ESANN'2006 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2006, d-side publi., ISBN 2-930307-06-4.

Synthesis of Maximum Margin and Multiview Learning using Unlabeled Data

Sandor Szedmak[1] and John Shawe-Taylor[1] *

1 - Electronics and Computer Science, ISIS Group
University of Southampton, SO17 1BJ, United Kingdom

Abstract. In this presentation we show the semi-supervised learning with two input sources can be transformed into a maximum margin problem to be similar to a binary SVM. Our formulation exploits the unlabeled data to reduce the complexity of the class of the learning functions. In order to measure how the complexity is decreased we use the Rademacher Complexity Theory. The corresponding optimization problem is convex and it is efficiently solvable for large-scale applications as well.

1 Introduction

Semi-supervised learning belongs to the main directions of the recent machine learning research. The exploitation of the unlabeled data is an attractive approach either to extend the capability of the known methods or to derive novel learning devices. In this presentation we give a synthesis of some earlier approaches and show an optimization framework with good statistical generalization capability. Our idea consists of:

Multiview learning, when two or more sources of the inputs are given with the same output, see in papers of Blum et al. [4] and Dasgupta et al. [5],

Maximum margin learning, where the unlabeled cases within the margin considered as errors, see in Bennett et al. [3],

Reduction of the learning class complexity, where the "closeness" of the learning functions is assumed on the unlabeled data, see in the conference paper about skeleton learning of Lugosi et al. [6] and Balcan et al. [1],

Combination of the multiview and maximum margin learning, which introduces new constraints to the optimization, see in Meng et al. [7].

Blending these ideas we arrive at an optimization framework, which can be solved efficiently at the level of computational complexity of a binary SVM problem. In the first part the optimization problem is presented and then the generalization performance is analyzed. We apply the Rademacher complexity to give an upper bound on the expected error. At the end, experimental results will illustrate the usefulness of the approach. The underlying theory is summarized in Appendix A and B. Further details can be found in [2].

*This work is supported by PASCAL Network of Excellence (IST-2002-506778) European Community IST Programmes. The authors want to thank Gabor Lugosi for very fruitful discussions.

2 Optimization Framework

In the semi-supervised, multiview learning, with two views, we are given a compound sample S_L of pairs of outputs and inputs $\{(y_i, (\mathbf{x}_i^1, \mathbf{x}_i^2)) : y_i \in \{-1, 1\},\ \mathbf{x}_i^k \in \mathcal{X}_k,\ i = 1, \ldots, m_L,\ k = 1, 2\}$ independently and identically generated by an unknown multivariate distribution $\mathcal{P}(Y, X)$, and a compound sample $S_U = \{(\mathbf{u}_i^1, \mathbf{u}_i^2)) : \mathbf{u}_i^k \in \mathcal{X}_k,\ i = 1, \ldots, m_U,\ k = 1, 2\}$ of unlabeled cases independently and identically chosen from the marginal distribution $\mathcal{P}(X)$ of $\mathcal{P}(Y, X)$. Furthermore, there is given a set of embedding of the inputs into Hilbert spaces by the functions $\phi_k : \mathcal{X}_k \to \mathcal{H}_{\phi_k},\ k = 1, 2$. The image vectors of these mappings are called feature vectors in the sequel.

The objective is to find linear functions $f_k(\mathbf{x}_k) = \mathbf{w}_k^T \boldsymbol{\phi}_{\boldsymbol{k}}(\mathbf{x}_k) + b_k$ which can predict the potential label value for any labeled and unlabeled cases. The decision function is then defined as $\frac{1}{2}\sum_k \text{sign}(f_k)$. In order to exploit the unlabeled data in the optimization problem we choose particular solutions of the SVM subproblems where the predictors give similar solutions on the unlabeled data.

