\|_{2,0}$ is the number of non-zero rows in $W$, of which not all the elements are zero.

It is NP-hard to solve the optimization problem of Eqn. (3). Therefore $\ell_{2,p}$-norm ($0 < p \leq 1$) can be adopted instead, resulting in a relaxed sparsity optimization problem:

$$
\min_W \|W\|_{2,p}, \quad s.t., XW = Y, \quad (4)
$$

where $0 < p \leq 1$.

Many studies (Liu, et al., 2009; Nie, et al., 2010; Obozinski, et al., 2006) assumed that the $\ell_{2,1}$-norm based problems is equivalent to $\ell_{2,0}$-norm based problems under certain conditions (Candes & Tao, 2005). In practice, $\ell_{2,p}$-norm ($0 < p < 1$) can lead to sparser solutions in most cases although it is non-convex (Chartrand & Staneva, 2008; Liu, et al., 2007; Zhang, et al., 2014).

It is desired that the classification model’s margins between classes are as large as possible for obtaining improved generalization performance. Accordingly, the equality constraint in Eqn. (4) is relaxed to be inequality constraints, i.e.,

$$
\min_{W,E} \|W\|_{2,p}, \quad s.t., Y \odot XW \geq 1, \quad (5)
$$

where $1 \in \mathbb{R}^{m \times c}$, $\odot$ is a Hadamard product operator for element-wise multiplication between two matrices of the same dimensions, and $\geq$ denotes that the elements of the left matrix are greater than or equal to their corresponding ones of the right matrix.

The optimization problem of Eqn. (5) can be reformulated by introducing a slack variable $E \in \mathbb{R}^{m \times c}$ whose elements have the same positive or negative sign as the corresponding elements of $Y$, as

$$
\min_{W,E} \|W\|_{2,p}, \quad s.t., XW = Y + E, \quad Y \odot E \geq 0 \quad (6)
$$

where $E \in \mathbb{R}^{m \times c}$.

To solve the optimization problem of Eqn. (6), we first solve the linear equation $XW = Y + E$ to obtain the solution space of $W$, and then directly search the solution space to find a solution with the minimum of $\|W\|_{2,p}$. Note that if $XW = Y + E$ is inconsistent, especially when the number of data samples $m$ is greater than the number of features $n$, least-square solution space of the equation can be used as a substitute. Actually, we only need to solve $XW = XX^+ (Y + E)$, where $X^+$ is pseudo-inverse. This equation is compatible when $X$ is row full rank, since $XX^+ - I$ on this occasion. Gaussian Elimination is a simple and efficient way to obtain the solution space of $W$. Without loss the generality, we assume that the first $m_0$ column vectors of $X$ are linearly independent, i.e.,

$$
\left[ X : XX^+ (Y + E) \right] \rightarrow \left[ X, X_2 : XX^+ (Y + E) \right] \left[ I \ M : \ N + LE \right]
$$

where the rank of $X$ is $m_0$, $X_1 \in \mathbb{R}^{m \times m_0}$, $X_2 \in \mathbb{R}^{m \times m_0}$, $I \in \mathbb{R}^{m_0 \times m_0}$, $M \in \mathbb{R}^{m_0 \times N}$, $N \in \mathbb{R}^{N \times m_0}$, and $D \in \mathbb{R}^{m_0 \times m_0}$ is product matrix of a series of elemental matrices, $L = DXX^+ \in \mathbb{R}^{m_0 \times m_0}$, and $n_0 = n + 1 - m_0$. Thus, the solution space of $W$ is
\[ W = PU + Q + F = \begin{bmatrix} M \\ I \end{bmatrix} U + \begin{bmatrix} N \\ 0 \end{bmatrix} + \begin{bmatrix} LE \\ 0 \end{bmatrix} \] (8)

where \( P \in \mathbb{R}^{n \times m_o}, Q \in \mathbb{R}^{n \times c}, F \in \mathbb{R}^{n \times c}, I \in \mathbb{R}^{m_o \times n_o}, \) and \( U \in \mathbb{R}^{n \times c} \) is an arbitrary matrix.

Finally, the optimization problem of Eqn. (5) can be reformulated as

\[
\min_{U,E} \left\| PU + Q + \begin{bmatrix} LE \\ 0 \end{bmatrix} \right\|_{2,p}, \quad s.t., \ Y \odot E \geq 0
\] (9)

where \( 0 < p \leq 1 \).

