Feature Selection via Scaling Factor Integrated Multi-Class Support Vector Machines

Jinglin Xu, Feiping Nie, Junwei Han
Northwestern Polytechnical University, Xian, 710072, P. R. China
{xujinglinlove, feipingnie, junweihan2010}@gmail.com

Abstract

In data mining, we often encounter high dimensional and noisy features, which may not only increase the load of computational resources but also result in the problem of model overfitting. Feature selection is often adopted to address this issue. In this paper, we propose a novel feature selection method based on multi-class SVM, which introduces the scaling factor with a flexible parameter to renewedly adjust the distribution of feature weights and select the most discriminative features. Concretely, the proposed method designs a scaling factor with $\frac{p}{2}$ power to control the distribution of weights adaptively and search optimal sparsity of weighting matrix. In addition, to solve the proposed model, we provide an alternative and iterative optimization method. It not only makes solutions of weighting matrix and scaling factor independently, but also provides a better way to address the problem of solving $\ell_{2,0}$-norm. Comprehensive experiments are conducted on six datasets to demonstrate that this work can obtain better performance compared with a number of existing state-of-the-art multi-class feature selection methods.

1 Introduction

With the rapid advance of computer techniques, immense quantities of high dimensional data have been yielded in many applications, such as computer vision, pattern recognition and data mining, which pose unprecedented challenges for learning tasks due to the curse of dimensionality. Those high dimensional, complex and irrelevant features often make learning models tend to over-fitting and become less comprehensible. In order to utilize this kind of data effectively, feature selection is an essential component to successful data mining [Liu and Motoda, 2007]. As a process of selecting a subset of original features or learning the feature weighting according to certain criteria, feature selection is an important and frequently used dimensionality reduction technique which not only reduces the number of features, removes irrelevant, redundant, or noisy data, but also brings the immediate effects for real applications, such as gait recognition [Kusakunniran *et al.*, 2010], expression recognition [Gupta and Xiao, 2011], image annotation [Huang *et al.*, 2012] and disease diagnosis [Zhu *et al.*, 2014].

During the past few decades, researchers have developed a large amount of feature selection algorithms. Based on the different computational models, feature selection methods can be roughly classified into three categories [Guyon and Elisseeff, 2003; Zhao *et al.*, 2010]: filter, wrapper and embedded models. The filter model utilizes the intrinsic properties or some global statistical information of the data, and evaluates features without any learning mechanism. Some popular feature selection methods based on filter model include Fisher Score (FS) [Duda *et al.*, 1973], ReliefF (RF) [Liu and Motoda, 2007], mRMR (MR) [Peng *et al.*, 2005], T-test (Ttest) [Montgomery *et al.*, 2009], Chi-Square (Chi) [Liu and Setiono, 1995], Information Gain (IG) [Cover and Thomas, 1991] and so on. Most of these methods are evaluated independently for each feature, therefore they cannot deal with redundant features effectively. The wrapper model [Kohavi and John, 1997] uses a given learning algorithm to select features according to its evaluation criteria. In spite of good performance, they often have expensive computational cost. The embedded model [Lal *et al.*, 2006] performs feature selection as a part of the training process, and often shows good performance since it is coupled with specific classifier. These different models are designed from different perspectives and all have their own strengths and weaknesses. In this paper, we propose a novel method based on embedded model for performing supervised feature selection.

Feature selection methods based on embedded model are coupled with specific classifiers, where Support Vector Machines (SVMs) are the most commonly used classifier. One of the standard embedded methods with SVMs is Recursive Feature Elimination (RFE) [Guyon *et al.*, 2002] which is based on the idea that the importance of a feature should be related to the magnitude of its weight and the feature with the smallest magnitude is removed. In order to reduce the time consumption, ℓ_1-SVM [Mangasarian, 2006] was proposed, which can obtain a sparse solution by changing the ℓ_2-norm regularization in SVM to an ℓ_1-norm. However, considering that the number of selected features using ℓ_1-SVM is upper bounded by the sample size, a Hybrid Huberized SVM (HHSVM) [Wang *et al.*, 2007] was presented via combining both ℓ_1-norm and ℓ_2-norm to form a more structured regularization. Unfortunately, the above methods were designed only for binary classification. Actually many classification problems in practice involve many categories, thus some works [Hou *et al.*, 2011; Yang *et al.*, 2011; He *et al.*, 2012; Han and Kim, 2015] were developed to deal with multi-class feature selection problem based on other models (such as iteratively re-weighted Least Squares and Half-quadratic optimization).

Although original SVMs can be extended to the multi-class case via one vs one or one vs all strategy, it ignores the correlation between classes, since the extended strategy just simply breaks the multi-class problem into several independent binary classification problems. To overcome this drawback, multi-class SVMs [Fan *et al.*, 2008; Chapelle and Keerthi, 2008] are invented to consider all classes simultaneously by solving a unified optimization problem. [Obozinski *et al.*, 2006] proposed $\ell_{2,1}$-norm based a multi-task learning method to take advantage of the structural sparsity to select the discriminative features across multi-class. However, this work used Least Square loss function instead of the hinge loss function, where the latter is usually better than the former in terms of feature selection and classification tasks [Cristianini and Shawe-Taylor, 2000]. Recently, the work [Cai *et al.*, 2011] proposed a new feature selection method that adopted the multi-class hinge loss with a structured regularization term for all classes without requiring further heuristic strategy. Although this work has been tackled based on the existing method, it doesn't consider scaling flexibility of each feature, and thus can not automatically eliminate irrelevant features efficiently.

