http://dx.doi.org/10.1090/conm/443/08561

Contemporary Mathematics
Volume **443**, 2007

Automatic Bias Correction Methods in Semi-supervised Learning

Hui Zou, Ji Zhu, Saharon Rosset, and Trevor Hastie

ABSTRACT. We consider the bias-correction problem in semi-supervised classification. Under certain situations where the labeled data are collected by biased sampling rather than random sampling, large margin algorithms are no longer consistent due to an inherent bias. Fortunately, with the unlabeled data we show that some bias-correction techniques can automatically correct the bias for large margin classifiers.

1. Introduction

In semi-supervised learning we have both labeled and unlabeled data. For example, in the text documents classification problem enormous amount of unlabeled documents are cheaply obtained; then some unlabeled documents are chosen and labeled by human experts as the training data for a supervised classifier. Labeling a document is time consuming and costly, thus the size of the labeled documents set is limited. In this kind of scenario, it is natural to ask whether the huge amount of unlabeled documents can help the classification. Seeger [12] and Zhu [20] gave a good literature review on this labeled-unlabeled data problem. Many positive results on the use of unlabeled data have been reported in the literature. For example, Nigam and McCallum [10] combined the Expectation-Maximization algorithm and a naive Bayes classifier for learning from both labeled and unlabeled documents in text classifications. They claimed that the use of unlabeled data reduced the classification error by up to 30%. Another popular algorithm is the transductive support vector machines [4]. However, the justification for combining labeled and unlabeled data is still unclear. Zhang and Oles [17] pointed out that the transductive SVM is unreliable since it may maximize the wrong margin. The wrong margin arguments can be viewed as a kind of modeling bias, i.e., the modeling assumption on the labeled data does not well match the model underlying the unlabeled data. To make this modeling bias argument more precise, Cozman et al. [2] showed that in semi-supervised learning of mixture models the unlabeled data can lead to an increase in classification error if the model assumptions are not exactly satisfied.

We study the use of unlabeled data in another biased situation where the labeled data are collected by biased sampling rather than random sampling. Biased sampling often occurs when it is hard to control the design of the experiments. Consider the document classification problem. Short documents are more likely to be labeled by the human experts than long documents. In medical studies,

some factors may affect patient's willingness to participate. When biased sampling occurs, supervised learning algorithms will produce inconsistent classifiers due to an inherent bias caused by biased sampling. Biased sampling is a more severe problem than the modeling bias aforementioned. Modeling bias could usually be eliminated by enlarging the model space or by replacing the strong parametric assumption with nonparametric models. However, biased sampling can not be fixed unless the modeler has prior information of the biased sampling procedure. Fortunately, in semi-supervised learning problems we have a huge amount of unlabeled data which contain rich information of the marginal distribution of the features. By taking advantage of the marginal information, we can use the unlabeled data in the supervised algorithms to correct the classification bias.

In the next section, we first highlight the important role of the random sampling assumption for the consistency of supervised classifiers. We then show the inconsistency caused by biased sampling. Section 3 describes some bias-correction techniques. In Section 4, we illustrate the proposed bias-correction methods by simulation. Section 5 contains a few concluding remarks.

2. The problem of biased sampling

2.1. Consistency and random sampling. Let $p(y, X)$ be a probability measure and $\delta(X) = \log\left(\frac{p(y=1|X)}{p(y=-1|X)}\right)$. Suppose $p(y, X)$ is the distribution generating future data in the classification problem, then $Sign(\delta(X))$ is the Bayes rule. The classification error of a classifier f is, under the 0-1 loss,

$$R_f = \int I\left(y \neq Sign(f(X))\right) dp(y, X). \tag{2.1}$$

The classification error of $Sign(\delta(X))$ is denoted as R_{Bayes}. A learning algorithm produces a classifier $\hat{f}_n$ based on training data $(y_i, X_i), i = 1, 2, \ldots, n$. $\hat{f}_n$ is said to be Bayes consistent if $R_{\hat{f}_n} \to R_{Bayes}$ in probability as $n \to \infty$.