3 An Iterative Algorithm for DSO-FS

We propose an iterative algorithm to solve the optimization problem of Eqn. (9) due to that no analytical solution is available. At each iteration step, we alternately optimize variables \( U \) and \( E \). An objective function with \( \ell_{2,p} \)-norm (0 < \( p \leq 1 \)) is non-smooth and non-convex when \( 0 < p < 1 \). To efficiently solve this problem, \( \ell_{2,p} \)-norm is reformulated from Frobenius-norm (F-norm) that is smooth and convex, as

\[
\|A\|_{2,p}^p = \|\Sigma A\|_2^p, \quad \text{(10)}
\]

where \( \Sigma \) is a diagonal matrix with the \( i \)-th diagonal element \( \Sigma_i = 1/\|a_i\|_2^{1-p/2} \), and \( \|a_i\|_2 \) is defined in Eqn. (1). Note that \( Z^* = \arg\min_z \|A(Z)\|_{2,p}^p = \arg\min_z \|A(Z)\|_{2,p}^2, \) where \( A(Z) \) is a function of \( Z \). The similar strategies has also been adopted in FOCal Underdetermined System Solver (FOCUSS) (Cotter et al., 2005; Gorodnitsky & Rao, 1997), and various optimization techniques for \( \ell_{2,1} \)-norm regularization based feature selection algorithms (Hou et al., 2011; Krishnapuram et al., 2005; Merchant et al., 2012; Nie et al., 2010; Xiang et al., 2012; Yi Yang et al., 2011) can be considered as its special case when \( p = 1 \). Our solution is summarized in Algorithm 1.

Algorithm 1. Feature Selection via Direct Sparsity Optimization (DSO-FS)

\[ \begin{align*}
\text{Input:} & \quad \text{data points } \{X_i\}_{i=1}^m (X_i \in \mathbb{R}^n) \text{ and their corresponding label } \{y_i\}_{i=1}^m; \text{ norm power } p; \text{ number of features } d \text{ to be selected.} \\
& \quad \text{Construct } X \text{ and } Y \text{ according to Eqn. (2), and } L, M, N, P \text{ and } Q \text{ according to Eqns. (7) and Eqn. (8).} \\
& \quad \text{Set } k = 0 \text{ and initialize } E^0 \text{ and } \Sigma^0 \in \mathbb{R}^{n \times n} \text{ with a zero matrix and an identity matrix respectively.} \\
& \text{repeat} \\
& \quad G = Q + \begin{bmatrix} LE^k \\ 0 \end{bmatrix} \\
& \quad U^{k+1} = \arg\min_U \left\| E^k (PU + G) \right\|_F^2 \\
& \quad V^k = -MU^{k+1} + N + LE^k. \\
& \text{Update diagonal matrix } A^k, \text{ where the } i\text{-th diagonal element is } \\
& \quad \frac{1}{\|w_i\|_2^{p/2}}. \\
\end{align*} \]

Note that at each iteration step, for solving \( U^{k+1} = \arg\min_U \|E^k (PU + G)\|_2^2 \), an analytical solution is available, i.e.,

\[
U^{k+1} = -(P^tSP)^{-1}P^tSG. \quad \text{(11)}
\]

According to Eqn. (8), Eqn. (11) can be reformulated as

\[
U^{k+1} = (M^tS_1M + S_2)^{-1}M^tS_1K, \quad \text{(12)}
\]

where \( S_1 \in \mathbb{R}^{m_0 \times n_0} \) is a diagonal matrix, of which the diagonal elements are the first \( m_0 \) elements of \( S \), \( S_2 \in \mathbb{R}^{n_0 \times m_0} \) is a diagonal matrix, of which the diagonal elements are the last \( n_0 \) elements of \( S \), and \( K = N + LE^k(\in \mathbb{R}^{m \times c}) \).

If \( m_0 < n_0 \), the optimization problem’s computational cost can be further reduced by a simple matrix operation

\[
U^{k+1} = T( MT + I )^{-1} K, \quad \text{(13)}
\]

where \( T = S_2^{-1}M^tS_1 \) and \( I \in \mathbb{R}^{m_0 \times m_0} \).

