Discussion of Boosting Papers

Peter L. Bartlett
Division of Computer Science and Department of Statistics
University of California, Berkeley
bartlett@stat.berkeley.edu

Michael I. Jordan
Division of Computer Science and Department of Statistics
University of California, Berkeley
jordan@stat.berkeley.edu

Jon D. McAuliffe
Department of Statistics
University of California, Berkeley
jon@stat.berkeley.edu

March 17, 2003

The authors have contributed three significant papers that provide, among other insights, an understanding of the consistency of several "large margin" methods for pattern classification. In two-class classification, the aim is to find a function $f : \mathcal{X} \to \mathbb{R}$ that accurately predicts a binary response variable $Y \in \{\pm 1\}$ using the covariate $X \in \mathcal{X}$, in the sense that $R(f) = \mathbf{E}\ell(Yf(X))$, the risk of the thresholded function, is minimized. Here, $\ell(z)$ denotes the indicator function of the event $z \leq 0$. *Large margin classification methods* use some loss function $\phi : \mathbb{R} \to \mathbb{R}$, typically convex, and seek a function f from some class $\mathcal{F}$ that minimizes the ϕ-risk, $R_\phi(f) = \mathbf{E}\phi(Yf(X))$, that is, the expected loss evaluated at the margin $Yf(X)$. These methods typically minimize the empirical ϕ-risk, $\hat{R}_\phi(f)$, or a regularized version thereof. Many successful pattern classification methods fall in this class, including AdaBoost and other greedy algorithms for forming ensembles of classifiers, and support vector machines. We can categorize them according to the loss function ϕ, the class of functions $\mathcal{F}$, and the algorithm used to approximately minimize R_ϕ.

The three papers in this issue demonstrate the consistency of various methods of this kind.

- The consistency result in the paper by Zhang applies to several loss functions, and concerns kernel methods, which choose a function f from a reproducing kernel Hilbert space $\mathcal{H}$ of functions on $\mathcal{X}$ to minimize a regularized empirical ϕ-risk,

$$\hat{R}_\phi(f) + C\|f\|_{\mathcal{H}},$$

  where $\|\cdot\|_{\mathcal{H}}$ is the Hilbert space norm. This is equivalent (for some λ) to choosing f from the function class

$$\mathcal{F}_k(\mathcal{H}, \lambda) = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq \lambda\}$$

  so as to minimize $\hat{R}_\phi(f)$.

- The paper by Lugosi and Vayatis considers the loss function $\phi(\alpha) = \exp(-\alpha)$, the function class

$$\mathcal{F}_b(\mathcal{G}, \lambda) = \left\{ \sum_i \alpha_i g_i : \|\alpha\|_1 \leq \lambda, g_i \in \mathcal{G} \right\},$$

  where $\mathcal{G} \in \{\pm 1\}^{\mathcal{X}}$ has finite VC-dimension, and an algorithm that minimizes empirical ϕ-risk. The Adaboost algorithm is similar, but without the constraint on the coefficients.

- The paper by Jiang also considers the exponential loss function, the function class

$$\mathcal{F}_b(k) = \left\{ \sum_{i=1}^{k} \alpha_i g_i : g_i \in \mathcal{G} \right\},$$

  and the Adaboost algorithm, which chooses the α_i, g_i sequentially, to greedily minimize empirical ϕ-risk.

We can identify three key steps in proving consistency results of this kind. The first involves a "comparison theorem," relating the excess risk $R(f) - R^*$ to the excess ϕ-risk, $R_\phi(f) - R_\phi^*$. Here,

R^* is the Bayes risk, that is, the infimum over all measurable f of $R(f)$, and $R^*_\phi = \inf_f R_\phi(f)$ is the analogous quantity for the ϕ-risk. A result of this kind is present in all three papers: Zhang's Theorem 2.1 gives an explicit inequality relating the two excess risks; Lugosi and Vayatis's Lemma 5 gives a limiting result; and Jiang's Lemma 1 gives a related comparison, via the $L_2(P)$ distance between f and $f^*_\phi = \arg\min_f R_\phi(f)$.

The second and third steps are more conventional in consistency proofs. The second step is to show that the functions used by the method are rich enough to approximate f^*_ϕ, the measurable function that minimizes ϕ-risk. As formulated above, this involves showing that

$$\bigcup_{\lambda>0} \mathcal{F}_k(\mathcal{H},\lambda), \bigcup_{\lambda>0} \mathcal{F}_b(\mathcal{G},\lambda), \text{ and } \bigcup_{k>0} \mathcal{F}_b(k)$$

are sufficiently rich.