Typically we assume that the training data (y_i, X_i) are independent and identically distributed (i.i.d.) samples from $p(y, X)$, which is referred to as random sampling in the literature. Although it is often assumed to be true without much care, the random sampling assumption is crucial for the consistency of $\hat{f}_n$. To better illustrate the point, we discuss the consistency of popular large margin classifiers. These classifiers are derived from a unified criterion

$$\hat{f}_n(X) = \arg\min_f \frac{1}{n} \sum_{i=1}^{n} \phi(y_i, f(X_i)) + \lambda_n J(f). \tag{2.2}$$

Usually ϕ is a convex surrogate for the 0-1 loss such that solving $\hat{f}_n(X)$ is a computationally tractable problem. If the regularization term $J(f) = \|f\|^2_{H_k}$, where H_k is a reproducing kernel Hilbert space (RKHS) with reproducing kernel K [16], then (2.2) is a kernel machine. For example, the support vector machine (SVM) uses the hinge loss and a reproducing kernel. Boosting minimizes the empirical ϕ risk using a different regularization method [3,7]. The exact form of regularization is not crucial in our arguments. In fact the loss function is the key in the consistency argument [1]. We need the following definition to study the limiting functional of $\hat{f}_n(X)$.

DEFINITION 2.1. Consider the infinity-sample ϕ risk $E[\phi(y, f(X))]$ where the expectation is taken with respect to some probability measure $p(y, X)$. Let the population minimizer of ϕ risk be

$$f_\infty(X) = \arg\min_f E[\phi(y, f(X))]. \tag{2.3}$$

[1] suggests that we should consider a convex loss function which produces $f_\infty(X)$ such that $Sign(f_\infty(X)) = Sign(\delta(X))$. Such a loss function is said to be classification-calibrated. For example, the hinge loss is classification-calibrated, so are the exponential loss and the logit loss.

It is now well known [13,18] that under the random sampling assumption, the classification-calibrated condition is sufficient for the consistency of $\hat{f}_n(X)$. The above arguments can be understood as follows. As $n \to \infty$, the empirical ϕ risk converges to $E_{Tr}[\phi(y, f(X))]$, where the expectation is taken with respect to the distribution of training data. The random sampling assumption implies that $E_{Tr}[\phi(y, f(X))]$ is the infinity-sample ϕ risk $E[\phi(y, f(X))]$. Thus $f_\infty(X)$ is the target function that $\hat{f}_n(X)$ tries to estimate. By assuming other technical conditions (such as λ_n vanishes at the right rate and H_k is rich) one could have $R_{\hat{f}_n} \to R_{f_\infty}$. Moreover, $Sign(f_\infty)$ is the Bayes rule as implied by the classification-calibrated condition, thus $\hat{f}_n(X)$ is Bayes consistent.

2.2. The classification bias under biased-sampling. As shown in the previous subsection, random sampling decides the target function of $\hat{f}_n$ and the classification-calibrated condition ensures the target function approximates the Bayes rule. However, the random sampling assumption is not necessarily true in many situations.

2.2.1. *Retrospective sampling.* Let us first consider a simple biased sampling scenario which is called the retrospective sampling in statistics [8]. It is easy to highlight the problems caused by biased sampling in this simple case.

Let n_+ and n_- be the sample sizes of classes $\{1\}$ and $\{-1\}$, respectively. Often $\frac{n_+}{n}$ and $\frac{n_-}{n}$ are set by the experimenter and are not random variables faithfully representing $p(y = 1)$ and $p(y = -1)$. Suppose class $\{-1\}$ indicates some disease and class $\{1\}$ means healthy. Features X are the medical measurements of each person. Since the disease is a rare event, random sampling is inefficient for gathering data to study the relation between the disease and X. A much more efficient retrospective sampling is often conducted in which all patients are included with a similar sized sample of healthy people. More generally, let us assume that a number of samples are randomly drawn from each class, while the weight of class $\{1\}$ corresponds a chosen prior π_c. π_c is not necessarily the true prior $p(y = 1)$, thus the random sampling assumption could be violated.

Let $p_c(y, X)$ be the distribution of the collected training samples. Given the label y, features X follow the true conditional distribution $p(X|y)$. Then $p_c(y, X)$ is given by

$$p_c(y = 1, X) = p(X|y = 1)\pi_c \quad \text{and} \quad p_c(y = -1, X) = p(X|y = -1)(1 - \pi_c). \tag{2.4}$$

Thus we can define

$$\delta_c(X) = \log\left(\frac{p_c(y = 1|X)}{p_c(y = -1|X)}\right) = \log\left(\frac{p(X|y = 1)\pi_c}{p(X|y = -1)(1 - \pi_c)}\right). \tag{2.5}$$

