Large Margin Distribution Learning

Zhi-Hua Zhou

National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
zhouzh@lamda.nju.edu.cn

Abstract. Support vector machines (SVMs) and Boosting are possibly the two most popular learning approaches during the past two decades. It is well known that the *margin* is a fundamental issue of SVMs, whereas recently the margin theory for Boosting has been defended, establishing a connection between these two mainstream approaches. The recent theoretical results disclosed that the *margin distribution* rather than a single margin is really crucial for the generalization performance, and suggested to optimize the margin distribution by maximizing the margin mean and minimizing the margin variance simultaneously. Inspired by this recognition, we advocate the *large margin distribution learning*, a promising research direction that has exhibited superiority in algorithm designs to traditional large margin learning.

1 Introduction

Support vector machines (SVMs) and Boosting have both been very popular during the past two decades. SVMs belong to the family of *large margin methods* [18] whereas Boosting belongs to the family of *ensemble methods* [22]. The former roots in the statistical learning theory [19], exploiting the kernel trick explicitly to handle nonlinearity with linear classifiers; the latter comes from the proof construction [13] to the theoretical problem that whether weakly learnable equals strongly learnable [8]. It is clearly that these two approaches were born with apparent differences.

The *margin* [19] is a fundamental issue of SVMs as an intuitive understanding of the behavior of SVMs is to search for a large margin separator in a RKHS (reproducing kernel Hilbert space). It is worth noting that there is also a long history of research trying to explain Boosting with a margin theory. Though there were twists and turns in this line of studies, recently the margin theory for Boosting has finally been defended [5], establishing a connection between these two mainstream learning approaches. It is interesting that in contrast to large margin methods that focus on the maximization of a single margin, the recent theoretical results disclosed that the *margin distribution* rather than a single margin is really crucial for the generalization performance, and suggested to optimize the margin distribution by maximizing the margin mean and minimizing the margin variance simultaneously. Inspired by this recognition, we advocate *large margin distribution learning*, a promising research direction that has already exhibited superiority in algorithm designs [21].

N. El Gayar et al. (Eds.): ANNPR 2014, LNAI 8774, pp. 1–11, 2014.

In this article, we will first briefly introduce the efforts on establishing the margin theory of Boosting, and then explain the basic idea of large margin distribution learning. After that, we will show some simple implementation of large margin distribution learning, followed by concluding remarks.

2 The Long March of Margin Theory for Boosting

Overfitting is among the most serious obstacles for learning approaches to achieve strong generalization performances, and great efforts have been devoted to mechanisms that help reduce overfitting risk, such as decision tree pruning, neural networks early stopping, minimum description length constraint, structural risk minimization, etc. It is typically believed that when the training error reaches zero (even much before that), the training process should be terminated because the further training will unnecessarily increase the model complexity and therefore, leading to overfitting. Indeed, according to the Occam's razor, if we have multiple hypotheses consistent with observations, then the simpler, the better.

However, for `AdaBoost`, the most famous representative of Boosting, it has been observed that the generalization performance can be improved further if the training process continues even after the training error reaches zero, though the ensemble model becomes more complicated owing to the inclusion of more base learners. This seems contradictory to previous knowledge, and thus, to understand why `AdaBoost` seems resistant to overfitting is the most fascinating fundamental theoretical issue in Boosting studies.

To explain this phenomenon, Schapire et al. [14] presented the margin theory for Boosting. Let $\mathcal{X}$ and $\mathcal{Y}$ denote the input and output spaces, respectively. A training set of size m is an *i.i.d.* sample $S = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_m, y_m)\}$ drawn according to D, an unknown underlying probability distribution over $\mathcal{X} \times \mathcal{Y}$. Denote $\Pr_D[\cdot]$ and $\Pr_S[\cdot]$ as the probability w.r.t. D and w.r.t. uniform distribution over S, respectively. Let $\mathcal{H}$ be a hypothesis space, and a base learner is a function $h\colon \mathcal{X} \to \mathcal{Y}$. Here, we focus on binary classification, i.e., $\mathcal{Y} = \{+1, -1\}$. Let $\mathcal{C}(\mathcal{H})$ denote the convex hull of $\mathcal{H}$, i.e., the ensemble model $f \in \mathcal{C}(\mathcal{H})$ is of the form

$$f = \sum\nolimits_i \alpha_i h_i \text{ with } \sum\nolimits_i \alpha_i = 1 \text{ and } \alpha_i \geq 0. \tag{1}$$