The large number of features and the small number of data samples form the serious challenge for classification. To address this challenge more efficiently, this paper develops a novel and general model based on multi-class SVM, which integrates a scaling factor $\boldsymbol{\theta}$ with $\frac{p}{2}$ power to renewedly adjust the distribution of feature weights and select the most discriminative features. During the process of weight allocation on each feature, the proposed model also shows its property. Concretely, when p is close to infinity, the property of $\ell_{2,\frac{2}{1+p}}$-norm is closest to that of $\ell_{2,0}$-norm. However, solving $\ell_{2,\frac{2}{1+p}}$-norm in the proposed model is more feasible than solving $\ell_{2,0}$-norm. When p is equal to one, the property of $\ell_{2,\frac{2}{1+p}}$-norm is equivalent to that of $\ell_{2,1}$-norm which is structured sparse. In addition, to solve the proposed model, we provide an alternative and iterative optimization method. Specifically, in the process of optimization, through mathematical transformations and Larange multiplier method, our optimization makes solutions of weighting matrix $\mathbf{W}$ and scaling factor $\boldsymbol{\theta}$ independently, which not only better addresses the difficulty of solving $\ell_{2,0}$-norm but also searches optimal p to ensure the proper sparsity of weighting matrix. What's more, the theoretical analysis including convergence is also provided. Comprehensive experiments are conducted on six datasets to demonstrate that this work can obtain better performance compared with a number of existing state-of-the-art multi-class feature selection methods.

2 Related Work

In this section, we briefly review some feature selection methods based on multi-class SVM [Chapelle and Keerthi, 2008; Fan *et al.*, 2008; Cai *et al.*, 2011] which are very close to the proposed methods.

2.1 Multi-class SVM by Crammer and Singer

The index i runs over the data samples $(i = 1,\ldots,n)$, v runs over the features $(v = 1,\ldots,d)$, and k runs over the classes $(k = 1,\ldots,c)$. $\mathbf{W} = (\mathbf{w}_1,\ldots,\mathbf{w}_c) \in \mathbf{R}^{d\times c}$ is the projection matrix. $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_n) \in \mathbf{R}^{d\times n}$ is the data matrix. $y_i \in \{1,\ldots,c\}$ is the label for each sample. The efficient approach developed by Crammer's formulation [Crammer and Singer, 2001] can be adopted to solve them, which is equivalent to solving the following optimization problem:

$$\begin{aligned}&\min_{\mathbf{w}_k,\xi_i}\frac{1}{2}\sum_{k=1}^{c}\mathbf{w}_k^T\mathbf{w}_k+\alpha\sum_{i=1}^{n}\xi_i\\&s.t.\mathbf{w}_{y_i}^T\mathbf{x}_i-\mathbf{w}_k^T\mathbf{x}_i\geq 1-\xi_i,\forall k\neq y_i,\xi_i\geq 0,i=1\ldots n\end{aligned}\tag{1}$$

where α is the regularization parameter to balance the loss and penalty, and the bias b_k for each class is absorbed by augmenting the vector $\mathbf{w}_k^T$ and each sample $\mathbf{x}_i$ with an additional dimension: $\mathbf{w}_k^T \leftarrow [\mathbf{w}_k^T, b_k], \mathbf{x}_i^T \leftarrow [\mathbf{x}_i^T, 1]$. The format of problem (1) can be written as:

$$\min_{\mathbf{W}}\sum_{i=1}^{n}(1-\mathbf{w}_{y_i}^T\mathbf{x}_i+\max_{k\neq y_i}\mathbf{w}_k^T\mathbf{x}_i)_+ +\beta\sum_{k=1}^{c}\|\mathbf{w}_k\|_2^2 \tag{2}$$

Given data matrix $\mathbf{X}$ and label matrix $\mathbf{Y} = \{y_{ik}\} \in \mathbf{R}^{n\times c}$, problem (2) can be generalized as:

$$\min_{\mathbf{W}}\frac{1}{2}\|\mathbf{W}\|_F^2 + Cf(\mathbf{W}^T\mathbf{X},\mathbf{Y}) \tag{3}$$

2.2 Multi-class SVM with $\ell_{2,1}$-norm Regularization

Recently, considering the superiority of $\ell_{2,1}$-norm in terms of [Nie *et al.*, 2010], the work [Cai *et al.*, 2011] combined the multi-class hinge loss with $\ell_{2,1}$-norm regularization term for the feature selection:

$$\min_{\mathbf{W}}\frac{1}{2}\|\mathbf{W}\|_{2,1} + Cf(\mathbf{W}^T\mathbf{X},\mathbf{Y}) \tag{4}$$

where $\|\mathbf{W}\|_{2,1} = \sum_{i=1}^{d}\|\mathbf{w}^i\|_2$, and $\mathbf{w}^i$ and $\mathbf{w}_j$ denote the i-th row and the j-th column vector of matrix $\mathbf{W}$, respectively. This work combines multi-class hinge loss with $\ell_{2,1}$-norm regularization and is solved by changing $\ell_{2,1}$-norm to trace optimization on $\mathbf{W}$.