Now suppose we still use the same algorithm in subsection 2.1 to build $\hat{f}_n$. Note that $p_c(y, X)$ is the distribution of the training data, thus as $n \to \infty$ the limit of empirical ϕ risk becomes

$$\int \phi(y, f(X)) dp_c(y, X). \tag{2.6}$$

ϕ is classification-calibrated, so the sign of the minimizer of (2.6) is equal to $Sign(\delta_c(X))$. Hence $\hat{f}_n$ approximates $Sign(\delta_c(X))$. However, the true distribution is $p(y, X)$ and the Bayes rule is $Sign(\delta(X))$, which implies that $\hat{f}_n$ approaches the wrong target function. Moreover, we see $\delta(X) = \log\left(\frac{p(X|y=1)\pi}{p(X|y=-1)(1-\pi)}\right)$ and

$$\delta_c(X) - \delta(X) = \log\left(\frac{\pi_c(1-\pi)}{\pi(1-\pi_c)}\right). \tag{2.7}$$

$\delta_c(X) - \delta(X)$ is the bias due to the biased sampling.

2.2.2. *Other biased sampling schemes.* We now consider more complex label sampling processes. Assume that both labeled and unlabeled data are i.i.d. draws from a joint distribution on (X, Y, s), where $s = 1$ means that the sample is labeled and $s = 0$ means the sample is unlabeled. For the labeled data, their distribution is $p_c(y, X) = p(y, X|s = 1)$. We are particularly interested in the case where $p(s|y, X)$ depends on the (potentially unobserved) response y, which is similar to the nonignorable nonresponse situation in the missing data literature [6].

Any consistent classifier built on the labeled data will consistently estimate the Bayes rule under $p_c(y, X)$. Denote the likelihood ratio function by

$$LR(y, X) = \frac{p(y, X|S = 1)}{p(y, X)}. \tag{2.8}$$

Then we have

$$\delta_c(X) - \delta(X) = \log\left(\frac{LR(y = 1, X)}{LR(y = -1, X)}\right) \equiv \text{Bias}(X). \tag{2.9}$$

The biased sampling happens whenever Bias(X) defined in (2.9) is a non-zero function of X. The consequence is that $R_{\delta_c(X)} > R_{\delta(X)} = R_{Bayes}$, thus *under biased sampling, $\hat{f}_n$ is asymptotically sub-optimal.*

It is important to note that the bias exists for all algorithm being consistent under random sampling. Thus it is desirable to have a general bias-correction technique that works for all the large margin classifiers.

3. Automatic Bias-Correction Methods

In this section we show that the bias can be estimated and eliminated if one uses the unlabeled data appropriately. In this sense the unlabeled data are always helpful in semi-supervised settings.

3.1. Inferring the bias. Suppose we have n labeled data and m unlabeled data. In semi-supervised setting $m \gg n$. Thus the unlabeled data provide rich information about the marginal distribution of X which is useful for inferring the sampling bias in the labeled data.

To motivate, let us first consider the simple yet important retrospective sampling case and present the bias estimation technique. In retrospective sampling, it

suffices to estimate π to infer the bias. We show a method for constructing some $\hat{\pi}$ using the unlabeled data such that

$$\hat{\pi} \to \pi \quad \text{in probability as } n \to \infty . \tag{3.1}$$

Our method is based on arguments from a method of moments. We evaluate the expectation of a function $g(X)$: $E_p[g] = \int g(X)p(X)dX$. By Bayes theorem

$$p(X) = p(X|y=1)\pi + p(X|y=-1)(1-\pi), \tag{3.2}$$

so we have

$$E_p[g] = \pi E_{p(X|y=1)}[g] + (1-\pi)E_{p(X|y=-1)}[g]. \tag{3.3}$$

Hence if $E_{p(X|y=1)}[g] \neq E_{p(X|y=-1)}[g]$, we have the identity

$$\pi = \frac{E_p[g] - E_{p(X|y=-1)}[g]}{E_{p(X|y=1)}[g] - E_{p(X|y=-1)}[g]}. \tag{3.4}$$

Using the empirical distribution (or the sample moments), the unlabeled data give a consistent estimate for $E_p[g]$. Similarly, $E_{p(X|y=1)}[g]$ and $E_{p(X|y=-1)}[g]$ can be consistently estimated by the labeled data. Therefore, a simple consistent estimate for π is given by

$$\hat{\pi} = \frac{\hat{E}_p[g] - \hat{E}_{p(X|y=-1)}[g]}{\hat{E}_{p(X|y=1)}[g] - \hat{E}_{p(X|y=-1)}[g]}. \tag{3.5}$$

Note that we do not use $g = 1$ in (3.5) to avoid a zero denominator. The derived estimate for π is fully non-parametric, requiring no assumption on $p(X)$ or $p(X|y)$. Estimating π of a mixture normal by equations like (3.4) is not new in statistics, e.g. see [9,14]. We show in this work that the method of moments can be generalized to estimate more complicated biased sampling schemes.