The third step is to choose a sequence of subsets $\mathcal{F}_n \subseteq \mathcal{F}$ with suitably restricted complexity as a function of the sample size n, so that the ϕ-risk of the estimated $\hat{f}_n \in \mathcal{F}_n$ converges to the minimal value, $\inf_{f\in\mathcal{F}_n} R_\phi(f)$. For example, in the cases considered in these three papers, the set $\mathcal{F}_n$ is defined as the set of combinations of k_n functions from $\mathcal{G}$, or the set of combinations of functions from $\mathcal{G}$ with the coefficient vector having one-norm no more than λ_n, or a ball of radius λ_n in an RKHS $\mathcal{H}$. (In the last case, $\hat{f}_n$ is chosen to minimize a combination of the empirical ϕ-risk and a regularization term involving the RKHS norm.)

This third step is a little more involved in the case of Jiang's consistency result, since that result involves an algorithm that does not minimize an objective function involving the empirical risk. Thus, it is essential to show that, under certain conditions, the algorithm finds a good function quickly.

It is interesting to consider what properties of the loss function ϕ allow comparison theorems, and hence consistency results, for large margin methods in general. Jiang's result is for the exponential loss function, and the proof exploits a smoothness assumption on the joint probability distribution that ensures that the optimal f^*_ϕ is continuous. Lugosi and Vayatis assume that ϕ is differentiable, strictly convex, monotonic, and has a certain limiting behavior. Zhang assumes that ϕ satisfies three conditions:

1. For any $\eta \neq 1/2$, any minimizer α^* of the conditional ϕ-risk, $\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)$ has the same sign as $\eta - 1/2$. Thus, a pointwise minimization of the conditional ϕ-risk leads to a function that gives the correct sign everywhere.

2. ϕ is convex.

3. The minimal conditional ϕ-risk,

$$H(\eta) = \inf_{\alpha\in\mathbb{R}} \left(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\right),$$

decreases polynomially with $|1/2 - \eta|$.

Note that the first condition is implied by Lugosi and Vayatis's assumptions (as can be verified by a short calculation), and thus holds a fortiori for the exponential function studied by Jiang. The condition is clearly the weakest possible condition that can be imposed on ϕ if we are to obtain consistency—if the minimizer of ϕ-risk yields the wrong sign at a given point, then it is easy to concoct a probability distribution that has zero excess ϕ-risk but non-zero excess risk. Surprisingly, it turns out that this condition is not only necessary but is also sufficient for obtaining a general comparison theorem—no other conditions are needed. We provide a brief overview of this result here; see Bartlett et al. (2003) for a detailed presentation.

We begin by defining the following functional transform of a loss function ϕ:

Definition 1. Given $\phi : \mathbb{R} \to [0, \infty)$, define the function $\tilde{\psi} : [0, 1] \to [0, \infty)$ by

$$\tilde{\psi}(\theta) = H^{-}\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right),$$

where

$$\begin{aligned} H(\eta) &= \inf_{\alpha\in\mathbb{R}} (\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)) \\ H^{-}(\eta) &= \inf_{\alpha:\alpha(2\eta-1)\leq 0} (\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)). \end{aligned}$$

The ψ-*transform* is defined to be the function $\psi : [0, 1] \to [0, \infty)$ that is the convex closure of $\tilde{\psi}$.

Note that it is straightforward to compute the ψ-transform for all of the examples of loss functions ϕ studied in the three papers in this issue.

The importance of the ψ-transform is shown by the following theorem.

Theorem 2. *For any nonnegative loss function ϕ, any measurable $f : \mathcal{X} \to \mathbb{R}$ and any probability distribution on $\mathcal{X} \times \{\pm 1\}$,*

$$\psi(R(f) - R^*) \leq R_\phi(f) - R^*_\phi.$$

This theorem establishes a general quantitative relationship between the excess ϕ-risk and the excess risk.

For this relationship to be useful in particular applications we need to show that ψ has particular properties—properties that arise from conditions that are imposed on ϕ. In particular, let us introduce the condition described above—that pointwise minimization of the conditional ϕ-risk leads to a function that gives the correct sign. We express this condition in the following way:

Definition 3. We say that ϕ is *classification-calibrated* if, for any $\eta \neq 1/2$,

$$H^{-}(\eta) > H(\eta).$$

Equivalently, ϕ is classification-calibrated if for any sequence (α_i) such that $\lim_{i\to\infty}\{\eta\phi(\alpha_i) + (1-\eta)\phi(-\alpha_i)\} = H(\eta)$, we have $\lim_{i\to\infty} \operatorname{sign}(\alpha_i(\eta-1/2)) = 1$. In particular, if the infimum $H(\eta)$ is achieved at a minimizing value α^*, then this value must have the correct sign. Thus, this condition is essentially an elaboration of Zhang's first condition. As pointed out by Lin (2001), it can be viewed as a variant of Fisher consistency that is appropriate for classification.