3 The Proposed Method

3.1 Formulation

In this paper, we tend to learn c classifications functions $f_k(\mathbf{x}) = \mathbf{w}_k^T\mathbf{x}, 1 \leq k \leq c$ and perform simultaneous feature selection, namely find a small set of features which are good for all the classifiers. Considering the flexibility of each feature and making $\mathbf{W}$ with optimal sparsity, we propose a novel framework of scaling factors for performing feature selection in the multi-class cases:

$$\begin{aligned}&\min_{\mathbf{W},\boldsymbol{\theta}}\frac{1}{2}\|\mathbf{W}\|_F^2+C\sum_{k=1}^{c}\sum_{i=1}^{n}(1-y_{ik}\sum_{v=1}^{d}\theta_v^{\frac{p}{2}}w_{vk}x_{vi})_+\\&s.t.\boldsymbol{\theta}\geq 0,\boldsymbol{\theta}^T\mathbf{1}=1\end{aligned}\tag{5}$$

where $\boldsymbol{\theta}$ is a vector which consists of d scaling factors $\theta_v(v=1...d)$, and $p\geq 1$ is the power exponent of θ_v. In the following part, we will further analyze how to solve this framework.

3.2 Optimization

Theorem 1. *Solving problem (5) can be transformed into solving the following problem:*

$$\min_{\mathbf{W}}\frac{1}{2}\|\mathbf{W}\|^2_{2,\frac{2}{1+p}}+C\sum_{k=1}^{c}\sum_{i=1}^{n}(1-y_{ik}\sum_{v=1}^{d}w_{vk}x_{vi})_+\tag{6}$$

Proof. Let $\tilde{w}_{vk}=\theta_v^{\frac{p}{2}}w_{vk}$, then $w_{vk}=\theta_v^{\frac{-p}{2}}\tilde{w}_{vk}$. After variable substitutions, problem (5) is equivalent to:

$$\begin{aligned}&\min_{\tilde{\mathbf{W}},\boldsymbol{\theta}}\frac{1}{2}\sum_{k=1}^{c}\sum_{v=1}^{d}\frac{\tilde{w}_{vk}^2}{\theta_v^p}+C\sum_{k=1}^{c}\sum_{i=1}^{n}(1-y_{ik}\sum_{v=1}^{d}\tilde{w}_{vk}x_{vi})_+\\&s.t.\boldsymbol{\theta}\geq 0,\boldsymbol{\theta}^T\mathbf{1}=1\end{aligned}\tag{7}$$

Making the change of variables $w_{vk}\leftarrow\tilde{w}_{vk}$, problem (7) can be rewritten as:

$$\begin{aligned}&\min_{\mathbf{W},\boldsymbol{\theta}}\frac{1}{2}\sum_{k=1}^{c}\sum_{v=1}^{d}\frac{w_{vk}^2}{\theta_v^p}+C\sum_{k=1}^{c}\sum_{i=1}^{n}(1-y_{ik}\sum_{v=1}^{d}w_{vk}x_{vi})_+\\&s.t.\boldsymbol{\theta}\geq 0,\boldsymbol{\theta}^T\mathbf{1}=1\end{aligned}\tag{8}$$

For updating scaling factors $\boldsymbol{\theta}$, we fix other variables and simplify the first term of problem (8) as follows:

$$\min_{\substack{\boldsymbol{\theta}^T\mathbf{1}=1\\\boldsymbol{\theta}\geq 0}}\sum_{k=1}^{c}\sum_{v=1}^{d}\frac{w_{vk}^2}{\theta_v^p}\Leftrightarrow\min_{\substack{\boldsymbol{\theta}^T\mathbf{1}=1\\\boldsymbol{\theta}\geq 0}}\sum_{v=1}^{d}\frac{\|\mathbf{w}^v\|_2^2}{\theta_v^p}\Leftrightarrow\min_{\substack{\boldsymbol{\theta}^T\mathbf{1}=1\\\boldsymbol{\theta}\geq 0}}\sum_{v=1}^{d}\theta_v^r h_v\tag{9}$$

where $r=-p(r\leq -1)$ and $h_v=\sum_{k=1}^{c}w_{vk}^2=\|\mathbf{w}^v\|_2^2$. The Lagrangian function of problem (9) is:

$$L(\theta_v,\lambda,\mu_v)=\sum_{v=1}^{d}\theta_v^r h_v-\lambda(\sum_{v=1}^{d}\theta_v-1)-\sum_{v=1}^{d}\mu_v\theta_v\tag{10}$$

where λ and $\mu_v\geq 0$ are the Lagrangian multipliers.

Taking derivatives of $L(\theta_v,\lambda,\mu_v)$ with respect to θ_v and setting it to zero, and then according to the KKT condition $\mu_v\theta_v=0$, we can calculate the solution of problem (9):

$$\theta_v=\frac{(h_v)^{\frac{1}{1-r}}}{\sum_{s=1}^{d}(h_s)^{\frac{1}{1-r}}}\tag{11}$$

Substituting original variables (p and $\mathbf{w}^v$) back to Eq.(11), there is:

$$\theta_v=\frac{\left(\|\mathbf{w}^v\|_2^2\right)^{\frac{1}{1+p}}}{\sum_{s=1}^{d}\left(\|\mathbf{w}^s\|_2^2\right)^{\frac{1}{1+p}}}\tag{12}$$

Algorithm 1 : Solving problem (5)

```
Input: Data $\mathbf{X}\in\mathbf{R}^{d\times n}$, label $\mathbf{Y}\in\mathbf{R}^{n\times c}$, parameters $C$ and $p$.
Initialization: Set $t=0$.
    Initialize $\mathbf{W}^{(0)}=\{w_{vk}=1\}$.
    Initialize scaled factor $\boldsymbol{\theta}$, where its $v$-th element is $\theta_v=\frac{1}{c}$.
Repeat
    1: Calculate scaled factor of each feature $\boldsymbol{\theta}_{t+1}$ by Eq.(12).
    2: Update feature weighting matrix $\mathbf{W}_{t+1}$ by solving problem (15) using Crammer's algorithm [Fan et al., 2008].
    3: $t=t+1$
Until converges.
Output: $\mathbf{W}$ and $\boldsymbol{\theta}$.
```