For more general label sampling distributions we need to estimate the bias given in (2.9). Note that

$$LR(y,X) = \frac{p(y,X|s=1)}{p(y,X)} = \frac{p(y,X,s=1)}{p(s=1)p(y,X)} = \frac{p(s=1|y,X)}{p(s=1)}. \tag{3.6}$$

Thus (3.6) says that

$$\text{Bias}(X) = \log\left(\frac{p(s=1|y=1,X)}{p(s=1|y=-1,X)}\right). \tag{3.7}$$

If we have a consistent estimate of $p(s=1|y,X)$, then we can estimate the bias consistently. Let $r(X) = e^{\text{Bias}(X)}$. In retrospective sampling, $r(X)$ can be consistently estimated by $\frac{n_+(1-\hat{\pi})}{n_-\hat{\pi}}$. It is worth pointing out that if $p(s|Y,X)$ is independent of y, then the bias becomes zero. That is why we only consider the nonignorable nonresponse situation in this work.

The moment-based inversion method can still be applied to estimate $p(s=1|y,X)$ if $p(s=1|y,X) = l_\theta(y,X)$ as a function of θ, as discussed in [11]. Let θ be a k vector. We suggest the following method to estimate θ

(1) Select k different real-valued functions $g_1(X), ..., g_k(X)$, and set the equation for each of $g_1, ..., g_k$

$$\sum_{i=1}^{n} \frac{g_j(X_i)}{l_\theta(y_i,X_i)} = \sum_{v=1}^{m+n} g_j(x_v) \quad j = 1, \ldots, k$$

(2) solve the resulting k equations to get an estimate of θ.

Then we estimate $r(X)$ by $\hat{r}(X) = \frac{l_{\hat{\theta}}(1,X)}{l_{\hat{\theta}}(-1,X)}$.

The key in the moment-based inversion method is to observe that we can evaluate the expectation of $g(X)$ using the unlabeled data. We can justify the consistency of $\hat{\theta}$ as follows. For simplicity we assume $k = 1$. From the moment equation we get

$$\hat{\theta} - \theta = \frac{\frac{1}{n}\sum_{v=1}^{m+n} g(X_v) - \frac{1}{n}\sum_{i=1}^{n} g(X_i) l_\theta^{-1}(y_i, X_i)}{\frac{1}{n}\sum_{i=1}^{n} g(X_i)\left(\frac{l_{\hat{\theta}}^{-1}(y_i,X_i) - l_\theta^{-1}(y_i,X_i))}{\hat{\theta}-\theta}\right)} \tag{3.8}$$

Note that

$$\frac{1}{n}\sum_{j=1}^{m+n} g(X_j) = \frac{m+n}{n}\frac{1}{n+m}\sum_{j=1}^{m+n} g(X_j) \to \frac{E[g(X)]}{p(s=1)} \tag{3.9}$$

$$\frac{1}{n}\sum_{i=1}^{n} \frac{g(X_i)}{l_\theta(y_i, X_i)} \to E_{p_c(y,X)}\left[\frac{g(X)p(s=1)}{p(s=1)l_\theta(y,X)}\right] = \frac{E[g(X)]}{p(s=1)}. \tag{3.10}$$

Under some regularity conditions, the denominator in (3.8) is approximately

$$E_{p_c(y,X)}\left[g(X)\frac{\partial l_\theta^{-1}(y,X)}{\partial\theta}\right] = -E[g(X)\frac{\dot{l}_\theta(y,X)}{l_\theta(y,X)}]\frac{1}{p(s=1)}, \tag{3.11}$$

where $\dot{l}_\theta(y,X) = \frac{\partial l_\theta^{-1}(y,X)}{\partial\theta}$. So we see $\hat{\theta}$ is consistent as long as (3.11) is not zero. It is good to require $E[\frac{\dot{l}_\theta(y,X)}{l_\theta(y,X)}|X] \neq 0$. Otherwise, for any function $g(X)$ we must have

$$E[g(X)\frac{\dot{l}_\theta(y,X)}{l_\theta(y,X)}] = E[g(X)E[\frac{\dot{l}_\theta(y,X)}{l_\theta(y,X)}|X]] = E[g(X)0] = 0. \tag{3.12}$$

Using $g(X) = 1$ seems to be a convenient choice. There are many other considerations in choosing a good g functions. On the other hand, small improvement in the estimation of θ (π) may not lead to significant changes in the bias-corrected classifiers, as shown in our experiments. In that sense it suffices to find a good consistent estimate of θ (π).