The matrix $\mathbf{Y}$ is a diagonal matrix of the labels $\{y_i\}$, and the matrices $\mathbf{X}_k$ and $\mathbf{U}_k$ comprise the labeled and unlabeled inputs in their rows for $k = 1, 2$. Applying the embedding functions ϕ_k on $\mathbf{X}_k$ and $\mathbf{U}_k$ gives matrices with the feature vectors in their rows, otherwise all vectors are column vectors.

The optimization problem formulating our learning framework is given by

$$\begin{array}{lll} & \min & \frac{1}{2}\sum_k \mathbf{w}_k^T\mathbf{w}_k + \mathbf{1}^T\sum_k C_k\boldsymbol{\xi}_k + C_\eta\mathbf{1}^T(\boldsymbol{\eta}^+ + \boldsymbol{\eta}^-) \\ & \text{w.r.t.} & (\mathbf{w}_k, b_k, \boldsymbol{\xi}_k),\ k = 1, 2,\ (\boldsymbol{\eta}^+, \boldsymbol{\eta}^-), \\ \text{Synthesis:} & \text{s.t.} & \sum_k (-1)^{k-1}(\boldsymbol{\phi}_k(\mathbf{U}_k)\mathbf{w}_k + b_k) = \boldsymbol{\eta}^+ - \boldsymbol{\eta}^-, \\ \text{Subproblems:} & & \mathbf{Y}(\boldsymbol{\phi}_k(\mathbf{X}_k)\mathbf{w}_k + b_k) \geq \mathbf{1} - \boldsymbol{\xi}_k, \\ & & \boldsymbol{\xi}_k \geq \mathbf{0},\ (\boldsymbol{\eta}^+, \boldsymbol{\eta}^-) \geq \mathbf{0},\ k = 1, 2. \end{array} \tag{1}$$

We use the acronym SVM_2K for the problem (1) in the sequel.

Introducing Lagrangian multipliers $\boldsymbol{\alpha}_k$ for any k to the SVM subproblems and $\boldsymbol{\beta}$ to the synthesis constraints then we can express the normal vector of the separating hyperplanes for any k by

$$\mathbf{w}_k = [\boldsymbol{\phi}_k(\mathbf{X}_k)^T, \boldsymbol{\phi}_k(\mathbf{U}_k)^T]\mathbf{g}_k,\ \text{where } \mathbf{g}_k^T = (\boldsymbol{\alpha}_k^T\mathbf{Y}, -(-1)^k\boldsymbol{\beta}^T),$$

then unfolding the Karush-Kuhn-Tucker conditions we can derive the corresponding dual problem. The kernel matrices in the dual have the structure

$$\mathbf{K}_k = \begin{bmatrix} \boldsymbol{\phi}_k(\mathbf{X}_k)\boldsymbol{\phi}_k(\mathbf{X}_k)^T & \boldsymbol{\phi}_k(\mathbf{X}_k)\boldsymbol{\phi}_k(\mathbf{U}_k)^T \\ \boldsymbol{\phi}_k(\mathbf{U}_k)\boldsymbol{\phi}_k(\mathbf{X}_k)^T & \boldsymbol{\phi}_k(\mathbf{U}_k)\boldsymbol{\phi}_k(\mathbf{U}_k)^T \end{bmatrix} = \begin{bmatrix} \mathbf{K}_k^{LL} & \mathbf{K}_k^{LU} \\ \mathbf{K}_k^{UL} & \mathbf{K}_k^{UU} \end{bmatrix} = \begin{bmatrix} \mathbf{K}_k^{L} \\ \mathbf{K}_k^{U} \end{bmatrix}.$$

Thus, the dual reads as

$$\begin{array}{ll} \min_{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}} & \frac{1}{2}\sum_k \mathbf{g}_k^T\mathbf{K}_k\mathbf{g}_k - \sum_k \mathbf{1}^T\boldsymbol{\alpha}_k \\ \text{s.t.} & \mathbf{1}^T\mathbf{g}_k = 0,\ \text{where } \mathbf{g}_k^T = (\boldsymbol{\alpha}_k^T\mathbf{Y} - (-1)^k\boldsymbol{\beta}^T), \\ & 0 \leq \boldsymbol{\alpha}_k \leq C_k,\ -C_\eta \leq \boldsymbol{\beta} \leq C_\eta,\ k = 1, 2. \end{array}$$

3 Theoretical Foundation

To illuminate the theoretical background of the presented learning method we use the theory of the Rademacher Complexity, see in [2]. We assume the notation taken from Section 2.