Using Eq.(12), the objective function of the first term in problem (8) can be further rewritten as:

$$\sum_{v=1}^{d}\frac{\|\mathbf{w}^v\|_2^2}{\theta_v^p}=\big(\sum_{v=1}^{d}\|\mathbf{w}^v\|_2^{\frac{2}{1+p}}\big)^{1+p}=\|\mathbf{W}\|^2_{2,\frac{2}{1+p}}\tag{13}$$

Therefore, we can arrive at problem (6) after submitting Eq.(13) into problem (8). This completes the proof and realizes rewriting problem (5) without an explicit $\boldsymbol{\theta}$ by using $\mathbf{W}$. □

For the proposed method, we utilize an alternative and iterative method to solve $\mathbf{W}$ and $\boldsymbol{\theta}$ in problem (5). When $\mathbf{W}$ is fixed, we can use Eq.(12) to update $\boldsymbol{\theta}$, and then update $\mathbf{W}$ when $\boldsymbol{\theta}$ has been obtained. When $\boldsymbol{\theta}$ is fixed, problem (5) becomes a classical multi-class SVM model and is simplified as:

$$\min_{\mathbf{W}}\frac{1}{2}\|\mathbf{W}\|_F^2+C\sum_{k=1}^{c}\sum_{i=1}^{n}(1-y_{ik}\sum_{v=1}^{d}w_{vk}\tilde{x}_{vi})_+\tag{14}$$

where $\tilde{x}_{vi}=\theta_v^{\frac{p}{2}}x_{vi}$ is known. Given $\widetilde{\mathbf{X}}=(\tilde{x}_{vi})$, problem (14) can be generalized as:

$$\min_{\mathbf{W}}\frac{1}{2}\|\mathbf{W}\|_F^2+Cf(\mathbf{W}^T\widetilde{\mathbf{X}},\mathbf{Y})\tag{15}$$

which can be solved by using [Fan *et al.*, 2008; Cai *et al.*, 2011]. For clarity, we summarize our method in Algorithm 1.

In above theoretical derivation, we transfer problem (5) into problem (6) which is a novel framework with flexible sparsity realized by tuning p in range of $[1,\infty)$. Concretely, when $p\to\infty$, $\|\mathbf{W}\|_{2,\frac{2}{1+p}}$ is closest to $\|\mathbf{W}\|_{2,0}$. However, solving $\|\mathbf{W}\|_{2,\frac{2}{1+p}}$ in the proposed model is much easier than solving $\|\mathbf{W}\|_{2,0}$. When $p=1$, $\|\mathbf{W}\|_{2,\frac{2}{1+p}}$ degrades as $\|\mathbf{W}\|_{2,1}$ which is structured sparse. At this time, problem (5) becomes:

$$\begin{aligned}&\min_{\mathbf{W},\boldsymbol{\theta}}\frac{1}{2}\|\mathbf{W}\|_F^2+C\sum_{k=1}^{c}\sum_{i=1}^{n}(1-y_{ik}\sum_{v=1}^{d}\sqrt{\theta_v}w_{vk}x_{vi})_+\\&s.t.\boldsymbol{\theta}\geq 0,\boldsymbol{\theta}^T\mathbf{1}=1\end{aligned}\tag{16}$$

which can be transformed into:

$$\min_{\mathbf{W}}\frac{1}{2}\|\mathbf{W}\|_{2,1}^2+C\sum_{k=1}^{c}\sum_{i=1}^{n}(1-y_{ik}\sum_{v=1}^{d}w_{vk}x_{vi})_+\tag{17}$$

Table 1: Datasets summary

Datasets	Property		
	♯Classes(c)	♯Samples(n)	♯Features(d)
LUNG	5	203	3312
CAR	11	174	9182
BrainT2	4	50	10367
SRBCT	4	83	2229
GLIOM	4	50	4434
MLLML	3	72	12582

For problem (16), we also use the alternative and iterative method to solve $\mathbf{W}$ and $\boldsymbol{\theta}$. If $\mathbf{W}$ is fixed, we can use Eq.(12) with $p=1$ to update $\boldsymbol{\theta}$, and then solve problem (15) to update $\mathbf{W}$ if $\boldsymbol{\theta}$ has been calculated.

3.3 Convergence

Theorem 2. *The Algorithm 1 monotonically decreases the objective of problem (5) in each iteration and converges to the optimum of the problem.*

Proof. Suppose that at the t-th iteration, we have obtained $\mathbf{W}_t$ and $\boldsymbol{\theta}_t$. In the t+1-th iteration, the updated $\mathbf{W}$ is denoted as $\mathbf{W}_{t+1}$. Considering that problem (5) can be rewritten as problem (6), we have:

$$\mathbf{W}_{t+1} = \arg\min_{\mathbf{W}} \frac{1}{2}\|\mathbf{W}_{t+1}\|^2_{2,\frac{2}{1+p}} + Cf(\mathbf{W}^T_{t+1}\mathbf{X}, \mathbf{Y}) \quad (18)$$