We call this ensemble model a voting classifier because the base learners are combined via voting (also called *additive model* in statistical literatures). Given an example $(\boldsymbol{x}, y)$, the *margin* w.r.t. the voting classifier $f = \sum \alpha_i h_i(\boldsymbol{x})$ is defined as $yf(\boldsymbol{x})$; in other words,

$$yf(\boldsymbol{x}) = \sum_{i\colon y = h_i(\boldsymbol{x})} \alpha_i - \sum_{i\colon y \neq h_i(\boldsymbol{x})} \alpha_i, \tag{2}$$

which shows the difference between the weights of base learners that classify $(\boldsymbol{x}, y)$ correctly and the weights of base learners that classify $(\boldsymbol{x}, y)$ incorrectly.

Based on the concept of margin, Schapire et al. [14] proved the first margin theorem for `AdaBoost` and upper bounded the generalization error as follows, where $\theta > 0$ is a threshold of margin over the training sample S.

Theorem 1. *(Schapire et al., 1998) For any $\delta > 0$ and $\theta > 0$, with probability at least $1 - \delta$ over the random choice of sample S with size m, every voting classifier $f \in \mathcal{C}(\mathcal{H})$ satisfies the following bound:*

$$\Pr_D[yf(\boldsymbol{x}) < 0] \leq \Pr_S[yf(\boldsymbol{x}) \leq \theta] + O\left(\frac{1}{\sqrt{m}}\left(\frac{\ln m \ln |\mathcal{H}|}{\theta^2} + \ln\frac{1}{\delta}\right)^{1/2}\right). \quad (3)$$

This theorem implies that, when other variables are fixed, the larger the margin over the training sample, the better the generalization performance; this offers an explanation to why AdaBoost tends to be resistant to overfitting: It is able to increase the margin even after the training error reaches zero.

The margin theory looks intuitive and reasonable, and thus, it attracted a lot of attention. Notice that Schapire et al.'s bound (3) depends heavily on the smallest margin, because $\Pr_S[yf(\boldsymbol{x}) \leq \theta]$ will be small if the smallest margin is large. Thus, Breiman [3] explicitly considered the *minimum margin*, $\hat{y}_1 f(\hat{\boldsymbol{x}}_1) = \min_{i\in\{1..m\}}\{y_i f(\boldsymbol{x}_i)\}$, and proved the following margin theorem:

Theorem 2. *(Breiman, 1999) For any $\delta > 0$, if $\theta = \hat{y}_1 f(\hat{\boldsymbol{x}}_1) > 4\sqrt{\frac{2}{|\mathcal{H}|}}$ and $R \leq 2m$, with probability at least $1 - \delta$ over the random choice of sample S with size m, every voting classifier $f \in \mathcal{C}(\mathcal{H})$ satisfies the following bound:*

$$\Pr_D[yf(\boldsymbol{x}) < 0] \leq R\Big(\ln(2m) + \ln\frac{1}{R} + 1\Big) + \frac{1}{m}\ln\frac{|\mathcal{H}|}{\delta}, \quad (4)$$

where $R = \frac{32\ln 2|\mathcal{H}|}{m\theta^2}$.

Breiman's minimum margin bound (4) is in $O(\ln m/m)$, sharper than Schapire et al.'s bound (3) that is in $O(\sqrt{\ln m/m})$. Thus, it was believed that the minimum margin is essential. Breiman [3] designed the `arc-gv` algorithm, a variant of `AdaBoost`, which directly maximizes the minimum margin. The margin theory would appear to predict that `arc-gv` should perform better than `AdaBoost`; however, empirical results show that though `arc-gv` does produce uniformly larger minimum margin than `AdaBoost`, its generalization error increases drastically in almost every case.[1] Thus, Breiman raised serious doubt about the margin theory, and almost sentenced the margin theory to death.