3.1 Analyzing the Semisupervised Multiview Learning

For SVM_2K, the two feature sets are $\big((\boldsymbol{\phi}_1(\mathbf{X}_1),(\boldsymbol{\phi}_1(\mathbf{U}_1)\big)$ and $\big((\boldsymbol{\phi}_2(\mathbf{X}_2),(\boldsymbol{\phi}_2(\mathbf{U}_2)\big)$.

An application of Theorem 3, see in the Appendix A, shows that

$$\begin{aligned} f_D(\mathbf{u}, f_1, f_2) &\stackrel{\text{def}}{=} \mathbb{E}_u[|\textstyle\sum_k (-1)^{k-1}(\mathbf{w}_k^T\boldsymbol{\phi}(\mathbf{u}_k)+b_k)|] \\ &\le \tfrac{1}{m_U}\mathbf{1}^T(\boldsymbol{\eta}^+ + \boldsymbol{\eta}^-) + \tfrac{2C}{m_U}\sqrt{\textstyle\sum_k tr\big(\mathbf{K}_k\big)} + 3\sqrt{\tfrac{2\ln(2/\delta)}{2m_U}} =: D. \end{aligned}$$

with probability at least $1-\delta$. We have assumed that $\|\mathbf{w}_k\|^2 + b_k^2 \le C^2$ for some prefixed C and any $k = 1, 2$. Hence, the class of functions we are considering when applying SVM_2K to this problem can be restricted to

$$\begin{aligned} \mathcal{F}_{C,D} \quad &= \Big\{ f \Big| f : (\mathbf{x}_k) \to \tfrac{1}{2}\textstyle\sum_{k=1}\big(\mathbf{K}_k(x_k)\mathbf{g}_k + b_k\big), \\ &\quad \mathbf{g}_k^T\mathbf{K}_k\mathbf{g}_k + b_k^2 \le C^2, \;\; k = 1,2, \;\; f_D(\mathbf{u}, f_1, f_2) \le D \Big\}, \\ \text{where} \quad &\mathbf{K}_k(\mathbf{x_k}) = \big[\boldsymbol{\phi}(\mathbf{x_k})^T\boldsymbol{\phi}(\mathbf{X}_k)^T, \boldsymbol{\phi}(\mathbf{x_k})^T\boldsymbol{\phi}(\mathbf{U}_k)^T\big]. \end{aligned}$$

The class $\mathcal{F}_{C,D}$ is clearly closed under negation.

Applying the usual Rademacher techniques for margin bounds on generalization we obtain the following result.

Theorem 1. *Fix $\delta \in (0,1)$ and let $\mathcal{F}_{C,D}$ be the class of functions described above. Let $(\mathbf{X}_k)$ labeled and $(\mathbf{U}_k)$ unlabeled samples be drawn independently according to a probability distribution $\mathcal{P}(X)$ for $k = 1, 2$. Then with probability at least $1-\delta$ over random draws of labeled samples of size m_L, every $f \in \mathcal{F}_{CD}$ satisfies*

$$P_{(x,y)\sim\mathcal{D}}(\operatorname{sign}(f(\mathbf{x})) \neq y) \le \frac{1}{2m_L}\mathbf{1}^T\sum_k \boldsymbol{\xi}_k + \hat{R}_\ell(\mathcal{F}_{C,D}) + 3\sqrt{\frac{\ln(2/\delta)}{2m_L}}.$$

It therefore remains to compute the empirical Rademacher complexity of $\mathcal{F}_{C,D}$, which is the critical discriminator between the bounds for the individual SVMs and that of the SVM_2K. The details are presented in the Appendix B.