For feature selection problems with the number of samples less than the number of features, i.e., \( m_0 \leq m < n_0 \), updating Eqn. (13) has reduced computational cost than Eqn. (11). It is worth noting that it is not necessary to directly calculate \( C = (MT + I)^{-1} K \), which is computationally more expensive than solving the following linear equation

\[
(MT + I)C = K. \quad \text{(14)}
\]

Although no analytical solution is available for \( E^{k+1} = \arg\min_E \|A^k(LE + H)\|_F^2 \) (s.t. \( Y \odot E \geq 0 \)), it is a smooth and convex optimization problem. Therefore, it can be efficiently solved by existing tools, such as CVX (CVX Research, 2011).

4 Convergence proof

Algorithm 1 makes \( \|W\|_p \) to monotonically decrease at every iteration step and finally converges.

Lemma 1. Given any two vectors \( a \) and \( b \), we have

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

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

Proof. Since \( \ln(x^2) \) is concave, we have
\[
\ln((1 - \theta)x_1^2 + \theta x_2^2) \geq (1 - \theta) \ln(x_1^2) + \theta \ln(x_2^2), \quad (16)
\]
where \(0 < \theta < 1\). The equality holds if and only if \(x_1 = x_2\), indicating that
\[
(1 - \theta)x_1^2 + \theta x_2^2 \geq x_1^2 - 2\theta x_1^2. \quad (17)
\]

Then, we have
\[
(1 - \theta)\|a\|_2^2 + \theta \|b\|_2^2 \geq \|a\|_2^2 - 2\theta \|b\|_2^2, \quad (18)
\]
where the equality holds if and only if \(a = b\). □

**Lemma 2.** Given an optimization problem:

\[
\min_Z \|S \Phi(Z)\|_F^2, \quad \text{s.t.} \quad Z \in \mathcal{F},
\]

where \(\Phi(Z)\) is a function of \(Z\), \(\mathcal{F}\) is the feasible region, and \(S\) is a diagonal matrix whose \(i\)-th diagonal element is \(1/\|\Phi(Z_0)\|_2^{1-p/2}\) (\(Z_0\) could be any object in \(\mathcal{F}\). \(\Phi(Z_0)_i\) is the \(i\)-th row vector of \(\Phi(Z_0)\) and \(0 < p \leq 2\)), we have

\[
\|\Phi(Z')\|_{2,p} \leq \|\Phi(Z_0)\|_{2,p},
\]

where \(Z'\) is the optimal solution of Eqn. (19) and the equality holds if and only if \(\Phi(Z') = \Phi(Z_0)\).

**Proof.** Since \(Z'\) is the optimal solution, we have

\[
\|S \Phi(Z')\|_F^2 \leq \|S \Phi(Z_0)\|_F^2.
\]

Then

\[
\sum_i \frac{\|\Phi(Z')_i\|_2^p}{\|\Phi(Z_0)_i\|_2^p} \leq \sum_i \|\Phi(Z_0)_i\|_2^p.
\]

where \(\Phi(Z_0)_i\) and \(\Phi(Z')_i\) are the \(i\)-th row vector of \(\Phi(Z_0)\) and \(\Phi(Z')\), respectively.

According to Lemma 1, we have

\[
\left(1 - \frac{p}{2}\right)\|\Phi(Z_0)\|_2^2 + \frac{p}{2} \|\Phi(Z')_i\|_2^2 \\
\geq \|\Phi(Z_0)\|_2^2 - p \|\Phi(Z')_i\|_2^p.
\]

Then dividing the both sides by \(\|\Phi(Z_0)\|_2^2 - p \|\Phi(Z')_i\|_2^p\), we have

\[
\|\Phi(Z')_i\|_2^p \leq \left(1 - \frac{p}{2}\right)\|\Phi(Z_0)\|_2^2 + \frac{p}{2} \frac{\|\Phi(Z')_i\|_2^2}{\|\Phi(Z_0)\|_2^{2-p}}
\]

It indicates that

\[
\sum_i \|\Phi(Z')_i\|_2^p \leq \left(1 - \frac{p}{2}\right)\sum_i \|\Phi(Z_0)\|_2^p + \frac{p}{2} \sum_i \frac{\|\Phi(Z')_i\|_2^2}{\|\Phi(Z_0)\|_2^{2-p}}.
\]

Combining Eqns. (21) and (24), we obtain

\[
\sum_i \|\Phi(Z')_i\|_2^2 \leq \sum_i \|\Phi(Z_0)\|_2^2.
\]

Therefore,

\[
\|\Phi(Z')\|_{2,p} \leq \|\Phi(Z_0)\|_{2,p},
\]

where the equality holds if and only if \(\Phi(Z') = \Phi(Z_0)\). □

**Theorem 1.** The sequence \(\{W^k\}\) produced via Algorithm 1 has the following properties: \(\|W^k\|_{2,p}\) is non-increasing at successive iteration steps and \(\{\|W^k\|_{2,p}\}\) converges to a limited value.