Finally, it should be pointed out that our method directly focuses on $l_\theta(y,X)$ without modeling $p(y|X)$. This is a significant advantage over the modeling approach, especially when $p(y|X)$ involves a lot of parameters. For example, suppose $p(y|X) = p_\xi(y|X)$ where X and ξ are both 50-dimension vectors, meanwhile θ is a scalar. Jointly modeling of (θ, ξ) will seek an estimator in a 51-dimension space, which is a much more difficult problem than estimating θ alone. Moreover, if $p(y|X) = p_\xi(y|X)$ is not the correct model, then jointly estimating (ξ, θ) could result in an inconsistent estimate of θ.

3.2. Estimation-calibrated classifiers. For many ϕ loss functions, not only the resulting classifier $Sign(f_\infty(X))$ agrees with the Bayes rule, but also $f_\infty(X)$ itself is related to $p(y=1|X)$ via a deterministic function. We give a formal definition to this phenomenon.

DEFINITION 3.1. A loss function ϕ is said to be *estimation-calibrated* with ψ, if $p(y=1|X) = \psi(f_\infty(X))$ for some continuous nondecreasing function $\psi : \mathbb{R} \to [0,1]$ such that $\psi(0) = \frac{1}{2}$ and $\psi(a) < \frac{1}{2}\ \forall\ a < 0$, $\psi(a) > \frac{1}{2}\ \forall\ a > 0$.

The next theorem describes a family of estimation-calibrated losses and gives the explicit expression of ψ for a given ϕ.

THEOREM 3.2. *Let ϕ be a margin-based loss function, i.e.,$\phi(y, f(X)) = \phi(yf(X))$. Suppose ϕ is strictly convex and ϕ' is continuous, then ϕ is estimation-calibrated if and only if $\phi'(0) < 0$.*

PROOF. For the *necessary condition*, note that any estimation-calibrated ϕ must be classification-calibrated, then the convexity of ϕ requires $\phi'(0) < 0$ by Theorem 6 in [1].

Sufficient condition. Let $a^* = \inf\{a : \phi'(a) = 0\}$, if it is an empty set, $a^* = \infty$. $\phi'(0) < 0$ implies $a^* > 0$. Fix a X, $f_\infty(X)$ minimizes

$$g(a) = p(y = 1|X)\phi(a) + p(y = 1|X)\phi(-a).$$

Differentiating g, we have

$$p(y = 1|X)\phi'(f_\infty(X)) = p(y = -1|X)\phi'(-f_\infty(X)),$$

thus $f_\infty(X) \in (-a^*, a^*)$ if $0 < p(y = 1|X) < 1$. We show $\frac{\phi'(-a)}{\phi'(a)}$ is increasing in $(-a^*, a^*)$. Let $a_1 > a_2$, then $\phi'(-a_1) < \phi'(-a_2)$ and $\phi'(a_1) > \phi'(a_2)$. Note that $\phi(a) < 0\ \forall a \in (-a^*, a^*)$, so $\frac{\phi'(-a_1)}{\phi'(a_1)} > \frac{\phi'(-a_2)}{\phi'(a_2)}$. Define ψ as follows:

$$\psi(a) = \frac{\phi'(-a)}{\phi'(-a) + \phi'(a)},\ a \in (-a^*, a^*). \tag{3.13}$$

If $a^* < \infty$, then let $\psi(a) = 1\ \forall a \geq a^*; \psi(a) = 0\ \forall a \leq -a^*$. ψ is the desired function in definition 2. Thus ϕ is estimation-calibrated. □

To illustrate theorem 3.2, let us consider kernel logistic regression where we take $\phi(t) = \log(1 + e^{-t})$. $\phi'(t) = -\frac{1}{1+e^t}$. Thus from the proof of theorem 3.2 we see $a^* = \infty$ and

$$\psi(\hat{f}_n(X)) = \frac{\phi'(-\hat{f}_n(X))}{\phi'(-\hat{f}_n(X)) + \phi'(\hat{f}_n(X))} = \frac{1}{1 + e^{-\hat{f}_n(X)}}. \tag{3.14}$$

Once we obtain a consistent estimate of the bias, we can use the following method to do the bias correction.