4 Experiments

In the experiment we used the image dataset[2] being commonly used for generic object recognition. In the cross-validation 5% percent of the cases are randomly

[2]Available at http://www.robots.ox.ac.uk/∼vgg/data/

chosen as labeled cases and all others were unlabeled. This random selection of the labeled cases was ten times repeated.

Table 1 shows the mean and the standard deviation of the accuracies and the Rademacher complexities computed for each image class by using SVM working on the concatenation of two feature sets and the same ones in case of the SVM_2K. It shows that the smaller Rademacher complexity can increase the mean and decrease the standard deviation of the accuracies in classification.

		Image classes		
		Airplanes	**Faces**	**Motorbikes**
SVM on	mean(std)	91.4(2.8)	96.8(1.3)	94.1(1.3)
two features	Rad. comp.	588.2	339.1	574.3
SVM_2K	mean(std)	**91.9**(2.1)	**97.4**(1.3)	**94.3**(1.1)
	Rad. comp.	236.4	34.4	197.4

Table 1: Accuracies(%), standard deviation and estimation of Rademacher Complexities of the SVM acting on two feature sets and the SVM_2K in three image classes. 5% of the cases were labeled.

A Appendix: Short Introduction of Rademacher Complexity Theory

We begin with the definitions required for Rademacher complexity, see for example Bartlett and Mendelson [2] (see also [8] for an introductory exposition).

Definition 2. *For a sample $S = \{x_1, \cdots, x_\ell\}$ generated by a distribution D on a set X and a real-valued function class $\mathcal{F}$ with a domain X, the empirical Rademacher complexity of $\mathcal{F}$ is the random variable*

$$\hat{R}_\ell(\mathcal{F}) = \mathbb{E}_\sigma\Big[\sup_{f\in\mathcal{F}}\Big|\frac{2}{\ell}\sum_{i=1}^{\ell}\sigma_i f(x_i)\Big|\Big|x_1, \cdots, x_\ell\Big]$$

where $\sigma = \{\sigma_1, \cdots, \sigma_\ell\}$ are independent uniform $\{\pm 1\}$-valued Rademacher random variables. The Rademacher complexity of $\mathcal{F}$ is

$$R_\ell(\mathcal{F}) = \mathbb{E}_S\big[\hat{R}_\ell(\mathcal{F})\big] = \mathbb{E}_{S,\sigma}\Big[\sup_{f\in\mathcal{F}}\Big|\frac{2}{\ell}\sum_{i=1}^{\ell}\sigma_i f(x_i)\Big|\Big]$$

We use $\mathbb{E}_\mathcal{D}$ to denote expectation with respect to a distribution $\mathcal{D}$ and $\mathbb{E}_S$ when the distribution is the uniform (empirical) distribution on a sample S.

Theorem 3. *Fix $\delta \in (0,1)$ and let $\mathcal{F}$ be a class of functions mapping from S to $[0,1]$. Let $(x_i)_{i=1}^{\ell}$ be drawn independently according to a probability distribution*

$\mathcal{D}$. Then with probability at least $1-\delta$ over random draws of samples of size ℓ, every $f \in \mathcal{F}$ satisfies

$$\mathbb{E}_{\mathcal{D}}\left[f\left(x\right)\right] \leq \mathbb{E}_{S}\left[f\left(x\right)\right] + R_{\ell}\left(\mathcal{F}\right) + 3\sqrt{\frac{\ln\left(2/\delta\right)}{2\ell}} \leq \mathbb{E}_{S}\left[f\left(x\right)\right] + \hat{R}_{\ell}\left(\mathcal{F}\right) + 3\sqrt{\frac{\ln\left(2/\delta\right)}{2\ell}}$$

Given a training set S the class of functions that we will primarily be considering are linear functions with bounded norm

$$\left\{x \to \sum_{i=1}^{\ell} \alpha_i \kappa\left(x_i, x\right) : \alpha' K \alpha \leq B^2\right\} \subseteq \left\{x \to \langle w, \phi\left(x\right)\rangle : \|w\| \leq B\right\} = \mathcal{F}_B,$$

where ϕ is the feature mapping corresponding to the kernel κ and K is the corresponding kernel matrix for the sample S. The following result bounds the Rademacher complexity of linear function classes.

