Compact Margin Machine*

Bo Dai[1] and Gang Niu[2]

[1] NLPR/LIAMA, Institute of Automation, Chinese Academy of Science,
Beijing 100190, P.R. China
bdai@nlpr.ia.ac.cn

[2] State Key Laboratory for Novel Software Technology, Nanjing University,
Nanjing 210093, P.R. China
niugang@ai.nju.edu.cn

Abstract. How to utilize data more sufficiently is a crucial consideration in machine learning. Semi-supervised learning uses both unlabeled data and labeled data for this reason. However, Semi-Supervised Support Vector Machine (S3VM) focuses on maximizing margin only, and it abandons the instances which are not support vectors. This fact motivates us to modify maximum margin criterion to incorporate the global information contained in both support vectors and common instances. In this paper, we propose a new method, whose special variant is a semi-supervised extension of Relative Margin Machine, to utilize data more sufficiently based on S3VM and LDA. We employ *Concave-Convex Procedure* to solve the optimization that makes it practical for large-scale datasets, and then give an error bound to guarantee the classifier's performance theoretically. The experimental results on several real-world datasets demonstrate the effectiveness of our method.

1 Introduction

Semi-supervised learning paradigm, which blossoms out to utilize data sufficiently, has attracted more and more attention in machine learning community [1]. Semi-Supervised Support Vector Machine(S3VM) [2, 3], the semi supervised extension of support vector machine, is a state-of-the-art semi-supervised learning algorithm. It uses all the data no matter labeled or not to detect the margin and achieves significant improvement in practice. However, S3VM abandons the instances which are not support vectors. The framework [4] utilizes general unlabeled data, but still wasteful for non-support vectors, which stops them from further improving the performance.

To utilize the data more sufficiently and efficiently, we notice that compactness of projected data provides global information and thus is important for classification. Linear Discriminant Analysis [5] is a successful algorithm to utilize the information. Algorithm in [6] whitens the data when seeking decision boundary. Gaussian Margin Machine [7] controls the projected data compactness under a distribution assumption. Universum SVM [8] constrains range by

* This work is supported in part by Natural Science Foundation of China (No. 60275025).

M.J. Zaki et al. (Eds.): PAKDD 2010, Part II, LNAI 6119, pp. 507–514, 2010.

data that is related but not belonged to any category. Another criterion that compresses the range of projected data is Relative Margin Machine[9] which is motivated from an affine invariance perspective and some probabilistic properties. Although these algorithms balance global and local information, these methods merely exploit labeled data but turn blind eyes to unlabeled data.

In this paper, we propose a method to use the information contained by instances sufficiently. After brief introductions to S3VM and LDA in Section 2, our framework incorporating LDA criterion with S3VM is presented in Section 3 and a relaxation of the criterion is given to make the optimization tractable. We also derive a special variant which can be viewed as the semi-supervised extension of Relative Margin Machine. The generalization bound is analyzed in Section 4. Finally, the experimental results are reported in Section 5.

2 Background

Suppose that we are given labeled data set $\mathcal{L} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_l, y_l)\}$ and unlabeled data set $\mathcal{U} = \{\mathbf{x}_{l+1}, \mathbf{x}_{l+2}, \ldots, \mathbf{x}_{l+u}\}$, both of which are drawn *i.i.d* from a certain data distribution $\mathcal{D}$, where $u \gg l$, $\mathbf{x}_i \in \mathbb{R}^d$ $(i = 1, 2, \ldots, l+u)$ and $y_j \in \{-1, 1\}, (j = 1, 2, \ldots, l)$ is the label of instance $\mathbf{x}_j$. The problem we want to solve is seeking a hypothesis $h : \mathbb{R}^d \rightarrow \{-1, +1\}$ which can classify the unlabeled data and unseen instances sampled from $\mathcal{D}$.

