Maximum Mean Discrepancy for Class Ratio Estimation: Convergence Bounds and Kernel Selection

Arun Iyer* ARUNIYER@YAHOO-INC.COM
Yahoo! Labs, Bangalore, Karnataka 560071 INDIA

Saketh Nath[1] SAKETH@CSE.IITB.AC.IN
Sunita Sarawagi[1] SUNITA@CSE.IITB.AC.IN
[1]IIT Bombay, Powai, Mumbai, Maharashtra 400076 INDIA

Abstract

In recent times, many real world applications have emerged that require estimates of class ratios in an unlabeled instance collection as opposed to labels of individual instances in the collection. In this paper we investigate the use of maximum mean discrepancy (MMD) in a reproducing kernel Hilbert space (RKHS) for estimating such ratios.

First, we theoretically analyze the MMD-based estimates. Our analysis establishes that, under some mild conditions, the estimate is statistically consistent. More importantly, it provides an upper bound on the error in the estimate in terms of intuitive geometric quantities like class separation and data spread. Next, we use the insights obtained from the theoretical analysis, to propose a novel convex formulation that automatically learns the kernel to be employed in the MMD-based estimation. We design an efficient cutting plane algorithm for solving this formulation. Finally, we empirically compare our estimator with several existing methods, and show significantly improved performance under varying datasets, class ratios, and training sizes.

1. Introduction

The goal of this work is to estimate the ratio of classes in any unlabeled test dataset given a labeled training dataset with an arbitrarily different ratio of classes. The closely related problem of creating a classifier with shifted class priors in the training and test set has been extensively studied. In contrast, our end goal is to estimate the ratio of classes and not the labels of individual instances in the unlabeled set. Many real-world applications have emerged that motivate this problem. As an example consider websites that serve user directed content like news, videos, or reviews. Each item (article, video, or product) is associated with many user comments. An analyst wants to estimate the fraction of comments that express positive/negative sentiments. The polarity of each comment is not of interest.

We now describe our problem formally. Let $\mathcal{X} = \{\mathbf{x} \in \mathbf{R}^d\}$ be the set of all instances and $\mathcal{Y} = \{0, 1, \ldots, c\}$ be the set of all labels. We are given a labeled dataset $D(\subset \mathcal{X} \times \mathcal{Y})$. Our goal is to design an estimator that for any given set $U(\subset \mathcal{X})$ can estimate the class ratios $\boldsymbol{\theta} = [\theta_0, \theta_1, \ldots, \theta_c]^\top$ where θ_y denotes the fraction of instances with class label y in U. We consider the case where the estimator will have to handle unlabeled data with widely varying values of $\boldsymbol{\theta}$, which might be different from those in the training data. As in all existing work on the topic (discussed next) we assume that the $\Pr(\mathbf{x}|y)$ distribution remains unchanged in the training and test distributions.

A baseline estimator is to use the labeled data D to train a classifier $C : \mathcal{X} \mapsto \mathcal{Y}$ using supervised learning techniques and estimate θ_y as $\frac{\hat{n}_y}{n_u}$ where n_u is the size of U and $\hat{n}_y$ is the number of instances of U labeled y by C. Since most supervised learning algorithms assume that the training and test data follow the same distribution, this method is unlikely to perform well. Many fixes have been proposed: for example, (Cortes & Mohri, 2004; Clémençon et al., 2009) propose to maximize the area under the ROC to make the classifier work for all possible class ratios, (Selvaraj et al., 2011) proposes transductive learning, while (Elkan, 2001) and (Lin et al., 2002) assume that the true class ratios are known, and propose to reweight instances or rebalance the decision cutoff. However, the primary goal of these methods is to improve per-instance accuracy, and class ratios are estimated using the same paradigm of aggregating from per-instance predictions.

Since we are not interested in the labels of individual in-

* This work was done by the author as part of the Sponsored PhD Program with IIT Bombay.

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

stances, we explore direct methods of estimating the class ratios. There have been three reported attempts for such direct estimation. The first is an EM-based approach (Saerens et al., 2002) that alternates between estimating the class ratios from per-instance predictions and 'correcting' the predictions using the estimated class ratios as priors. The second approach (Plessis & Sugiyama, 2012), proposes to minimize various f-divergence measures between the test distribution and a model distribution parameterized on the unknown $\boldsymbol{\theta}$s and instance-level ratios of two distributions. This method is developed in a semi-supervised setting and involves solving an elaborate optimization problem for each test set. The third, most recent approach (Zhang et al., 2013), is based on minimizing the maximum mean discrepancy (MMD) over functions in a reproducing kernel Hilbert space (RKHS) induced by a kernel K. The MMD-based approach has several advantages: it is applicable to arbitrary domains since it does not assume any parametric density on the data unlike conventional mixture modeling approaches (Titterington, 1983; Woodward et al., 1984). Because of this property, MMD has been successfully deployed in other problems including covariance shift (Gretton et al., 2009) and two-sample test (Gretton et al., 2012a). When deployed for class ratio estimation, it gives rise to a convex QP over a small number of variables, which is efficiently solvable. However, the approach has not been understood theoretically, empirical comparisons with other class ratio estimation methods are lacking, and kernel selection has not been adequately addressed.