Seven years later, Reyzin and Schapire [12] found that, amazingly, Breiman had not controlled the model complexity well in experiments. To study the margin, one must fix the model complexity of base learners as it is meaningless to compare the margins of models with different complexities. In his experiments, Breiman [3] used `CART` decision trees, and considering that each decision tree leaf corresponds to an equivalent class in the instance space, Breiman tried to fix the model complexity by using trees with fixed number of leaves. Reyzin and Schapire found that the trees of `arc-gv` are generally deeper than that

[1] Similar empirical evidences have been reported by other researchers such as [7].

of `AdaBoost`, and they argued that trees with different heights may be with different model complexities. Then, they repeated Breiman's experiments using *decision stumps* with two leaves and observed that, comparing to `AdaBoost`, `arc-gv` is with larger minimum margin but smaller margin distribution. Thus, they claimed that the minimum margin is not essential, while the margin distribution characterized by the average or median margin is important.

Though Reyzin and Schapire showed that the empirical attack of Breiman is not deadly, it is far from validating the essentiality of margin distribution, because Breiman's generalization bound based on the minimum margin is quite tight. To enable the margin theory to gets renascence, it is crucial to have a sharper bound based on margin distribution.

For this purpose, Wang et al. [20] presented a sharper bound in term of the *Emargin*, i.e., $\arg\inf_{q\in\{q_0,q_0+\frac{1}{m},\cdots,1\}} KL^{-1}(q;u[\hat{\theta}(q)])$, as follows:

Theorem 3. *(Wang et al., 2008) For any $\delta > 0$, if $8 < |\mathcal{H}| < \infty$, with probability at least $1-\delta$ over the random choice of sample S with size $m > 1$, every voting classifier $f \in \mathcal{C}(\mathcal{H})$ satisfies the following bound:*

$$\Pr_D[yf(\boldsymbol{x}) < 0] \leq \frac{\ln|\mathcal{H}|}{m} + \inf_{q\in\{q_0,q_0+\frac{1}{m},\cdots,1\}} KL^{-1}(q;u[\hat{\theta}(q)]), \tag{5}$$

where $q_0 = \Pr_S\left[yf(\boldsymbol{x}) \leq \sqrt{8/|\mathcal{H}|}\right] < 1$, $u[\hat{\theta}(q)] = \frac{1}{m}\left(\frac{8\ln|\mathcal{H}|}{\hat{\theta}^2(q)}\ln\frac{2m^2}{\ln|\mathcal{H}|} + \ln|\mathcal{H}| + \ln\frac{m}{\delta}\right)$, $\hat{\theta}(q) = \sup\left\{\theta \in \left(\sqrt{8/|\mathcal{H}|},1\right]\colon \Pr_S[yf(\boldsymbol{x}) \leq \theta] \leq q\right\}$.

Here $KL^{-1}(q;u) = \inf_w \{w\colon w \geq q \text{ and } KL(q||w) \geq u\}$ is the inverse of the KL divergence $KL(q||\cdot)$ for a fixed q. Notice that the factors considered by (5) are different from that considered by (3) and (4). Though (5) was believed to be a generalization bound based on margin distribution, the Emargin is too unintuitive to inspire algorithm design.

Several years later, Gao and Zhou [5] revealed that both the minimum margin and Emargin are special cases of the *k-th margin*, which is still a single margin. Fortunately, they proved a sharper generalization bound based on margin distribution as follows by considering the same factors as in (3) and (4).