2.1 Semi-supervised SVM(S3VM)

Many semi-supervised methods find the suitable hypothesis by minimizing the criterion which utilize both labeled data and unlabeled data as

$$\min_f \|f\|_{\mathcal{H}_K} + C_1 \sum\nolimits_{i=1}^{l} \ell_1(\mathbf{x}_i, y_i; f) + C_2 \sum\nolimits_{i=l+1}^{l+u} \ell_2(\mathbf{x}_i; f) \tag{1}$$

where $\mathcal{H}$ is the reproducing kernel Hilbert space introduced by kernel function $\mathbb{K}$, ℓ_1 is a common classification loss function, and ℓ_2 is another loss function which utilizes unlabeled data only. S3VM employs *hinge loss* as ℓ_1 and *symmetric hinge loss* as ℓ_2. C_1 and C_2 are the parameters to balance the loss between labeled data, unlabeled data and function complexity. We take the set of linear form of $h = sign(f(\mathbf{x}))$, where $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b(\mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R})$, as the hypothesis space. So the formulation of (1) could be transformed as follow:

$$\begin{aligned} &\min_{\mathbf{w},b,\eta,\xi \geq 0} \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum\nolimits_{i=1}^{l} \eta_i + C_2 \sum\nolimits_{i=l}^{l+u} \xi_i \\ &\text{s.t.} \quad y_i f(\mathbf{x}_i) \geq 1 - \eta_i, i = 1, \ldots, l, \quad |f(\mathbf{x}_i)| \geq 1 - \xi_i, i = l+1, \ldots, l+u. \end{aligned} \tag{2}$$

2.2 Linear Discriminative Analysis(LDA)

The basic principle of LDA is projecting the data into a subspace in which the instances in different categories can be scattered and the instances in the same category can be concentrated together. Let $\mathbf{m}_i = \frac{1}{n_i}\sum_{j\in\mathcal{C}_i}\mathbf{x}_j, i \in \{-1,+1\}$ is the sample mean of class $\mathcal{C}_i$ where n_i is the number of samples in $\mathcal{C}_i$, the mean of projected instances is given by $\hat{\mathbf{m}}_i = \frac{1}{n_i}\sum_{j\in\mathcal{C}_i} f(\mathbf{x}_j) = \mathbf{w}^T\mathbf{m}_i + b, i \in \{-1,+1\}$. So the distance between the projected means, *between-class distance*, is $\|\hat{\mathbf{m}}_{-1} - \hat{\mathbf{m}}_1\| = \|\mathbf{w}^T\mathbf{m}_{-1} - \mathbf{w}^T\mathbf{m}_1\|$. LDA seeks the largest *between-class distance* relative to total *within-class variance* which is defined by $\sum_{i=-1,+1}\mathbf{s}_i$ where $\mathbf{s}_i = \sum_{j\in\mathcal{C}_i}(f(\mathbf{x}_j) - \hat{\mathbf{m}}_i)^2$. The LDA criterion is defined as the ratio of the *between-class distance* to total *within-class variance*, $\mathcal{J}_{\mathbf{w}} = \frac{\|\hat{\mathbf{m}}_{-1}-\hat{\mathbf{m}}_1\|}{\sum_{i=-1,+1}\mathbf{s}_i}$.

3 Compact Margin Machine

The data compactness after projecting is very important in reflecting the data structure and can indeed help in classifying the data. It is necessary to restrict the range of projected data while seeking the largest margin. Within-class variance defined in LDA provides us a natural way to measure the projected data compactness. We modify $\mathbf{s}_i$ as $\sum_{j\in\mathcal{C}_i}\max\{0, |f(\mathbf{x}_j) - \hat{\mathbf{m}}_i| - \varepsilon\}, i = \{-1,+1\}$ instead of $2-norm$ form and introduce it as $\sum_{i=1}^{l+u}\ell_c(\mathbf{x}_i;\mathcal{L},\mathcal{U},f) = \sum_{i=-1,+1}\mathbf{s}_i$ to S3VM. We propose the model, *Compact Margin Machine*, as follow ($f(\mathbf{x})$ stands for $\mathbf{w}^T\mathbf{x} + b$):

$$\min_{\mathbf{w},b,\{\eta,\xi,\zeta\}\geq 0, d_i=\{0,1\}} \frac{1}{2}\|\mathbf{w}\|^2 + C_1\sum_{i=1}^{l}\eta_i + C_2\sum_{i=l+1}^{l+u}\xi_i + C_3\sum_{i=1}^{l+u}(\zeta_{1,i} + \zeta_{0,i}) \tag{3}$$
$$\text{s.t. } y_i f(\mathbf{x}_i) \geq 1 - \eta_i, i = 1,\ldots,l \quad |f(\mathbf{x}_i)| \geq 1 - \xi_i, i = l+1,\ldots,l+u$$