In this paper we address the above limitations of the MMD approach. Specifically, we make these contributions:

We theoretically analyze the MMD-based estimator for arbitrary number of classes. Under some mild conditions, we show that the estimator is statistically consistent. In addition to asymptotic convergence rates, we derive empirical bounds that involve intuitive geometric quantities and motivate a kernel learning method. Analysis of the MMD-based estimator is non-trivial and requires bounding techniques that exploit the nature of the MMD objective in addition to typical concentration inequalities employed in learning theory. We are aware of no work that bounds the error of class ratio estimates by any other method.

We use the insights obtained from the theoretical analysis, to propose a novel convex formulation for selecting the best kernel to be employed in the MMD-based estimation procedure. Our kernel learning formulation turns out to be an instance of a Semi-Definite Program (SDP) with infinitely many linear constraints.

Since it is expected that at optimality only a few of the linear constraints are active, we propose a cutting-plane based algorithm for solving this formulation. At every iteration, the SDP restricted to the current constraint-set is solved using a simple projected sub-gradient descent algorithm. We are aware of no prior work on kernel selection for this problem.

We present an extensive evaluation of several existing methods, both from the direct and per-instance aggregation family, under varying true class ratios and training sizes. We obtain up to 60% reduction in class ratio estimation errors over the best existing method.

Outline: In Section 2, we present an overview of the MMD-based approach and analyze it theoretically. In Section 3, we provide our formulation for learning a kernel function for improved estimates. In Section 4 we present empirical comparisons and conclude in Section 5.

2. The Maximum Mean Discrepancy approach

The core idea in maximum mean discrepancy (MMD) in a reproducing kernel Hilbert space (RKHS) is to match two distributions based on the mean of features in the Hilbert space induced by a kernel K. This is justified because when K is universal there is an injection between the space of distributions and the space of mean feature vectors lying in its RKHS. From a practical perspective too, the MMD approach is appealing because unlike other parametric density estimation methods, it can be applied to arbitrary domains and to high-dimensional data, and is computationally tractable. This approach was earlier used in the covariance shift problem (Gretton et al., 2009), the two-sample problem (Gretton et al., 2012a), and recently in (Zhang et al., 2013) for estimating class ratios.

Let K be a universal kernel and $\mathcal{H}$ denote the RKHS induced by K. Let $\Phi : \mathcal{X} \mapsto \mathcal{H}$ denote the canonical feature map induced by the kernel on the RKHS. We expect the test distribution $P_U(\mathbf{x})$ to match the distribution $Q(\mathbf{x}) = \sum_y P_D(\mathbf{x}|y)\theta_y$ where θ_y denotes the unknown probability of class y instances in U's distribution and $P_D(\mathbf{x}|y)$ denotes the training distribution for class y. This holds because as mentioned earlier, we assume that

$$P_U(\mathbf{x}|y) = P_D(\mathbf{x}|y),\ \forall y \in \mathcal{Y} \qquad (\mathcal{A}1)$$

Let $\bar{\Phi}_y$ and $\bar{\Phi}_u$ denote the true means of the feature vectors of the y-th class and unlabeled data respectively. That is,

$$\bar{\Phi}_y = \mathbb{E}_{P_D(\mathbf{x}|y)}\Phi(\mathbf{x}), \qquad \bar{\Phi}_U = \mathbb{E}_{P_U(\mathbf{x})}\Phi(\mathbf{x})$$