Theorem 4. *(Gao and Zhou, 2013) For any $\delta > 0$, with probability at least $1-\delta$ over the random choice of sample S with size $m \geq 5$, every voting classifier $f \in \mathcal{C}(\mathcal{H})$ satisfies the following bound:*

$$\Pr_D[yf(\boldsymbol{x}) < 0] \leq \frac{2}{m} + \inf_{\theta\in(0,1]}\left[\Pr_S[yf(\boldsymbol{x}) < \theta] + \frac{7\mu + 3\sqrt{3\mu}}{3m} + \sqrt{\frac{3\mu}{m}\Pr_S[yf(\boldsymbol{x}) < \theta]}\right], \tag{6}$$

where $\mu = \frac{8}{\theta^2}\ln m \ln(2|\mathcal{H}|) + \ln\frac{2|\mathcal{H}|}{\delta}$.

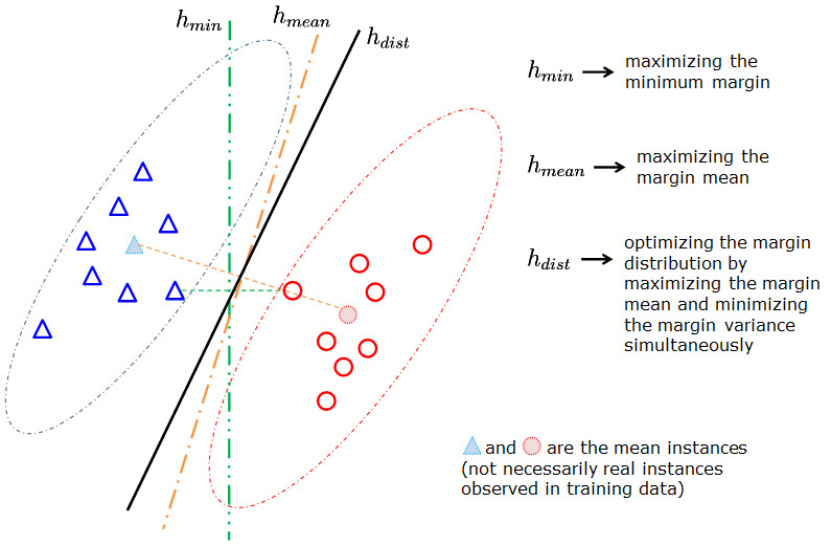

Fig. 1. A simple illustration of linear separators optimizing the minimum margin, margin mean and margin distribution, respectively

This result proves the essentiality of margin distribution to generalization performance. Thus, the margin theory for Boosting finally stands.[2]

Now, it is clear that the margin distribution can be improved further even after the training error reaches zero, and therefore, the generalization performance of `AdaBoost` can be improved further if the training process continues. This also implies that overfitting will finally occur, although very late, since the margin distribution cannot be improved endlessly. As for the contradictory to the Occam's razor, now our understanding is that the complexity of ensemble models is related to not only the number of learners but also the structural relation between the learners; thus, including more base learners in an ensemble does not necessarily lead to a higher model complexity. This is likely to be relevant to the diversity issue of ensemble methods [22], and theoretical exploration of this point may offer model complexity some new comprehension.

3 Optimizing Margin Distribution

Fig. 1 provides a simple illustration. Suppose we are trying to separate two categories of data points, i.e., red circles and blue triangles. For simplicity, consider

[2] Notice that instead of considering the whole function space, there are some studies about data-dependent margin-based generalization bounds, based on techniques such as the empirical cover number [15], empirical fat-shattering dimension [2] and Rademacher and Gaussian complexities [9,10]. Some of these bounds are proven to be sharper than (3), but hard to show sharper than (4)-(6). Moreover, they fail to explain the resistance of `AdaBoost` to overfitting.

the separable case. First, we can see that classifiers maximizing the minimum margin, the margin mean[3] and the margin distribution, respectively, are usually significantly different. For example, in Fig. 1 the classifier trying to maximize the minimum margin will favor the separator h_{min}, the classifier trying to maximize the margin mean will favor the separator h_{mean}, whereas the classifier trying to maximize the margin distribution will favor h_{dist}. Second, the classifier optimizing the margin distribution can be intuitively better as the predictive confidence of h_{dist} on most data points are larger than the predictive confidence of h_{min} and h_{mean}.