Theorem 4. *[2] If $\kappa : X \times X \to R$ is a kernel, and $S = \{x_1, \cdots, x_\ell\}$ is a i.i.d. sample from X, then the empirical Rademacher complexity of the class $\mathcal{F}_B$ satisfies $\hat{R}_{\ell}\left(\mathcal{F}\right) \leq \frac{2B}{\ell}\sqrt{tr\left(K\right)}$.*

B Appendix: Empirical Rademacher Complexity of $\mathcal{F}_{C,D}$

We define an auxiliary function of the weight vectors $\bar{\mathbf{w}}_k = (\mathbf{w}_k, b_k)$, $k = 1, 2$,

$$D(\bar{\mathbf{w}}_1, \bar{\mathbf{w}}_2) \stackrel{def}{=} \mathbb{E}_{\mathcal{D}}\Big[\big|\sum_{k}(-1)^{k-1}(\boldsymbol{\phi}_k(\mathbf{U}_k)\mathbf{w}_k + b_k)\big|\Big],$$

and the Rademacher complexity of the class $\mathcal{F}_{C,D}$

$$\begin{aligned}\hat{R}_{\ell}\left(\mathcal{F}_{C,D}\right) &= \mathbb{E}_{\sigma}\left[\sup_{f\in\mathcal{F}_{C,D}}\left|\tfrac{2}{m_L}\textstyle\sum_{i=1}^{m_L}\sigma_i f\left(\mathbf{x}_i\right)\right|\right] \\ &= \mathbb{E}_{\sigma}\left[\sup_{\substack{\left\|\bar{w}_k\right\|\leq C,\ k=1,2 \\ D(\bar{w}_1,\bar{w}_2)\leq D}}\left|\tfrac{2}{m_L}\textstyle\sum_{k=1}\boldsymbol{\sigma}^T\left(\mathbf{K}_k(\mathbf{x}_k)\mathbf{g}_k + b_k\right)\right|\right].\end{aligned}$$

Based on the reversed version of the basic Rademacher complexity theorem where the roles of the empirical and true expectations are swapped:

Theorem 5. *Fix $\delta \in (0,1)$ and let $\mathcal{F}$ be a class of functions mapping from S to $[0,1]$. Let $(x_i)_{i=1}^{\ell}$ be drawn independently according to a probability distribution $\mathcal{D}$. Then with probability at least $1-\delta$ over random draws of samples of size ℓ, every $f \in \mathcal{F}$ satisfies*

$$\mathbb{E}_{S}\left[f\left(x\right)\right] \leq \mathbb{E}_{\mathcal{D}}\left[f\left(x\right)\right] + R_{\ell}\left(\mathcal{F}\right) + 3\sqrt{\frac{\ln\left(2/\delta\right)}{2\ell}} \leq \mathbb{E}_{\mathcal{D}}\left[f\left(x\right)\right] + \hat{R}_{\ell}\left(\mathcal{F}\right) + 3\sqrt{\frac{\ln\left(2/\delta\right)}{2\ell}}.$$

The proof tracks that of Theorem 3 but is omitted through lack of space.