$$\frac{y_i+1}{2}|f(\mathbf{x}_i) - \frac{\sum_{i=l+1}^{l+u}(1-d_i)f(\mathbf{x}_i) + \sum_{i=1}^{l}(\frac{y_i+1}{2})f(\mathbf{x}_i)}{2u(1-r) + \sum_{i=1}^{l}(\frac{y_i+1}{2})}| \leq \varepsilon + \zeta_{0,i} \tag{4}$$

$$\frac{1-y_i}{2}|f(\mathbf{x}_i) - \frac{\sum_{i=l+1}^{l+u}d_i f(\mathbf{x}_i) + \sum_{i=1}^{l}(\frac{1-y_i}{2})f(\mathbf{x}_i)}{u(2r-1) + \sum_{i=1}^{l}(\frac{1-y_i}{2})}| \leq \varepsilon + \zeta_{1,i} \tag{5}$$

$$(1-d_i)|f(\mathbf{x}_i) - \frac{\sum_{i=l+1}^{l+u}(1-d_i)f(\mathbf{x}_i) + \sum_{i=1}^{l}(\frac{y_i+1}{2})f(\mathbf{x}_i)}{2u(1-r) + \sum_{i=1}^{l}(\frac{y_i+1}{2})}| \leq \varepsilon + \zeta_{0,i} \tag{6}$$

$$d_i|f(\mathbf{x}_i) - \frac{\sum_{i=l+1}^{l+u}d_i f(\mathbf{x}_i) + \sum_{i=1}^{l}(\frac{1-y_i}{2})f(\mathbf{x}_i)}{u(2r-1) + \sum_{i=1}^{l}(\frac{1-y_i}{2})}| \leq \varepsilon + \zeta_{1,i} \tag{7}$$

where we introduce the class balancing constraint $\frac{1}{u}\sum_{i=l+1}^{l+u}d_i = 2r - 1$ to avoid the trivial solution that assigns all the instances the same label [10]. We can use branch-and-bound algorithms to search the global optimal solution. However, the computational complexity is too high. We relax the constraints to make the optimization process easier and it can be proved that our relaxed constraints imply upper bounds of original loss functions.

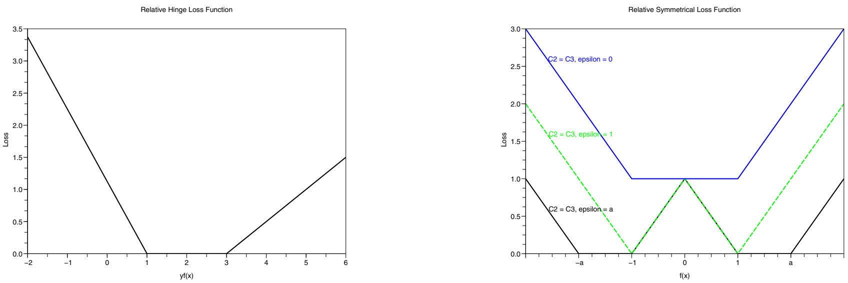

Fig. 1. Relaxed compact ε-hinge loss and relaxed compact ε-symmetric hinge loss

Proposition 1. *By replacing the constraints (4)-(7) with* $|\mathbf{w}^T\mathbf{x}_i+b| \le \frac{1}{2}(\varepsilon+\zeta_i)$, *the loss will be no less than* ℓ_c *defined above.*

Proof. As the constraints (4)-(7) have the same form, without loss of generality, we consider the constraint (7) and the others are similar. Obviously, if there is a solution $\mathbf{w}$ that satisfies $|\mathbf{w}^T\mathbf{x}_i + b| \le \frac{1}{2}(\varepsilon + \zeta_i)$, then we have

$$d_i\Big|\Big((\mathbf{w}^T\mathbf{x}_i + b) - \frac{\sum_{j=l+1}^{l+u} d_j(\mathbf{w}^T\mathbf{x}_j + b) + \sum_{j=1}^{l}(\frac{1-y_j}{2})(\mathbf{w}^T\mathbf{x}_j + b)}{\sum_{j=l}^{l+u} d_j + \sum_{j=1}^{l}(\frac{1-y_j}{2})}\Big)\Big| \le$$

$$d_i\Big(|\mathbf{w}^T\mathbf{x}_i + b| + \Big|\frac{\sum_{j=l+1}^{l+u} d_j(\mathbf{w}^T\mathbf{x}_j + b) + \sum_{j=1}^{l}(\frac{1-y_j}{2})(\mathbf{w}^T\mathbf{x}_j + b)}{\sum_{j=l}^{l+u} d_j + \sum_{j=1}^{l}(\frac{1-y_j}{2})}\Big|\Big) \le$$

$$|\mathbf{w}^T\mathbf{x}_i + b| + \frac{1}{n_j}\sum_{j\in\mathcal{C}_{y_i}}|\mathbf{w}^T\mathbf{x}_j + b| \le \varepsilon + \frac{1}{2}\zeta_i + \frac{1}{2n_j}\sum_{j\in\mathcal{C}_{y_i}}\zeta_j$$