Fig. 2 shows a more complicated case where there are outliers or noisy data points. If we insist on optimizing the minimum margin, in Fig. 2 the classifier will almost be dominated by the outliers or noisy data points. If we try to optimize the margin distribution instead, the influence of the outliers or noisy data points will diminish automatically. In other words, classifiers optimizing the margin distribution will be more robust than classifiers optimizing the minimum margin. Moreover, optimizing the margin distribution can also accommodate class imbalance and unequal misclassification costs naturally.

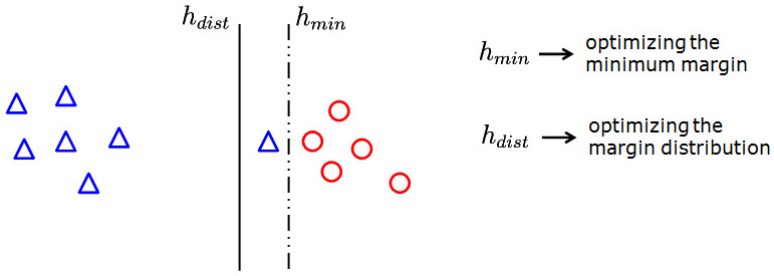

Fig. 2. Another illustration of linear separators with outliers or noisy data points

Notice that though the theoretical results proving the essentiality of margin distribution in Section 3 were derived for Boosting, the implications are far beyond Boosting. There are many learning approaches trying to optimize actually a single margin, particularly the minimum margin; the most famous representatives are SVMs.

For SVMs, $f(\boldsymbol{x}) = \boldsymbol{w}^\top \phi(\boldsymbol{x})$ where $\boldsymbol{w}$ is a linear predictor, $\phi(\boldsymbol{x})$ is a feature mapping of $\boldsymbol{x}$ induced by a kernel k, i.e., $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{x}_j)$. Given an example $(\boldsymbol{x}, y)$, similar to that in Section 2, the margin γ w.r.t. f is defined as $yf(\boldsymbol{x})$ [4, 19]:

$$\gamma = yf(\boldsymbol{x}) = y\boldsymbol{w}^\top \phi(\boldsymbol{x}). \tag{7}$$

[3] Notice that the mean instances are not necessarily observed in training data.

The SVMs formulation for separable case (hard-margin SVMs) is indeed a maximization of the minimum margin, i.e., $\min\{\gamma_i\}_{i=1}^m$:

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} \\ \text{s.t.} \quad & y_i \boldsymbol{w}^\top \phi(\boldsymbol{x}_i) \geq 1 \\ & i = 1, \cdots, m. \end{aligned} \tag{8}$$

The formulation for non-separable case (soft-margin SVMs) introduces the slack variables $\boldsymbol{\xi} = [\xi_1, \cdots, \xi_m]^\top$ to measure the losses of different instances, where C is a trading-off parameter:

$$\begin{aligned} \min_{\boldsymbol{w}, \boldsymbol{\xi}} \quad & \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + C \sum_{i=1}^{m} \xi_i \\ \text{s.t.} \quad & y_i \boldsymbol{w}^\top \phi(\boldsymbol{x}_i) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \ i = 1, \cdots, m. \end{aligned} \tag{9}$$

There exists a constant $\bar{C}$ such that (9) can be equivalently reformulated as follows, showing that the soft-margin SVMs are maximizing the k-th margin (i.e., the k-th smallest margin) [5]:

$$\begin{aligned} \max_{\boldsymbol{w}} \quad & \gamma_0 - \bar{C} \sum\nolimits_{i=1}^{m} \xi_i \\ \text{s.t.} \quad & \gamma_i \geq \gamma_0 - \xi_i, \\ & \xi_i \geq 0, \ i = 1, \cdots, m. \end{aligned} \tag{10}$$

Hence, both hard-margin and soft-margin SVMs are indeed trying to optimize a single margin. It is very likely that they can be improved by replacing the optimization of a single margin by the optimization of margin distribution, while keeping the other parts of their solution strategies unchanged; this also applies to other large margin methods. Thus, the large margin distribution learning offers a promising way to derive more powerful learning approaches by simple adaptations.