LEMMA 3.3. *Suppose ϕ is estimation-calibrated with ψ and assume the technical conditions in subsection 2.1. Then Sign $\left(\psi(\hat{f}_n(X)) - \frac{\hat{r}(X)}{1+\hat{r}(X)}\right)$ is Bayes consistent.*

PROOF. Under the technical conditions we know $\hat{f}_n(X)$ converges to the population minimizer of ϕ risk under the distribution $p_c(y, X)$ [13,18]. Thus, by the continuity of ψ we have $\psi(\hat{f}_n(X)) \to p_c(y = 1|X)$. Or $\log\left(\frac{\psi(\hat{f}_n(X))}{1-\psi(\hat{f}_n(X))}\right) \to \delta_c(X)$. From $\hat{r}(X) \to$ Bias we see $\log\left(\frac{\psi(\hat{f}_n(X))}{1-\psi(\hat{f}_n(X))}\right) - \hat{r}(X) \to \delta(X)$. Then note that $Sign\left(\log\left(\frac{\psi(\hat{f}_n(X))}{1-\psi(\hat{f}_n(X))}\right) - \hat{r}(X)\right) = Sign\left(\psi(\hat{f}_n(X)) - \frac{\hat{r}(X)}{1+\hat{r}(X)}\right)$. □

3.3. Weighted loss. It is easy to see that any estimation-calibrated loss must be classification-calibrated, but the reverse is not always true. For example, the hinge loss population minimizer is exactly the Bayes rule, providing no information about $p(y = 1|X)$. This is actually regarded as a major drawback of SVMs [3,19]. Even if ϕ is not estimation-calibrated, the consistency can be ensured by a weighted loss technique, as long as ϕ is classification-calibrated.

Define $w_{y,X} = \frac{p(y,X)}{p(y,X|s=1)}$. Then we have the following identity

$$E_{p_c(y,X)}[\phi(y,X)w_{y,X}] = E_{p(y,X)}[\phi(y,X)]. \tag{3.15}$$

The above identity comes from the importance sampling idea.

If we knew the weights, we could consider the weighted ϕ loss

$$\frac{1}{n}\sum_{i=1}^{n} w_{y_i,X_i}\phi(y_i,f_i). \tag{3.16}$$

Then combining the law of large numbers and (3.15) yields

$$\lim \frac{1}{n}\sum_{i=1}^{n} w_{y_i,X_i}\phi(y_i,f_i) = E_{p(y,X)}[\phi(y,X)]. \tag{3.17}$$

Therefore, the minimizer of the weighted ϕ loss will try to estimate f_∞, the right target function.

The weights are unknown in the semi-supervised problem. We use the *plug-in principle* to replace the weights with their consistent estimates to construct the weighted loss. For example, in the retrospective sampling case, consistent estimates can be $\hat{w}_{1,X} = \frac{n\hat{\pi}}{n_+}$ and $\hat{w}_{-1,X} = \frac{n(1-\hat{\pi})}{n_-}$, where n_+ and n_- denote the number of labeled data in class $\{1\}$ and $\{-1\}$, respectively. In general, note that $w_{y,X} = \frac{p(s=1)}{p(s=1|y,X)}$. An obvious consistent estimate of $p(s=1)$ is $\frac{n}{n+m}$. If we assume $p(s=1|y,X) = l_\theta(y,X)$, then θ can be estimated by the method in section 3.1, so we estimate the importance sampling weights by $\hat{w}_{y_i,X_i} = \frac{\frac{n}{n+m}}{l_{\hat{\theta}}(y_i,X_i)}$. A bias-corrected classifier is then obtained by minimizing the weighted ϕ loss

$$\hat{f}_n^w(X) = \arg\min_f \{\frac{1}{n}\sum_{i=1}^{n} \hat{w}_{y_i,X_i}\phi(y_i,f_i) + \lambda_n J(f)\}. \tag{3.18}$$

When the classifier is constructed via penalized likelihood method, such as penalized logistic regression, the weighted loss becomes weighted likelihood. In fact, the weighted loss technique has its root in statistical theory. Vardi [15] considered the non-parametric maximum likelihood estimate of some CDF F on the basis of samples from weighted version of F, with a known weight function. Vardi's method replaces the usual non-parametric maximum likelihood criterion with its weighted version. Lin et al. [5] proposed a similar technique for SVMs under nonstandard situations. In their work the weights are known constants. Our work has three additional contributions: (1) it works for all classification-calibrated loss functions, not limited to the hinge loss of the SVM; (2) the weights are adaptively estimated by using the unlabeled data; and (3) the weights can be feature-dependent.

4. Experiments

In this section we use simulation to demonstrate the efficacy of the proposed bias-correction techniques.