Referring to the way of argumentation for [Chang *et al.*, 2014], through rewriting Eq.(18) we have:

$$\begin{aligned} &\frac{1}{2}\|\mathbf{W}_{t+1}\|^2_{2,\frac{2}{1+p}} + Cf(\mathbf{W}^T_{t+1}\mathbf{X}, \mathbf{Y}) \\ &\leq \frac{1}{2}\|\mathbf{W}_{t}\|^2_{2,\frac{2}{1+p}} + Cf(\mathbf{W}^T_{t}\mathbf{X}, \mathbf{Y}) \end{aligned} \quad (19)$$

Thus, Eq.(19) proves that problem (5) and its variant problem (6) are lower bounded and their objective function value decreases after each iteration. □

4 Experiment

4.1 Experiment Setup

In our experiments, six different public datasets are adopted to illustrate the performance of different feature selection methods. These datasets include five microarray datasets (LUNG, CAR, SRBCT, GLIOM and MLLML)[1] and one primary tumor dataset (BrainT2)[2]. These datasets are widely used by many previous feature selection methods such as [Padungweang *et al.*, 2012; Liu *et al.*, 2013; Hong *et al.*, 2016] to evaluate their performance. All datasets are standardized to be zero-mean and normalized by standard deviation, and satisfy that in each dataset the number of samples is much less than the number of the features. We show more information of the datasets in Table 1.

In order to demonstrate the effectiveness of the proposed method ($p > 1$, denoted as square $\ell_{2,\frac{2}{1+p}}$-norm (SL2P)) and its simple version ($p=1$, denoted as square $\ell_{2,1}$-norm (SL21)) for feature selection, we compare them with six classical feature selection methods (i.e. IG [Cover and Thomas, 1991], Ttest [Montgomery *et al.*, 2009], FS [Duda *et al.*, 1973], Chi [Liu and Setiono, 1995], RF [Liu and Motoda, 2007] and MR [Peng *et al.*, 2005]) as well as two widely used methods which are multi-class hinge loss with ℓ_2-norm regularization (denoted as CS) [Fan *et al.*, 2008] and multi-class hinge loss with $\ell_{2,1}$-norm regularization (denoted as L21) [Cai *et al.*, 2011]. Because our work concerns multi-class feature selection, we do not compare binary feature selection method (like ℓ_1-SVM).

In addition, for each dataset, we randomly sample 50% instances as training data and the remaining are used as test data. We tune parameters based on the training set and predict classification accuracy based on the test set. We choose the classification accuracy as metric, achieved by Crammer and Singer's SVM [Fan *et al.*, 2008] classifier using the selected features. The numbers of selected features are set as $\{10,20,30,40,50,60,70,80\}$, respectively. A good feature selection algorithm should select a few of features that result in high classification accuracy.

4.2 Experiment Results

In our experiments, we utilize the Crammer's method [Fan *et al.*, 2008] for solving multi-class SVM to evaluate the effectiveness of selected features. As shown in Table 2 and 3, in terms of the classification accuracy, it is obvious that proposed methods (SL21 and SL2P) outperform all other feature selection methods on six datasets.

In Tables 2 and 3, for almost all the datasets, the proposed SL2P method can achieve significant improvements compared with six filter feature selection methods. These results demonstrate that incorporating feature selection as a part of the training process and obtaining feature weighting analytically from a learning mechanism are more efficient. In contrast, filter feature selection methods select features without involving any learning mechanism.

Furthermore, compared with two multi-class SVM methods (CS and L21), the proposed SL21 method obtains better results on LUNG, SRBCT, GLIOM and MLLML datasets. For example, in Table 2, the proposed SL21 method averagely achieves 12.6316%, 12.6984%, 4.1667% and 12.7451% improvements, respectively. Similarly, compared with CS and L21 methods, the proposed SL2P method can obtain the best results on LUNG, CAR, BrainT2, SRBCT, GLIOM and MLLML datasets. This is because the $\ell^2_{2,\frac{2}{1+p}}$-norm is more efficient than other norms (Frobenius-norm and $\ell_{2,1}$), by renewedly scaling factors and tuning parameter p to allocate each feature weight with optimal sparsity. For example, in Table 2, the SL2P method averagely achieves 11.5789%, 20.9459%, 14.2858%, 15.8730%, 12.5001% and 16.6667% improvements, respectively.

Besides, the proposed SL2P method can achieve higher classification accuracy than the proposed SL21 method. For example, in Table 3, it can achieve 3.1579%, 1.3513%,