For weight vectors $\{\bar{\mathbf{w}}_k\}$ satisfying $D(\bar{\mathbf{w}}_1, \bar{\mathbf{w}}_2) \leq D$, an application of Theorem 5 shows that with probability at least $1-\delta$ we have

$$\begin{aligned}\hat{D}(\bar{\mathbf{w}}_1, \bar{\mathbf{w}}_2) &\stackrel{def}{=} \mathbb{E}_S\left[\left|\textstyle\sum_k (-1)^{k-1}(\phi(\mathbf{U}_k)\mathbf{w}_k + b_k)\right|\right] \\ &\leq D + \tfrac{2C}{m_U}\sqrt{\textstyle\sum_k \operatorname{tr}(\mathbf{K}_k)} + 3\sqrt{\tfrac{2\ln(2/\delta)}{2m_U}} \\ &\leq \tfrac{1}{m_U}\mathbf{1}^T(\eta^+ + \eta^-) + \tfrac{2C}{m_U}\sqrt{\textstyle\sum_k \operatorname{tr}(\mathbf{K}_k)} + 3\sqrt{\tfrac{\ln(2/\delta)}{2m_U}} =: \hat{D}.\end{aligned}$$

The above result shows that the Rademacher complexity of $\mathcal{F}_{C,D}$ with probability greater than $1-\delta$ satisfies

$$\hat{R}_\ell\left(\mathcal{F}_{C,D}\right) \leq \mathbb{E}_\sigma\left[\sup_{\substack{\|\bar{w}_k\|\leq C,\ k=1,2,\\ \hat{D}(\bar{w}_1,\bar{w}_2)\leq \hat{D}}} \left|\frac{2}{m_L}\boldsymbol{\sigma}^T \sum_k [\phi_k(\mathbf{X}_k)\mathbf{w}_k + \mathbf{1}b_k]\right|\right],$$

where $\boldsymbol{\sigma} \in \{-1,+1\}^{m_L}$. Note that the expression in square brackets is concentrated under the uniform distribution of Rademacher variables. Hence, we can estimate the complexity for randomly chosen instantiation $\hat{\sigma}$ of the the Rademacher variables σ. We now must find the value of $\{\bar{\mathbf{w}}_k\}$ that maximizes

$$\begin{aligned}\max\quad & \tfrac{1}{m_L}\left|\left[\textstyle\sum_k \boldsymbol{\sigma}^T\phi_k\left(\mathbf{X}_k\right)\mathbf{w}_k + \sum_k \boldsymbol{\sigma}^T\mathbf{1}b_k\right]\right| = \tfrac{1}{m_L}\left|\textstyle\sum_k \boldsymbol{\sigma}^T\left(\mathbf{K}_k^L\mathbf{g}_k + b_k\right)\right| \\ \text{s.t.}\quad & \mathbf{g}_k^T\mathbf{K}_k\mathbf{g}_k + b_k^2 \leq C^2,\ k=1,2, \\ & \tfrac{1}{m_U}\mathbf{1}^T\left|(\textstyle\sum_k (-1)^{k-1}\left(\mathbf{K}_k^U\mathbf{g}_k + b_k\right)\right| \leq \hat{D}.\end{aligned}$$

The expected value of the objective function computed on randomly chosen $\hat{\sigma}$'s is the estimate of the Rademacher complexity.

References

[1] M.F. Balcan and A. Blum. A pac-style model for learning from labeled and unlabeled data. In *COLT*, pages 111–126, 2005.

[2] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[3] K. Bennett and A. Demirez. Semi-supervised support vector machines. In M.S. Kearns, S.A. Solla, and D.A. Cohn, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 368–374. MIT Press, Cambridge, MA, 1998.

[4] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100. 1998.

[5] S. Dasgupta, M.L. Littman, and D. McAllester. Pac generalization bounds for co-training. In *Advances in Neural Information Processing Systems (NIPS)*. 2001.

[6] G. Lugosi and M. Pinter. A data-dependent skeleton estimate for learning. In *In Proceedings of the 9th Annual Conference on Computational Learning Theory, New York,*, pages 51–58. Association for Computing Machinery, 1996.

[7] H. Meng, J. Shawe-Taylor, S. Szedmak, and J.R.D. Farquhar. Support vector machine to synthesise kernels. In *Sheffield Machine Learning Workshop Proceedings, Lecture Notes in Computer Science*. Springer, 2005.

[8] J. Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.