The inequality is derived from triangle inequality. We obtain that the relaxed constraints provide an upper bound and $\sum_{i=1}^{n}\ell_c(\mathbf{x}_i;\mathcal{L},\mathcal{U},f) \le \sum_{i=1}^{n}\zeta_i$. □

Proposition 1 suggests that the relaxed constrains are helpful in reflecting the projected data compactness. Mathematically, CMM is relaxed as:

$$\begin{aligned}\min_{\mathbf{w},b,\{\eta,\xi,\zeta\}\ge 0}\ &\frac{1}{2}\|\mathbf{w}\|^2 + C_1\sum_{i=1}^{l}\eta_i + C_2\sum_{i=l+1}^{l+u}\xi_i + C_3\sum_{i=1}^{l+u}\zeta_i\\ \text{s.t.}\ \ &y_i(\mathbf{w}^T\mathbf{x}_i + b) \ge 1 - \eta_i, \quad i = 1,\ldots,l\\ &|\mathbf{w}^T\mathbf{x}_i + b| \ge 1 - \xi_i, \quad i = l+1,\ldots,l+u\\ &|\mathbf{w}^T\mathbf{x}_i + b| \le \varepsilon + \zeta_i, \quad i = 1,\ldots,l+u\end{aligned} \tag{8}$$

Note that relaxed constraints modify both of the loss functions for labeled and unlabeled data given by S3VM actually. The loss function for the labeled data is $\ell_1^{'} = \max\{0, 1 - yf(x)\} + \frac{C_3}{C_1}\max\{0, |f(x)| - \varepsilon\}$ which we named as *relaxed compact ε-hinge loss*. On the other hand, the loss function of the unlabeled data is adapted as $\ell_2^{'} = \max\{0, 1 - |f(x)|\} + \frac{C_3}{C_2}\max\{0, |f(x)| - \varepsilon\}$, named as *relaxed compact ε-symmetric hinge loss*. The loss functions are shown above. It is not easy to solve the optimization problem (8) directly because of the non-convex property of *relaxed compact ε-symmetric hinge loss*. As [2], Mix Integer Programming can find the global optimal solution of (8). However, the computational complexity of MIP is usually very high. By employing *Concave-Convex Procedure(CCCP)* [11], (8) can be solved effciently [12]. We set $\mathbf{x}_i = \mathbf{x}_{i-u}, i = l+u+1, \ldots, l+2u,\ y_i = 1, i = l+1, \ldots, l+u,\ y_i = -1, i = l+u+1, \ldots, l+2u$ and rewrite the relaxed Compact Margin Machine criterion (8) as the sum of a convex part

$$\mathcal{J}_{vex} = \tfrac{1}{2}\|\mathbf{w}\|^2 + C_1 \textstyle\sum_{i=1}^{l} \max\{0, 1 - y_i f(\mathbf{x}_i)\} + C_2 \sum_{i=l+1}^{l+2u} \max\{0, 1 - y_i f(\mathbf{x}_i)\} \\ + C_3 \textstyle\sum_{i=1}^{l+2u} \max\{0, |f(\mathbf{x}_i)| - \varepsilon\}$$