**Proof.** Supposing we have obtained the solution \(U^k\), \(E^k\), and the objective function \(W^k\) at the \((k + 1)\)-th iteration step, we solve the optimization problem \(\min_u \|\Sigma^k(\mathcal{P}U + G)\|_F^2\) to obtain \(U^{k+1}\) by fixing \(E^k\). According to Lemma 2, we have

\[
\|V^k\|_{2,s} = \left\|PU^{k+1} + Q + \left[\begin{array}{c}LE^k \vdots \mathbf{0} \end{array}\right]\right\|_{2,p} \\
\leq \|W^k\|_{2,p}.
\]

Then we fix \(U^{k+1}\), and solve the optimization problem \(\min_{E} \|A^k(LE + H)\|_F^2\) to obtain \(E^{k+1}\). According to Lemma 2, we have

\[
\|W^{k+1}\|_{2,p} = \left\|PU^{k+1} + Q + \left[\begin{array}{c}LE^{k+1} \vdots \mathbf{0} \end{array}\right]\right\|_{2,p} \\
\leq \|V^k\|_{2,p}.
\]

Combining (27) and (28), we obtain

\[
\|W^{k+1}\|_{2,p} \leq \|W^k\|_{2,p},
\]

where the equality holds if and only if \(W^{k+1} = V^k = W^k\). Since the lower bound of \(\|W\|_{2,p}\) is limited, \(\{\|W^k\|_{2,p}\}\) will converge. □

**Theorem 2.** If sequences \(\{W^k\}\) and \(\{E^k\}\) produced in Algorithm 1 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.

Due to space limitations, the \(p\) supplementary document (Peng & Fan, 2016).

It is worth noting that Lemma 2 works for sparsity regularization based feature selection algorithms, i.e.,

\[
\min_{Z} \sum_i \|A_i Z + B_i\|_2 + \lambda \|Z\|_{2,p},
\]

where \(\lambda\) is the regularization coefficient.

Eqn. (30) can be reformulated as

\[
\min_{Z} \|[(A_1 Z + B_1)^T, ..., (A_i Z + B_i)^T, ..., A_i Z\|^2_{2,p}.
\]

The Robust Feature Selection (RFS) (Nie, et al., 2010; Xiang, et al., 2012) is a special case of Eqn. (30) when \(i = 1\) and \(p = 1\).
5 Experiments

5.1 Experimental datasets and settings

The proposed algorithm has been evaluated based on 6 publicly available datasets. In particular, 2 datasets were obtained from UCI, including ISOLET and SEMEION. ISOLET is a speech recognition data set with 7797 samples from 26 classes, and each sample has 617 features; SEMEION contains 1593 handwritten images from ~80 people, stretched in a rectangular box of $16 \times 16$. For these 2 datasets, the number of features is less than the number of their samples. Another 2 datasets were microarray data, including LUNG and CLSUB-111. In particular, LUNG consists of 203 samples, each having 3312 genes with standard deviations larger than 50 expression units (Bhattacharjee et al., 2001); CLSUB-111 consists of 111 samples with 11340 features (featureselection.asu.edu). Our algorithm has also been validated based on 2 image datasets, including UMIST and AR. In particular, UMIST includes face images with a resolution of $56 \times 46$ from 20 different people, and AR has 130 samples with 2400 features.

We compared our method with sparsity regularization based feature selection methods, including $\ell_1$-SVM (Bradley & Mangasarian, 1998; Mangasarian, 2006) and $\ell_2.1$-norm based Robust Feature Selection (RFS) (Nie et al., 2010; Xiang et al., 2012), with an objective function $J(W) = \|XW - Y\|_{2,1} + \lambda \|W\|_{2,1}$. We also compared our algorithm with well-known filter methods, including ReliefF (Kira & Rendell, 1992) and mRMR (Peng et al., 2005).