To accomplish large margin distribution learning, we need to understand how to optimize the margin distribution. Reyzin and Schapire [12] suggested to maximize the average or median margin, and there are also efforts on maximizing the average margin or weighted combination margin [1, 6, 11]. These arguments, however, are all heuristics without theoretical justification.

In addition to (6), Gao and Zhou [5] proved anther form of their margin theorem, disclosing that the average or median mean is not enough, and to characterize the margin distribution, it is important to consider not only the *margin mean* but also the *margin variance*. This suggests a new direction for algorithm design, i.e., to optimize the margin distribution by maximizing the margin mean and minimizing the margin variance simultaneously. This argument has got supported empirically by some recent Boosting studies [16, 17].

4 A Simple Implementation of Large Margin Distribution Learning

For a straightforward implementation of large margin distribution learning, as an example, we adapt the simple SVMs formulation (8) to the optimization of margin distribution [21].

Denote $\boldsymbol{X} = [\phi(\boldsymbol{x}_1), \cdots, \phi(\boldsymbol{x}_m)]$ as the matrix whose i-th column is $\phi(\boldsymbol{x}_i)$, $\boldsymbol{y} = [y_1, \cdots, y_m]^\top$, and $\boldsymbol{Y}$ as a $m \times m$ diagonal matrix whose diagonal elements are $y_1, \cdots, y_m$. According to the definition in (7), the margin mean is

$$\bar{\gamma} = \frac{1}{m} \sum_{i=1}^{m} y_i \boldsymbol{w}^\top \phi(\boldsymbol{x}_i) = \frac{1}{m} (\boldsymbol{X}\boldsymbol{y})^\top \boldsymbol{w}, \tag{11}$$

and the margin variance is

$$\begin{aligned} \hat{\gamma} &= \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} (y_i \boldsymbol{w}^\top \phi(\boldsymbol{x}_i) - y_j \boldsymbol{w}^\top \phi(\boldsymbol{x}_j))^2 \\ &= \frac{2}{m^2} (m \boldsymbol{w}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{w} - \boldsymbol{w}^\top \boldsymbol{X}\boldsymbol{y}\boldsymbol{y}^\top \boldsymbol{X}^\top \boldsymbol{w}). \end{aligned} \tag{12}$$

By incorporating into (8) the maximization of margin mean and the minimization of margin variance simultaneously, we get the hard-margin LDM (Large Margin distribution Machine) formulation [21]:

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \frac{1}{2} \boldsymbol{w}^\top \boldsymbol{w} + \lambda_1 \hat{\gamma} - \lambda_2 \bar{\gamma} \\ \text{s.t.} \quad & y_i \boldsymbol{w}^\top \phi(\boldsymbol{x}_i) \geq 1 \\ & i = 1, \cdots, m, \end{aligned} \tag{13}$$

where λ_1 and λ_2 are trading-off parameters. It is evident that (8) is a special case of (13) when λ_1 and λ_2 equal zero.

Similarly, we have the soft-margin LDM which degenerates to (10) when λ_1 and λ_2 equals zero:

$$\begin{aligned} \min_{\boldsymbol{w}, \boldsymbol{\xi}} \quad & \frac{1}{2} \boldsymbol{w}^\top \boldsymbol{w} + \lambda_1 \hat{\gamma} - \lambda_2 \bar{\gamma} + C \sum_{i=1}^{m} \xi_i \\ \text{s.t.} \quad & y_i \boldsymbol{w}^\top \phi(\boldsymbol{x}_i) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \ i = 1, \cdots, m. \end{aligned} \tag{14}$$

Notice that in (14) the influence of the $C \sum_{i=1}^{m} \xi_i$ term can be subsumed by the λ_1 and λ_2 terms, whereas we keep it to let (14) and (10) look similar such that it is easy to perceive that adapting the soft-margin SVMs to the optimization of margin distribution is quite straightforward.