and a concave part $\mathcal{J}_{cav} = -C_2 \sum_{i=l+1}^{l+2u} \max\{0, \delta - y_i f(\mathbf{x}_i)\}$ where $\delta \in (-1, 0]$.

```
Initialize $\boldsymbol{\theta}^0 = (\mathbf{w}^0, b^0)$ with a standard SVM solution on labeled data
Set
    $\beta_i = \begin{cases} C_2 \text{ if } y_i f_{\theta^0}(\mathbf{x}_i) < \delta \quad \textit{and} \quad i \geq l+1 \\ 0 \quad \text{otherwise} \end{cases}$
while $\boldsymbol{\beta}^{t+1} \neq \boldsymbol{\beta}^t$ do
    solve the convex problem where $\mathbb{K}$ is kernel
    $\min_{\boldsymbol{\alpha},\gamma,\hat{\gamma}\geq 0} \frac{1}{2}\big((\boldsymbol{\alpha}-\boldsymbol{\beta}) \odot \mathbf{y} - \gamma + \hat{\gamma}\big)^T \mathbb{K}\big((\boldsymbol{\alpha}-\boldsymbol{\beta}) \odot \mathbf{y} - \gamma + \hat{\gamma}\big) - \boldsymbol{\alpha}^T\mathbf{1} + \varepsilon\gamma^T\mathbf{1} + \varepsilon\hat{\gamma}^T\mathbf{1}$
    subject to $(\boldsymbol{\beta} - \alpha)^T\mathbf{y} + \gamma^T\mathbf{1} - \hat{\gamma}^T\mathbf{1} = 0, \hat{\gamma} + \gamma \leq C_3\mathbf{1}$
        $\alpha_i \leq C_1, i = 1, \ldots, l, \alpha_i \leq C_2, i = 1+1, \ldots, l+2u$
    compute $b^{t+1}$ by $f_{\boldsymbol{\theta}^{t+1}}(\mathbf{x}_i) = \big((\boldsymbol{\alpha} - \boldsymbol{\beta}) \odot \mathbf{y} - \gamma + \hat{\gamma}\big)^T \mathbb{K}_{i\bullet} + b^{t+1}$
    compute $y_i f_{\boldsymbol{\theta}^{t+1}}(\mathbf{x}_i), i = l+1, \ldots, l+2u$ by $KKT$ conditions
    alternate $\boldsymbol{\beta}$
        $\beta_i = \begin{cases} C_2 \text{ if } y_i f_{\theta^{t+1}}(\mathbf{x}_i) < \delta \quad \textit{and} \quad i \geq l+1 \\ 0 \quad \text{otherwise} \end{cases}$
end
```

Algorithm 1. Concave-Convex Procedure for Relaxed CMM

At each iteration, *CCCP* solves $\min_{\mathbf{w},b} \mathcal{J}_{vex}(\mathbf{w}, b) + \mathcal{J}_{cav}'(\mathbf{w}^t, b^t) \cdot (\mathbf{w}, b)$ until convergence. The convergence of *CCCP* has been shown by [13]. The steps of algorithm are shown in Algorithm 1.

3.1 Special Variant and the Relationship to RMM

In this section, we derive a special variant from (8) which can be solved more efficiently. We set the parameters $C_2 = C_3$, $\varepsilon = 0$ in (8), so the loss function of unlabeled data is the blue one in Fig.1. The mathematical form is as following:

$$\begin{aligned} \min_{\mathbf{w},b,\{\eta,\zeta\}\geq 0} & \frac{1}{2}\|\mathbf{w}\|_2^2 + C_1\sum_{i=1}^{l}\eta_i + C_2\sum_{i=1}^{l+u}(\zeta_i+\hat{\zeta}_i) \\ \text{s.t. } & y_i(\mathbf{w}^T\mathbf{x}_i+b)\geq 1-\eta_i, \quad i=1,\ldots,l \\ & \mathbf{w}^T\mathbf{x}_i+b\leq\varepsilon+\zeta_i, -(\mathbf{w}^T\mathbf{x}_i+b)\leq\varepsilon+\hat{\zeta}_i, \quad i=1,\ldots,l+u \end{aligned} \tag{9}$$

This special variant of relaxed Compact Margin Machine utilizes the labeled data to control the margin and both the labeled and unlabeled data to compress the range of projected data. It can be viewed as a semi-supervised extension of RMM. RMM introduces the relaxed constraints (8) from the other motivations such as affine invariance perspective and some probabilistic properties. From this aspect, we can conclude that our model achieves the same properties that are given by RMM.

4 Theoretical Analysis

In this section, we derive the empirical transductive Rademacher complexity [14] for function classes relaxed Compact Margin Machine which can be plugged into uniform Rademacher error bound directly.