[1] http://csse.szu.edu.cn/staff/zhuzx/Datasets.html

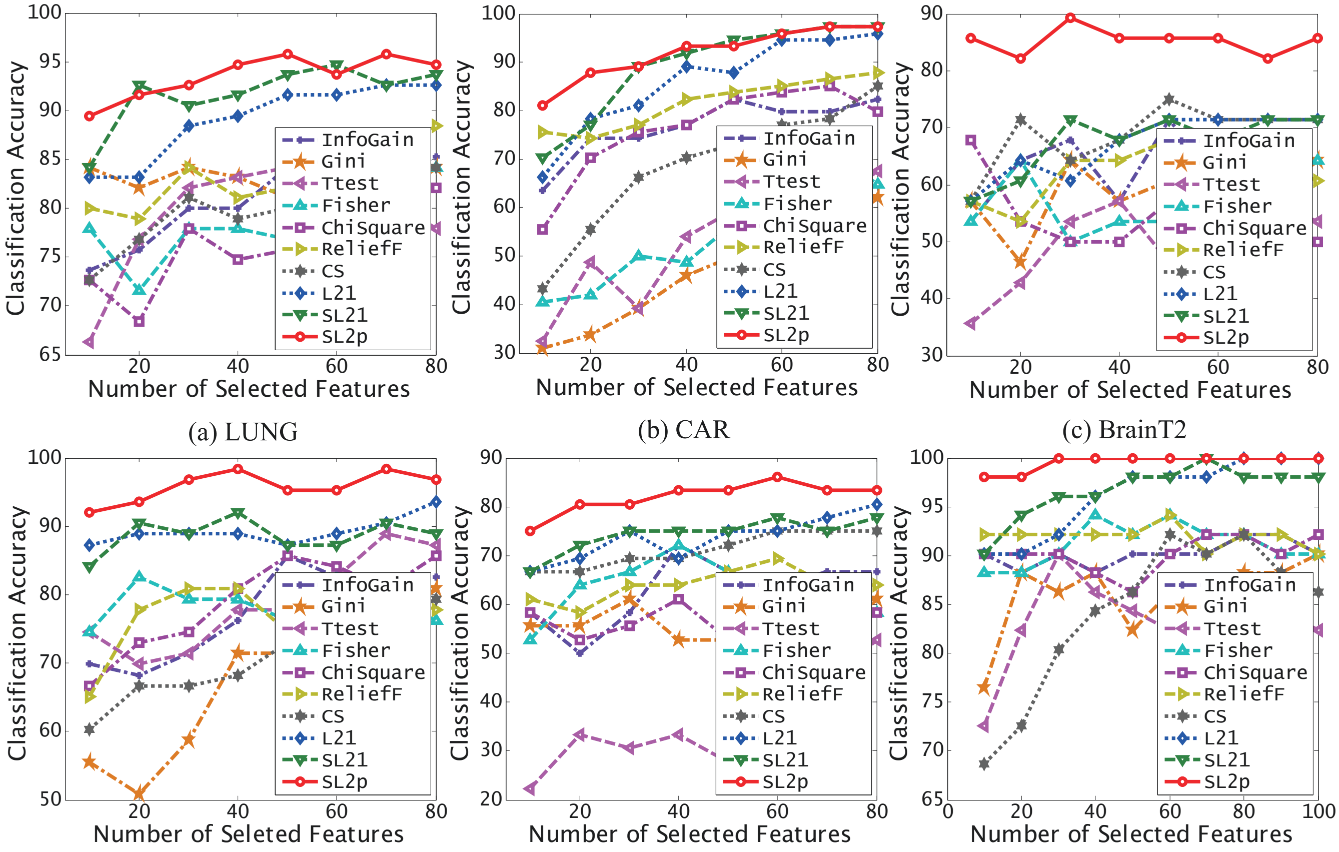
[2] https://archive.ics.uci.edu/ml/datasets/Primary+Tumor

Table 2: Classification accuracy of SVM as classifier on top 20 selected features.

	Methods									
Datasets	IG	Ttest	FS	Chi	RF	MR	CS	L21	SL21	SL2P
LUNG	75.7895	76.8421	71.5789	68.4211	78.9474	83.1579	76.8421	83.1579	**92.6316**	91.5789
CAR	74.3243	48.6486	41.8919	70.2703	74.3243	66.2162	55.4054	78.3784	77.0270	**87.8378**
BrainT2	64.2857	42.8571	64.2857	53.5714	53.5714	64.2857	71.4286	64.2857	60.7143	**82.1429**
SRBCT	68.2540	69.8413	82.5397	73.0159	77.7778	80.9524	66.6667	88.8889	90.4762	**93.6508**
GLIOM	50.0000	33.3333	63.8889	52.7778	58.3333	63.8889	66.6667	69.4444	72.2222	**80.5556**
MLLML	88.2353	82.3529	88.2353	90.1961	92.1569	88.2353	72.5490	90.1961	94.1176	**98.0392**

Table 3: Classification accuracy of SVM as classifier on top 40 selected features.

	Methods									
Datasets	IG	Ttest	FS	Chi	RF	MR	CS	L21	SL21	SL2P
LUNG	80.0000	83.1579	77.8947	74.7368	81.0526	86.3158	78.9474	89.4737	91.5789	**93.6842**
CAR	77.0270	54.0541	48.6486	77.0270	82.4324	67.5676	70.2703	89.1892	91.8919	**93.2432**
BrainT2	57.1429	57.1429	53.5714	50.0000	64.2857	64.2857	67.8571	67.8571	67.8571	**85.7143**
SRBCT	76.1905	77.7778	79.3651	80.9524	80.9524	80.9524	68.2540	88.8889	92.0635	**98.4127**
GLIOM	72.2222	33.3333	72.2222	61.1111	63.8889	63.8889	69.4444	69.4444	75.0000	**83.3333**
MLLML	88.2353	86.2745	94.1176	88.2353	92.1569	86.2745	84.3137	96.0784	96.0784	**100.000**