Solving (13) and (14) is not difficult. For example, by substituting (11)-(12), (14) leads to a quadratic programming problem:

$$\min_{\boldsymbol{w},\boldsymbol{\xi}} \ \frac{1}{2}\boldsymbol{w}^\top\boldsymbol{w} + \frac{2\lambda_1}{m^2}(m\boldsymbol{w}^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{w} - \boldsymbol{w}^\top\boldsymbol{X}\boldsymbol{y}\boldsymbol{y}^\top\boldsymbol{X}^\top\boldsymbol{w}) - \lambda_2\frac{1}{m}(\boldsymbol{X}\boldsymbol{y})^\top\boldsymbol{w} + C\sum_{i=1}^{m}\xi_i \tag{15}$$

$$\text{s.t.} \ \ y_i\boldsymbol{w}^\top\phi(\boldsymbol{x}_i) \geq 1-\xi_i, \quad \xi_i \geq 0, \ i=1,\cdots,m. \tag{16}$$

A dual coordinate descent method for kernel LDM and an average stochastic gradient descent method for large-scale linear kernel LDM have been developed, with details in [21]. Table 1 shows some experimental results of comparing LDM to SVM, where it can be seen that LDM is significantly better on more than half of the experimental datasets and never worse than SVM. Such a simple implementation of large margin distribution learning also exhibits superior performance to many other related methods [1, 6, 11] in experiments [21].

Table 1. Comparing predictive accuracy (mean±std.) of SVM and LDM. •/◦ indicates the performance of LDM is significantly better/worse than SVM (paired t-tests at 95% significance level). The win/tie/loss counts are summarized in the last row.

Data sets	Linear kernel		RBF kernel	
	SVM	LDM	SVM	LDM
promoters	.723±.071	.721±.069	.684±.100	.715±.074•
planning-relax	.683±.031	.706±.034•	.708±.035	.707±.034
colic	.814±.035	.832±.026•	.822±.033	.841±.018•
parkinsons	.846±.038	.865±.030•	.929±.029	.927±.029
colic.ORIG	.618±.027	.619±.042	.638±.043	.641±.044
sonar	.725±.039	.736±.036	.842±.034	.846±.032
vote	.934±.022	.970±.014•	.946±.016	.968±.013•
house	.942±.015	.968±.011•	.953±.020	.964±.013•
heart	.799±.029	.791±.030	.808±.025	.822±.029•
breast-cancer	.717±.033	.725±.027•	.729±.030	.753±.027•
haberman	.734±.030	.738±.020	.727±.024	.731±.027
vehicle	.959±.012	.959±.013	.992±.007	.993±.006
clean1	.803±.035	.814±.019•	.890±.020	.891±.024
wdbc	.963±.012	.968±.011•	.951±.011	.961±.010•
isolet	.995±.003	.997±.002•	.998±.002	.998±.002
credit-a	.861±.014	.864±.013•	.858±.014	.861±.013
austra	.857±.013	.859±.015	.853±.013	.857±.014•
australian	.844±.019	.866±.014•	.815±.014	.854±.016•
fourclass	.724±.014	.723±.014	.998±.003	.998±.003
german	.711±.030	.738±.016•	.731±.019	.743±.016•
w/t/l (SVM vs. LDM)	0/8/12		0/10/10	

5 Conclusion

Recently the margin theory for Boosting has been defended [5], showing that the *margin* is not only a fundamental issue of SVMs but also an essential factor of Boosting. In contrast to previous belief on single margins such as the minimum margin optimized by SVMs, the recent theoretical results disclosed that the *margin distribution* rather than a single margin is crucial for the generalization performance. Inspired by this recognition, in this article we advocate *large margin distribution learning*. We also briefly introduce how the SVMs can be easily adapted to large margin distribution learning by maximizing the margin mean and minimizing the margin variance simultaneously, while such a simple implementation leads to the LDMs that exhibit superior performance to SVMs [21]. Overall, large margin distribution learning exhibits a promising direction to derive powerful learning approaches.