Firstly, we define the function class of S3VM as $\mathscr{H}_{\mathcal{D}} = \{\mathbf{w}^T\mathbf{x} \mid \frac{1}{2}\mathbf{w}^T\mathbf{w} \leq \mathcal{D}\}$ and the function class of *relaxed Compact Margin Machine* as $\mathscr{F}_{\mathcal{D},\mathcal{C}_3} = \{\mathbf{w}^T\mathbf{x} \mid \frac{1}{2}\mathbf{w}^T\mathbf{w}+\frac{\mathcal{C}_3}{2}\|\mathbf{w}^T\mathbf{z}_i\|_2^2 \leq \mathcal{D}, 1\leq i\leq n\}$ where $\mathscr{Z} = \{\mathbf{z}_1,\ldots,\mathbf{z}_n\}$ is an extra dataset drawn from the same distribution for convenience of proof. However, in practice, we may use training dataset and testing dataset as the extra dataset. We can derive the empirical transductive Rademacher complexity of relaxed CMM and S3VM.

Lemma 2. *The transductive Rademacher complexity of* Semi-Supervised SVM $R_{l+u}(\mathscr{H}_{\mathcal{D}}) \leq 2\sqrt{\frac{D}{lu}\sum_{i=1}^{r}\lambda_i}$*, where* $\{\lambda_i\}_{i=1}^{r}$ *are the singular eigenvalues of Gram matrix of the data.* □

Lemma 3. *The transductive Rademacher complexity of* relaxed Compact Margin Machine $R_{l+u}(\mathscr{F}_{\mathcal{D},\mathcal{C}_3}) \leq \min_{\alpha\geq 0}\frac{1}{l+u}\sum_{i=1}^{l+u}\mathbf{x}_i\mathbb{K}_{\alpha,\mathcal{C}_3}\mathbf{x}_i + \frac{lu}{(m+u)^2}D\sum_{i=1}^{n}\alpha_i$*, where* $\mathbb{K}_{\alpha,C_3} = \sum_{i=1}^{n}\alpha_i\mathbf{I} + \mathcal{C}_3\sum_{i}^{n}\alpha_i\mathbf{z}_i\mathbf{z}_i^T$. □

Based on theorems in [14], the following corollary provides an upper bound on the error rate:

Corollary 4. *Fix* $\gamma > 0$*, let* $\mathscr{F}$ *be the set of function. Let* $c_0 = \sqrt{\frac{32\ln 4e}{3}}$ *and* $Q = \frac{l+u}{lu}$*. Then with probability at least* $1-\delta$ *over the training set, the following bound holds:*

$$P[y \neq sign(f(x))] \leq \frac{1}{n\gamma}\sum_{i=1}^{l}\xi_i + \frac{R_{l+u}(\mathscr{F})}{\gamma} + c_0 Q\sqrt{\min\{l,u\}} + \sqrt{2Q\ln\frac{1}{\delta}}$$

where $\xi_i = \max\{0, 1-y_i f(x_i)\}$*.*

Table 1. Experimental Results (Accuracy in percentage)

Performance Comparison on MNIST (mean%)								
num	#label/class	5	10	20	40	50	100	200
400	SVM	71.19	89.76	92.36	94.12	94.41	95.32	95.50
	LDA	84.31	89.34	92.13	93.16	92.85	93.23	93.84
	RMM	84.62	89.90	92.58	94.43	94.81	95.95	**96.13**
	TSVM	87.18	93.54	**95.11**	95.15	95.20	95.23	95.50
	LapSVM	**88.83**	91.68	93.22	94.52	94.63	95.11	95.50
	CMM	88.29	**93.80**	95.07	**95.60**	**95.76**	**96.10**	**96.13**
600	TSVM	90.32	93.59	94.76	95.20	95.26	95.72	96.22
	LapSVM	**91.23**	92.31	93.50	94.33	94.69	95.39	96.47
	CMM	90.73	**94.62**	**95.27**	**95.83**	**96.01**	**96.33**	**96.72**
800	TSVM	**94.52**	93.86	94.65	95.41	95.45	96.12	96.42
	LapSVM	92.13	93.44	94.13	95.00	95.20	95.87	96.67
	CMM	91.37	**94.99**	**95.70**	**96.10**	**96.31**	**96.63**	**96.89**

Performance Comparison on TDT2 (mean%)					
num	#label/class	5	15	50	150
300	SVM	87.09	93.62	96.38	98.00
	LDA	93.44	96.33	94.45	94.74
	RMM	94.39	96.84	97.60	**98.55**
	TSVM	**96.48**	95.95	97.73	98.00
	LapSVM	89.12	95.79	97.52	98.00
	CMM	94.90	**97.72**	**98.52**	**98.55**
600	TSVM	**96.23**	95.63	96.11	97.11
	LapSVM	88.50	93.24	95.89	97.69
	CMM	95.45	**97.20**	**97.38**	**97.78**