(a) LUNG

(b) CAR

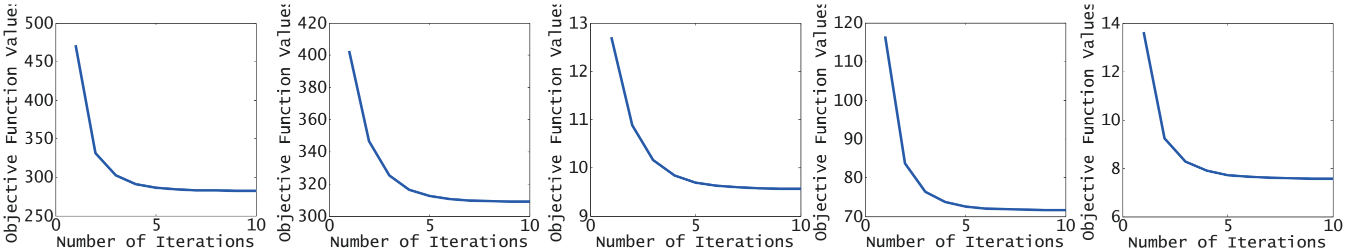
(c) BrainT2

(d) SRBCT

(e) GLIOM

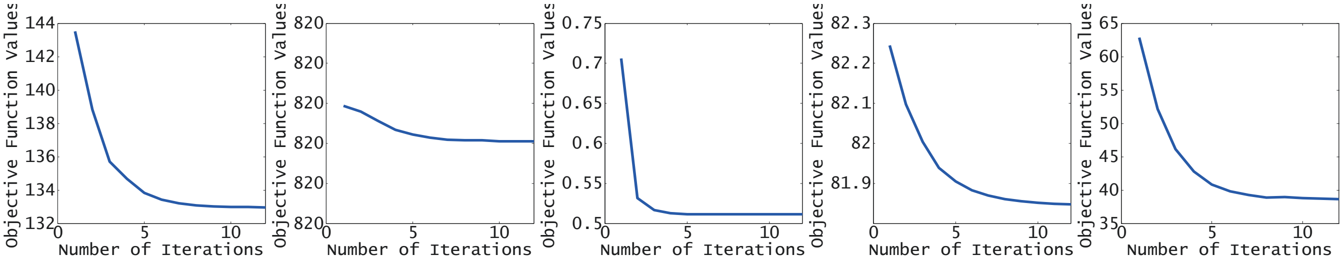
(f) MLLML

Figure 1: Comparisons of ten feature selection methods on LUNG, CAR, BrainT2, SRBCT, GLIOM and MLLML datasets, respectively, in terms of classification accuracy using SVM as classifier.



(a) LUNG (b) CAR (c) BrainT2 (d) GLIOM (e) MLLML

Figure 2: Convergence behaviors of the proposed method SL21 for top 40 features on LUNG, CAR, BrainT2, SRBCT, GLIOM and MLLML datasets, respectively.



(a) LUNG (b) CAR (c) BrainT2 (d) GLIOM (e) MLLML

Figure 3: Convergence behaviors of the proposed method SL2P for top 40 features on LUNG, CAR, BrainT2, SRBCT, GLIOM and MLLML datasets, respectively.

17.8572%, 6.3492%, 8.3333% and 3.9216% improvements, respectively on all datasets. These results illustrate that when flexibly tuning parameter p in $(1, \infty)$, the SL2P method can obtain more proper sparsity than the SL21 method. Thus, due to the introduction of parameter p, the SL2P method not only makes the optimal sparsity of the weighting matrix adaptively, but also naturally selects the most discriminative features.

Finally, we tune the number of selected features instead of only picking 20 or 40 features and show their experimental results in Figure 1. It can be seen that two proposed methods SL21 (green dotted line marked as lower triangular) and SL2p (red solid line marked as circle) can achieve higher classification accuracy and converge within 50 iterations. What's more, to validate the convergence analysis of the proposed SL21 and SL2P methods, we present their convergence behavior curves in Figures 2 and 3, respectively.

4.3 Parameters Setup

In our experiments, for six filter-based methods (IG, Ttest, F-S, Chi, RF and MR), their parameters can be tuned according to [Zhao *et al.*, 2010]. In the two multi-class SVMs methods (CS and L21), there is only one parameter C which can be adjusted on training data by grid search. Concretely, from 10^{-1} to 10^{1} with 0.1 step, we search for optimal parameters with the highest accuracy. For the proposed SL21 and SL2P methods, two parameters C and p need to be tuned. The parameter C controls the trade-off between discrimination and sparsity, and the parameter p emphasizes the proper sparsity of the feature weighting. They all play important roles in our methods. For $p = 1$, there is only one parameter C in Algorithm 1. For $p > 1$, there are two parameters C and p in Algorithm 1. We also vary these two parameters from 10^{-1} to 10^{1} with 0.1 step and search for optimal parameters corresponding to the highest accuracy, where their results satisfy constraints of $\boldsymbol{\theta}$ in the end. All classification results are conducted on each selected feature set from six datasets and methods by Crammer and Singer's SVM [Fan *et al.*, 2008].

5 Conclusion

To tackle the challenge of high feature dimensionality and noisy features in multi-class classification, this work proposes a novel and general feature selection framework, the square $\ell_{2,\frac{2}{1+p}}$-norm method, via flexibly scaling factor for each feature, where parameter p is the power exponent of scaling factor. The proposed methods capture a small number of more discriminative features across multiple classes by tuning different values of p, and shows the connection between $\ell_{2,\frac{2}{1+p}}$-norm method and $\ell_{2,1}$-norm method. Furthermore, the theoretical proof of the convergence has also been provided. Compared with the state-of-the-art methods, the proposed methods can always achieve the best performance.

Acknowledgments

This work was supported in part by the National Science Foundation of China under Grants 61522207 and 61473231.

References

[Cai *et al.*, 2011] Xiao Cai, Feiping Nie, Heng Huang, and Chris Ding. Multi-class $\ell_{2,1}$-norm support vector machine. In *ICDM*, pages 91–100, 2011.

[Chang *et al.*, 2014] X. Chang, F. Nie, Y. Yang, and H. Huang. A convex formulation for semi-supervised multi-label feature selection. In *AAAI*, pages 1171–1177, 2014.