Model 1. Following the mixture model setup in [3], we designed $p(X|y=1)$ and $p(X|y=-1)$ as a mixture of 10 Gaussians such that the Bayes decision boundary is nonlinear. We set $p(y=1) = 0.75$ and simulated 5000 unlabeled data (X) from $p(X) = 0.75 \cdot p(X|y=1) + 0.25 \cdot p(X|y=-1)$. We generated 100 labeled data (y,X) points from each class. The SVM and kernel logistic regression (KLR) [19] classifiers were fitted on the labeled data. We used the 5000 unlabeled data to

TABLE 1. Model 1: estimate of π.

π	$g(X)$	$\hat{\pi}$	MSE($\hat{\pi}$)
0.75	$x_1 - x_2$	0.762	0.003

TABLE 2. Model 1: error rates of KLR and SVM before and after bias-correction.

KLR	WKLR	KLR-EC	SVM	WSVM
0.200	0.154	0.150	0.202	0.148

help us estimate π. Based on the discussion in section 3.1 we used $g(X) = x_1 - x_2$ to estimate π, and the estimated $\hat{\pi}$ was used in the weighted SVM (WSVM) and weighted KLR (WKLR). For each classifier, its test error rate was calculated by the true model $p(y, X)$. The above process was repeated 100 times.

Table 1 shows that the moment estimator is accurate. From Table 2, we see that the SVM and KLR based on the biased data have errors around 0.20. However, bias-correction reduces the misclassification errors of KLR and the SVM to 0.15. Moreover, the KLR using the estimation-calibrated bias-correction formula (KLR-EC) performs similarly to the weighted KLR.

Model 2. In the second example we considered a biased sampling scheme which depends on the response y as well as features X. Like in model 1, we designed $p(X|y = 1)$ and $p(X|y = -1)$ as a mixture of 10 Gaussians and we let $p(y = 1) = 0.5$ such that this model is exactly identical to the one used in [3] (Chap 2). Let us first generate 5000 data (y, X) from $p(y, X)$ described in model 1. Then we kept y based on probability

$$p(s = 1|y, X) = \frac{1}{2 + y} e^{-\theta(|x_1| + |x_2|)}$$

with $\theta = 2$. We ended up with about 155 labeled data. Before bias-correction, the SVM and KLR perform very poorly with error rates 0.41 and 0.45, respectively. Note that with 160 random samples from $p(y, X)$ the error rates of KLR and the SVM are 0.31. We tried three functions, $g_1(X) = 1$, $g_2(X) = \frac{1}{|x_1| + |x_2|}$ and $g_3(X) = |x_1| + |x_2|$, to estimate θ and then applied the bias-correction methods on KLR and the SVM. The simulation was repeated 100 times.

As can be seen from Table 3, g_1 and g_2 give very similar estimates and their estimators have smaller mean squared errors than the estimator by g_3. But these three estimators give almost identical bias-correction results. This observation suggests that as long as we could obtain a reasonably good estimator of θ, the choice of g function is not critical for the performance of the bias-corrected classifier. With estimated weights, the bias-correction techniques reduce the error rates of KLR and the SVM to about 0.34.

5. Discussion

We have emphasized the importance of the sampling scheme, which is often considered nuisance in machine learning. One should be cautious when the sampling distribution is not the target distribution of learning. The biased sampling problem has been discussed in statistics for a long time. The existing work mainly focuses on the bias correction method for the traditional discriminative methods such as

TABLE 3. Model 2: estimates of θ and error rates of KLR and the SVM after bias-correction. Before bias-correction, the error rates of KLR and the SVM are 0.45 and 0.42.

$g(X)$	$\hat{\theta}$	MSE($\hat{\theta}$)	WKLR	KLR-EC	WSVM
1	2.278	0.124	0.349	0.338	0.339
$\frac{1}{\lvert x_1\rvert+\lvert x_2\rvert}$	2.280	0.134	0.349	0.338	0.338
$\lvert x_1\rvert+\lvert x_2\rvert$	2.325	0.178	0.349	0.338	0.340

logistic regression and Fisher's LDA, see [8,9] and references therein. Their methods require strong assumptions on $p(y|X)$ but modern margin-based classifiers are fully nonparametric. So it is not clear if their methods could work well with large margin classifiers. We have considered a new approach based on the method of moments to take advantage of the enormous amount of unlabeled data to estimate the bias caused by biased sampling. We have also considered two bias-correction techniques for the large margin classifiers. Numerical experiments are very encouraging.