5 Experiment

The proposed algorithm is evaluated on two benchmark datasets, MNIST[1] and TDT2[2], to illustrate the effectiveness. We compare our model with SVM, Kernel LDA, RMM, LapSVM and S3VM. We implement our algorithm, RMM and LapSVM based on CVX[3]. The SVM and S3VM are solved by SVM^{light} [4]. Among all the experiments, we select the best kernel by 10-fold cross-validation. The other parameters are also tuned beforehand. We chose the digits "8" vs "9" from MNIST dataset for binary classification problem because these two digits are difficult to discriminate visually. In TDT2 dataset, we chose categories randomly. We conduct experiments on several different amount of labeled and unlabeled data. The final test accuracy is given as the average of 10 independent trials on the test dataset.

[1] The MNIST dataset can be download from `http://yann.lecun.com/exdb/mnist/`

[2] The TDT2 dataset can is available at `http://www.nist.gov/speech/tests/tdt/tdt98`

[3] The CVX matlab code be obtained from `http://www.stanford.edu/~boyd/cvx/`

[4] The package of SVM^{light} can be download from `http://svmlight.joachims.org/`

The training data are selected randomly each time. We further select a certain number of labeled data from the selected training dataset randomly.

The weakness of large margin criterion can be observed from the results. When labeled data are rare, the algorithms that compress the range of projected data, such as LDA and RMM, perform better than SVM. Our semi-supervised algorithm achieves significant improvement on the two datasets with benefit from the compact margin. Fortunately, our algorithm suffers the slightest depression compared with others when the unlabeled data hurt the classifiers.

6 Conclusion

We propose a novel semi-supervised algorithm which can utilize the data sufficiently. To make our method capable to handle large-scale applications, we employ Concave-Convex Procedure to solve the non-convex problem. A semi-supervised extension of RMM can be derived from our model. Moreover, we provide theoretical analyses to guarantee the performance of our algorithm, and finally experimental results show that the proposed algorithm improves the accuracy. A efficient global optimization method for exact solution is our future direction.

References

[1] Chapelle, O., Schölkopf, B., Zien, A. (eds.): Semi-Supervised Learning. MIT Press, Cambridge (2006)

[2] Bennett, K.P., Demiriz, A.: Semi-supervised support vector machines. In: 12th NIPS, pp. 368–374 (1998)

[3] Joachims, T.: Transductive inference for text classification using support vector machines. In: 16th ICML, pp. 200–209 (1999)

[4] Huang, K., Xu, Z., King, I., Lyu, M.R.: Semi-supervised learning from general unlabeled data. In: Perner, P. (ed.) ICDM 2008. LNCS (LNAI), vol. 5077, pp. 273–282. Springer, Heidelberg (2008)

[5] Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals Eugen. 7, 179–188 (1936)

[6] Xiong, R., Cherkassky, V.: A combined svm and lda approach for classification. In: IJCNN, pp. 157–171 (2005)

[7] Crammer, K., Mohri, M., Pereira, F.: Gaussian margin machines. In: 12th AISTATS (2009)

[8] Weston, J., Collobert, R., Sinz, F., Bottou, L., Vapnik, V.: Inference with the universum. In: 23rd ICML, pp. 1009–1016 (2006)

[9] Shivaswamy, P., Jebara, T.: Relative margin machines. In: 22nd NIPS (2008)

[10] Chapelle, O., Sindhwani, V., Keerthi, S.S.: Optimization techniques for semi-supervised support vector machines. JMLR 9, 203–233 (2008)

[11] Yuille, A.L., Rangarajan, A.: The concave-convex procedure (cccp). In: 15th NIPS. MIT Press, Cambridge (2001)

[12] Collobert, R., Sinz, F., Weston, J., Bottou, L.: Large scale transductive svms. JMLR 7, 1687–1712 (2006)

[13] Sriperumbudur, B., Lanckriet, G.: On the convergence of the concave-convex procedure. In: 23rd NIPS (2009)

[14] El-Yaniv, R., Pechyony, D.: Transductive rademacher complexity and its applications. In: 20th COLT, pp. 157–171 (2007)