We have the following result:

Theorem 4. *The following conditions are equivalent:*

1. *ϕ is classification-calibrated.*
2. *For any sequence (θ_i) in $[0,1]$,*
$$\psi(\theta_i) \to 0 \quad \textit{if and only if} \quad \theta_i \to 0.$$
3. *For every sequence of measurable functions $f_i : \mathcal{X} \to \mathbb{R}$ and every probability distribution P,*
$$R_\phi(f_i) \to R_\phi^* \quad \textit{implies} \quad R(f_i) \to R^*.$$

Thus we see that we obtain a meaningful general comparison theorem under the weakest possible condition on the loss function ϕ. In addition, it can be shown that for a given ϕ, the ψ-transform is optimal in the sense that everywhere on its domain, the bound given by Theorem 2 cannot be improved in general.

Note in particular that we have not assumed that ϕ is convex. If we do assume that ϕ is convex then we can say more—in particular, the function $\tilde{\psi}$ in Definition 1 is then necessarily closed and convex, and thus the ψ-transform is specified directly via the variational representation $\psi(\theta) = H^-((1+\theta)/2) - H((1+\theta)/2)$. Moreover, if ϕ is convex, then it is possible to show that it is classification-calibrated if and only if it is differentiable at 0 and $\phi'(0) < 0$.

The comparison theorem in Theorem 2, and the analogous comparison theorems in the three papers in this issue, suggest a general framework for studying pattern classification methods that involve a surrogate loss function. It is common to view the excess risk as a combination of an estimation term and an approximation term:

$$R(f) - R^* = \left(R(f) - \inf_{g\in\mathcal{F}} R(g) \right) + \left(\inf_{g\in\mathcal{F}} R(g) - R^* \right).$$

However, choosing a function with risk near minimal over a class $\mathcal{F}$—that is, finding an f for which the estimation term above is close to zero—is, in a minimax setting, equivalent to the problem of minimizing empirical risk. For typical classes $\mathcal{F}$ of interest, this problem is computationally infeasible. Even worse, for the function classes typically used by boosting and kernel methods, the estimation term in this expression does not converge to zero for the minimizer of the empirical risk. On the other hand, the comparison theorems we are considering suggest splitting the upper bound on excess risk into an estimation term and an approximation term:

$$\psi(R(f) - R^*) \le R_\phi(f) - R_\phi^* = \left(R_\phi(f) - \inf_{g\in\mathcal{F}} R_\phi(g) \right) + \left(\inf_{g\in\mathcal{F}} R_\phi(g) - R_\phi^* \right). \tag{1}$$

We can view the function ψ provided by the comparison theorem as quantifying the penalty incurred by using the surrogate loss function ϕ in place of the 0-1 loss, and linking the excess risk to the approximation error and estimation error associated with the ϕ-risk.

In many cases it is possible to minimize the ϕ-risk efficiently over a convex class $\mathcal{F}$, and hence find an $f \in \mathcal{F}$ for which this upper bound on risk is near minimal. This holds despite the fact that finding an $f \in \mathcal{F}$ with near-minimal risk is typically computationally infeasible.

Another interesting question raised by Theorem 2 and by the papers in this issue is that of convergence rates. Zhang's paper makes a start in this direction for kernel methods, and this is continued in his more recent work with Mannor and Meir concerning boosting methods (Mannor et al., 2002). Recently, Tsybakov (2001) has considered empirical risk minimization in pattern classification problems with low noise—specifically, where the P_X-probability that $P(Y = 1|X)$ is near $1/2$ is small. He showed that the risk of the empirical minimizer converges to its minimal value surprisingly quickly in these cases. It turns out that, under Tsybakov's low noise condition, the relationship between excess risk and excess ϕ-risk presented in Theorem 2 can be improved (Bartlett et al., 2003). In that case, if the loss function ϕ is uniformly convex and $\mathcal{F}$ is convex, then the excess risk converges to its minimal value (the approximation error term in (1)) surprisingly quickly.

The problem of classification has been a fruitful domain in which to explore connections between statistical and computational science. Efficient algorithms can be designed to solve large-scale classification problems by exploiting tools from convex optimization, and the statistical consequences of using these tools are beginning to be understood. The three papers in this issue represent significant progress on the general problem of incorporating considerations of computational complexity in statistical theory, providing hints of general tradeoffs between statistical accuracy and computational resources that are only beginning to be explored.

References

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. Technical report, Department of Statistics, University of California, Berkeley, 2003.

Y. Lin. A note on margin-based loss functions in classification. Technical Report 1044r, Department of Statistics, University of Wisconsin, 2001.

S. Mannor, R. Meir, and T. Zhang. The consistency of greedy algorithms for classification. In *Proceedings of the Annual Conference on Computational Learning Theory*, pages 319–333, 2002.

A. Tsybakov. Optimal aggregation of classifiers in statistical learning. Technical Report PMA-682, Université Paris VI, 2001.