The feature selection methods were evaluated based on their classification accuracy. Particularly, linear SVM (Chang & Lin, 2011) was adopted to build classifiers based on the selected features. The parameter $C$ of linear SVM classifiers were tuned using a cross-validation strategy by searching a candidate set $\{10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. The regularization parameter of $\ell_1$-SVM and RFS were tuned using the same cross-validation strategy by searching a candidate set $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10^1, 10^2, 10^3\}$.

In our experiments, we first normalized all the data to have 0 mean and unit standard deviation for each feature. 10 trials were performed on each dataset. In each trial, the samples of each dataset were randomly split into training and testing subsets with a ratio of 6:4. For tuning parameters, a 3-fold was used for datasets with less than 200 training samples, and an 8-fold cross-validation was used for other datasets.

5.2 Effect of parameter $p$

To investigate how the classification performance is affected by the parameter $p$ of DSO-FS, we performed experiments based on datasets ISOLET, LUNG, and UMIST. We obtained solutions of $W$ with different settings of $p \in [0.1, 0.3, 0.5, 0.7, 0.9, 1]$ using Algorithm 1 based on the above 3 datasets, and then selected top-ranked features according to $\ell_2$-norm $\|W\|_2$ to build classifiers. Figure 1 shows the classification accuracy as a function of the number of the selected features and the parameter $p$. The results shown in Figure 1 indicated that $p$ played an important role in the classification.

5.3 Comparisons with state-of-the-art methods

Since the classification performance of linear SVM classifiers built on the features selected by our method is hinged on the parameter $p$, we used a cross-validation strategy to select an optimal value from $[0.1, 0.3, 0.5, 0.7, 0.9, 1]$ for our method. Figure 2 shows the average classification performance of classifiers built on the features selected by different methods in 10 trials. In particular, the average classification accuracy is shown as a function of the number of features used in the classification model. Compared with other methods, the proposed method achieved higher classification accuracy on most datasets, indicating that our method had overall better performance than other algorithms.

Table 1 summarizes mean and standard deviation of the classification rates in 10 trials for classifiers built on the top 100 features selected by the algorithms under comparison. These results demonstrated that our algorithm had the overall best classification accuracy on all the 6 datasets.

![Figure 1](image_url)  
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.
proposed an regression of class labels (Bartlett, et al., 2004). W
features, rather than data-fitting loss functions built upon the
could improve the generalization ability of the selected
constraints to obtain large between-class margins, which
feature selection method adopts data-fitting inequality
trade-off between a data-fitting loss function term and a
regularization based algorithms

6 Conclusions and Discussions

In this paper, a novel feature selection algorithm via
direct sparsity optimization was proposed. Our method
directly optimizes a larget margin linear classification
model’s sparsity with \( \ell_{2,p} \)-norm (\( 0 < p \leq 1 \)) subject to a
data-fitting constraint, different from the sparse
regularization based algorithms that typically take a
trade-off between a data-fitting loss function term and a
sparsity term. Since little is known about such a trade-off’s
impact on the feature selection, empirical parameter tuning is
typically adopted for choosing the trade-off parameter.
Our feature selection method adopts data-fitting inequality
constraints to obtain large between-class margins, which
could improve the generalization ability of the selected
features, rather than data-fitting loss functions built upon the
regression of class labels (Bartlett, et al., 2004). We also
proposed an efficient algorithm to solve the non-convex
( \( 0 < p < 1 \) ) and non-smooth optimization problem
associated with the feature selection problem. Extensive
experiments based on 6 datasets have demonstrated that
the proposed method could achieve better performance
than state-of-the-art feature selection algorithms. Furthermore, our algorithm can be easily extended for
solving other sparsity regularization algorithms. In
particular, our algorithm could be used to solve \( \ell_0 \) and
\( \ell_{2,0} \) based optimization problems subject to linear
constraints by setting \( p \) close to 0. Furthermore, similar to
SVM, we could also add soft-margin to our algorithm for
better performance in the future work.

Acknowledgments

This work was supported in part by National Key Basic
Research and Development Program of China (No.
2015CB856404), NSFC (No. 81271514, 61473296), and
NIH grants MH070365 and AG014971.
References