It should be pointed out that choosing the right g functions is the key to in the moment-based inversion approach, especially when the dimension of θ increases. Qualitatively, we would like to use the g functions yielding moment functions that have a unique, stable and robust solution. However, this is the problem for many inverse problems and it is very difficult to give concrete instructions for selecting the good g functions in general.

In this paper we have assumed that the biased sampling scheme follows a parametric model and used the moment-based inversion method to estimate the unknown parameter. The moment-based inversion idea can also be effectively used in the nonparametric models. We could construct a cost function based on the moment equations like those in Section 3 and then employ the kernel methods or boosting algorithms to find a regularized minimizer of the cost function. This is an interesting direction deserving further exploration.

References

[1] Bartlett, P., Jordan, M. & McAuliffe, J. (2003) Convexity, classification, and risk bounds. *Technical Report, Division of Computer Science and Department of Statistics, U.C. Berkeley.*

[2] Cozman, F. G. & Cohen, I. & Cirelo, M. C. (2003) Semi-Supervised learning of mixture models. *Proceedings of the Twentieth International Conference on Machine Learning*, 99-106.

[3] Hastie, T., Tibshirani, R. & Friedman, J. (2001) *The Elements of Statistical Learning.* Springer-Verlag, New York.

[4] Joachims, T. (1999) Transductive inference for text classification using support vector machines. *Proceedings of the Sixteenth International Conference on Machine Learning* 350-358. San Francisco: Morgan Kaufmann.

[5] Lin, Y., Lee, Y. & Wahba, G. (2000) Support vector machines for classification in nonstandard situations. *Machine Learning*, **46**, 191-202.

[6] Little, R. & Rubin, D. (2002) *Statistical Analysis with Missing Data, 2nd Ed.* . Wiley & Sons.

[7] Lugosi, G. & Vayatis, N. (2004) On the Bayes risk consistency of regularized boosting methods. *Annals of Statistics*, **32**, 30-55.

[8] McCullagh, P. & Nelder, J. A. (1989) *Generalized Linear Models, 2nd edition.* Chapman & Hall.

[9] McLachlan, G. (1992) *Discriminant Analysis and Statistical Pattern Recognition.* Wiley.

[10] Nigam, K., McCallum, A. K., Thrun, S. & Mitchell, T. (2000) Text classification from labeled and unlabeled documents using EM. *Machine Learning*, **39** 103-134.

[11] Rosset, S., Zhu, J., Zou, H. & Hastie, T. (2005) A Method for Inferring Label Sampling Mechanisms in Semi-Supervised Learning. In *Advances in Neural Information Processing Systems*, **17**.

[12] Seeger, M. (2002) Learning with labeled and unlabeled data. *Technical Report, Institute for Adaptive and Neural Computation, University of Edinburgh.*

[13] Steinwart, I. (2003) Consistency of support vector machines and other regularized kernel classifiers. *Technical Report, Los Alamos National Laboratory, Los Alamos, NM*

[14] Titterington, D., Smith, A. & Makov, U (1985) *Statistical Analysis of Finite Mixture Distributions.* Wiley.

[15] Vardi, Y. (1985) Empirical distributions in selection bias models. *Annals of Statistics*, **13**, 178-203.

[16] Wahba, G. (1999) Support vector machine, reproducing kernel Hilbert spaces and the randomized GACV. *Advances in Kernel Methods-Support Vector Learning*, 69-88, MIT press.

[17] Zhang, T. & Oles, F. (2000) A probability analysis on the value of unlabeled data for classification problems. *Int. Joint Conf. on Machine Learning*, 1191-1198.

[18] Zhang, T. (2004) Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, **32**, 56-85.

[19] Zhu, J. & Hastie, T. (2002) Kernel logistic regression and the import vector machine. In *Advances in Neural Information Processing Systems*, **14**.

[20] Zhu, X (2005) Semi-Supervised Learning Literature Survey. Computer Sciences TR 1530, University of Wisconsin - Madison.

SCHOOL OF STATISTICS, UNIVERSITY OF MINNESOTA, MINNEAPOLIS, MN 55455
E-mail address: hzou@stat.umn.edu

DEPARTMENT OF STATISTICS, UNIVERSITY OF MICHIGAN,ANN ARBOR, MI 48109
E-mail address: jizhu@umich.edu

IBM T.J. WATSON RESEARCH CENTER, YORKTOWN HEIGHTS, NY 10598
E-mail address: srosset@us.ibm.com

DEPARTMENT OF STATISTICS, STANFORD UNIVERSITY, STANFORD, CA 94305
E-mail address: hastie@stat.stanford.edu