Acknowledgments. This article summarizes the author's keynote talk at the ANNPR'2014, Montreal, Canada. The author was supported by the National Science Foundation of China (61333014, 61321491).

References

1. Aiolli, F., San Martino, G., Sperduti, A.: A kernel method for the optimization of the margin distribution. In: Proceedings of the 18th International Conference on Artificial Neural Networks, Prague, Czech, pp. 305–314 (2008)
2. Antos, A., Kégl, B., Linder, T., Lugosi, G.: Data-dependent margin-based generalization bounds for classification. Journal of Machine Learning Research 3, 73–98 (2002)
3. Breiman, L.: Prediction games and arcing classifiers. Neural Computation 11(7), 1493–1517 (1999)
4. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)
5. Gao, W., Zhou, Z.-H.: On the doubt about margin explanation of boosting. Artificial Intelligence 199-200, 22–44 (2013) (arXiv:1009.3613, September 2010)
6. Garg, A., Roth, D.: Margin distribution and learning algorithms. In: Proceedings of the 20th International Conference on Machine Learning, Washington, DC, pp. 210–217 (2003)
7. Grove, A.J., Schuurmans, D.: Boosting in the limit: Maximizing the margin of learned ensembles. In: Proceedings of the 15th National Conference on Artificial Intelligence, Menlo Park, CA, pp. 692–699 (1998)
8. Kearns, M., Valiant, L.G.: Cryptographic limitations on learning boolean formulae and finite automata. In: Proceedings of the 21st Annual ACM Symposium on Theory of Computing, Seattle, WA, pp. 433–444 (1989)
9. Koltchinskii, L., Panchanko, D.: Empirical margin distributions and bounding the generalization error of combined classifiers. Annuals of Statistics 30(1), 1–50 (2002)
10. Koltchinskii, L., Panchanko, D.: Complexities of convex combinations and bounding the generalization error in classification. Annuals of Statistics 33(4), 1455–1496 (2005)

11. Pelckmans, K., Suykens, J., Moor, B.D.: A risk minimization principle for a class of parzen estimators. In: Platt, J.C., Koller, D., Singer, Y., Roweis, S. (eds.) Advances in Neural Information Processing Systems 20, pp. 1137–1144. MIT Press, Cambridge (2008)
12. Reyzin, L., Schapire, R.E.: How boosting the margin can also boost classifier complexity. In: Proceeding of 23rd International Conference on Machine Learning, Pittsburgh, PA, pp. 753–760 (2006)
13. Schapire, R.E.: The strength of weak learnability. Machine Learning 5(2), 197–227 (1990)
14. Schapire, R.E., Freund, Y., Bartlett, P.L., Lee, W.S.: Boosting the margin: A new explanation for the effectives of voting methods. Annuals of Statistics 26(5), 1651–1686 (1998)
15. Shawe-Taylor, J., Williamson, R.C.: Generalization performance of classifiers in terms of observed covering numbers. In: Proceedings of the 4th European Conference on Computational Learning Theory, Nordkirchen, Germany, pp. 153–167 (1999)
16. Shen, C., Li, H.: Boosting through optimization of margin distributions. IEEE Transactions on Neural Networks 21(4), 659–666 (2010)
17. Shivaswamy, P.K., Jebara, T.: Variance penalizing AdaBoost. In: Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F.C.N., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 24, pp. 1908–1916. MIT Press, Cambridge (2011)
18. Smola, A.J., Bartlett, P.L., Schölkopf, B., Schuurmans, D. (eds.): Advances in Large Margin Classifiers. MIT Press, Cambridge (2000)
19. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1995)
20. Wang, L., Sugiyama, M., Yang, C., Zhou, Z.-H., Feng, J.: On the margin explanation of boosting algorithm. In: Proceedings of the 21st Annual Conference on Learning Theory, Helsinki, Finland, pp. 479–490 (2008)
21. Zhang, T., Zhou, Z.-H.: Large margin distribution machine. In: Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, New York, NY (2014)
22. Zhou, Z.-H.: Ensemble Methods: Foundations and Algorithms. CRC Press, Boca Raton (2012)