[Chapelle and Keerthi, 2008] Olivier Chapelle and S Sathiya Keerthi. Multi-class feature selection with support vector machines. In *PASA*, 2008.

[Cover and Thomas, 1991] Thomas M Cover and Joy A Thomas. Elements of information theory. *New York*, 1991.

[Crammer and Singer, 2001] K. Crammer and Y. Singer. On the algorithmic interpretation of multiclass kernel-based vector machines. *JMLR*, 2(2):2001, 2001.

[Cristianini and Shawe-Taylor, 2000] Nello Cristianini and John Shawe-Taylor. An introduction to support vector machines, 2000.

[Duda *et al.*, 1973] Richard O Duda, Peter E Hart, David G Stork, et al. *Pattern classification*, volume 2. Wiley New York, 1973.

[Fan *et al.*, 2008] Rong-En Fan, Kai-Wei Chang, Cho-Jui H-sieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *JMLR*, 9:1871–1874, 2008.

[Gupta and Xiao, 2011] Mithun Das Gupta and Jing Xiao. Non-negative matrix factorization as a feature selection tool for maximum margin classifiers. In *CVPR*, pages 2841–2848, 2011.

[Guyon and Elisseeff, 2003] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *JMLR*, 3:1157–1182, 2003.

[Guyon *et al.*, 2002] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *ML*, 46(1-3):389–422, 2002.

[Han and Kim, 2015] Dongyoon Han and Junmo Kim. Unsupervised simultaneous orthogonal basis clustering feature selection. In *CVPR*, pages 5016–5023, 2015.

[He *et al.*, 2012] Ran He, Tieniu Tan, Liang Wang, and Wei-Shi Zheng. $l_{2,1}$ regularized correntropy for robust feature selection. In *CVPR*, pages 2504–2511, 2012.

[Hong *et al.*, 2016] Tao Hong, Hou Chenping, Nie Feiping, Jiao Yuanyuan, and Yi Dongyun. Effective discriminative feature selection with nontrivial solution. *TNNLS*, 27(4):796–808, 2016.

[Hou *et al.*, 2011] Chenping Hou, Feiping Nie, Dongyun Yi, and Yi Wu. Feature selection via joint embedding learning and sparse regression. In *IJCAI*, pages 1324–1329, 2011.

[Huang *et al.*, 2012] Heng Huang, C. Ding, Deguang Kong, and Haifeng Zhao. Multi-label relieff and f-statistic feature selections for image annotation. In *CVPR*, pages 2352–2359, 2012.

[Kohavi and John, 1997] Ron Kohavi and George H John. Wrappers for feature subset selection. *AI*, 97(1):273–324, 1997.

[Kusakunniran *et al.*, 2010] Worapan Kusakunniran, Qiang Wu, Jian Zhang, and Hongdong Li. Support vector regression for multi-view gait recognition based on local motion feature selection. In *CVPR*, pages 974–981, 2010.

[Lal *et al.*, 2006] Thomas Navin Lal, Olivier Chapelle, Jason Weston, and André Elisseeff. Embedded methods. In *FE*, pages 137–165. 2006.

[Liu and Motoda, 2007] Huan Liu and Hiroshi Motoda. *Computational methods of feature selection*. CRC Press, 2007.

[Liu and Setiono, 1995] Huan Liu and Rudy Setiono. Chi2: Feature selection and discretization of numeric attributes. In *TWAI*, pages 388–391, 1995.

[Liu *et al.*, 2013] Xinwang Liu, Lei Wang, Jian Zhang, Jianping Yin, and Huan Liu. Global and local structure preservation for feature selection. *TNNLS*, 25(6):1083–1095, 2013.

[Mangasarian, 2006] Olvi L Mangasarian. Exact ℓ_1-norm support vector machines via unconstrained convex differentiable minimization. *JMLR*, 7:1517–1530, 2006.

[Montgomery *et al.*, 2009] Douglas C Montgomery, George C Runger, and Norma F Hubele. *Engineering statistics*. John Wiley & Sons, 2009.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris H. Q. Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *NIPS*, pages 1813–1821, 2010.

[Obozinski *et al.*, 2006] Guillaume Obozinski, Ben Taskar, and Michael Jordan. Multi-task feature selection. *SD*, 2, 2006.

[Padungweang *et al.*, 2012] P Padungweang, C Lursinsap, and K Sunat. A discrimination analysis for unsupervised feature selection via optic diffraction principle. *TNNLS*, 23(10):1587–1600, 2012.

[Peng *et al.*, 2005] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *TPAMI*, 27(8):1226–1238, 2005.

[Wang *et al.*, 2007] Li Wang, Ji Zhu, and Hui Zou. Hybrid huberized support vector machines for microarray classification. In *ICML*, pages 983–990, 2007.

[Yang *et al.*, 2011] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou. $l_{2,1}$-norm regularized discriminative feature selection for unsupervised learning. In *IJCAI*, pages 1589–1594, 2011.

[Zhao *et al.*, 2010] Zheng Zhao, Fred Morstatter, Shashvata Sharma, Salem Alelyani, Aneeth Anand, and Huan Liu. Advancing feature selection research. *ASU*, pages 1–28, 2010.

[Zhu *et al.*, 2014] Xiaofeng Zhu, Heung-Il Suk, and Dinggang Shen. Matrix-similarity based loss function and feature selection for alzheimer's disease diagnosis. In *CVPR*, pages 3089–3096, 2014.