The true mean feature vector for the $Q(\mathbf{x})$ distribution is then $\sum_{y\in\mathcal{Y}} \theta_y \bar{\Phi}_y$. To match $Q(\mathbf{x})$ and $P_U(\mathbf{x})$, the MMD approach minimizes the distance between the two means. This gives rise to the following optimization problem over the unknown $\boldsymbol{\theta}$s:

$$\min_{\boldsymbol{\theta}:\theta_y \geq 0, \sum_{y=0}^{c}\theta_y = 1} \| \sum_{y\in\mathcal{Y}} \theta_y \bar{\Phi}_y - \bar{\Phi}_U \|^2 \qquad (1)$$

or, after rewriting using $\theta_0 = 1 - \sum_{y=1}^{c}\theta_y$ as

$$\min_{\boldsymbol{\theta}\in\wedge^c} \ \left\|\bar{A}\boldsymbol{\theta} - \bar{a}\right\|^2 \tag{2}$$

where $\bar{A} = [\bar{\Phi}_1 - \bar{\Phi}_0 \ \ldots \ \bar{\Phi}_c - \bar{\Phi}_0]$ and $\bar{a} = [\bar{\Phi}_U - \bar{\Phi}_0]$ and $\wedge^c$ denotes the new feasibility set $\{\theta_y \geq 0, \sum_{y=1}^{c}\theta_y \leq 1\}$. Let $\boldsymbol{\theta}^*$ denote the solution of the above, which is unique because of the following two assumptions:

$$\begin{array}{ll} K \text{ is universal} & (\mathcal{A}2) \\ \forall\boldsymbol{\theta} \neq \boldsymbol{\theta}', \sum_y \theta_y \Pr(\mathbf{x}|y) \neq \sum_y \theta'_y \Pr(\mathbf{x}|y) & (\mathcal{A}3) \end{array}$$

$\mathcal{A}2$ implies that there is an injection from the space of distributions to the space of mean feature vectors. $\mathcal{A}3$ is a standard identifiability assumption on $\boldsymbol{\theta}$ without which the class ratio estimation problem is undefined and no algorithm can identify the true class ratio.

However Equation 2 is impossible to solve as $\bar{\Phi}_y$ and $\bar{\Phi}_u$ are unknowns. So, we approximate them by substituting sample means from sample U of $P_U(\mathbf{x})$, and labeled sample D of $P_D(\mathbf{x}|y)$ calculated as

$$\widehat{\Phi}_y(n_y) = \sum_{(\mathbf{x},y)\in D}\frac{\Phi(\mathbf{x})}{n_y}, \quad \widehat{\Phi}_U(n_u) = \sum_{\mathbf{x}\in U}\frac{\Phi(\mathbf{x})}{n_u}$$

Here, n_u denotes the number of instances in U, and n_y denotes the number of instances labeled y in D. The empirical version of the MMD above is:

$$\min_{\boldsymbol{\theta}\in\wedge^c} \ \left\|\widehat{A}(n)\boldsymbol{\theta} - \widehat{a}(n)\right\|^2 \tag{3}$$

where $\widehat{A}(n) = [\widehat{\Phi}_1(n_1) - \widehat{\Phi}_0(n_0) \ \ldots \ \widehat{\Phi}_c(n_c) - \widehat{\Phi}_0(n_0)]$ and $\widehat{a}(n) = [\widehat{\Phi}_U(n_u) - \widehat{\Phi}_0(n_0)]$. We call the solution to this the MMD estimate $\widehat{\theta}(n)$. In the paper we sometimes drop the (n) argument from $\widehat{A}, \widehat{\Phi}, \widehat{a}, \widehat{\theta}$ to reduce clutter.

We can apply the Kernel trick on this objective and rewrite it in kernel form as follows:

$$\min_{\boldsymbol{\theta}\in\wedge^c} \ \boldsymbol{\theta}^\top[\widehat{A}^\top\widehat{A}]\ \boldsymbol{\theta} - 2\boldsymbol{\theta}^\top[\widehat{A}^\top\widehat{a}] \tag{4}$$

where an entry yy' of $[\widehat{A}^\top\widehat{A}] = \widehat{\Phi}_y^\top\widehat{\Phi}_{y'} - \widehat{\Phi}_y^\top\widehat{\Phi}_0 - \widehat{\Phi}_0^\top\widehat{\Phi}_{y'} + \widehat{\Phi}_0^\top\widehat{\Phi}_0$ and each $\widehat{\Phi}_y^\top\widehat{\Phi}_{y'}$ can be written in terms of a kernel as $\frac{1}{n_y n_{y'}}\sum_{(\mathbf{x},y),(\mathbf{x}',y')\in D} K(\mathbf{x},\mathbf{x}')$. Similarly an entry y of $[\widehat{A}^\top\widehat{a}] = \widehat{\Phi}_y^\top\widehat{\Phi}_U - \widehat{\Phi}_y^\top\widehat{\Phi}_0 - \widehat{\Phi}_0^\top\widehat{\Phi}_U + \widehat{\Phi}_0^\top\widehat{\Phi}_0$ can be written in terms of kernel. Since the objective is convex in $\boldsymbol{\theta}$ with only the simplex constraints, algorithms such as Mirror Descent (Beck & Teboulle, 2003) can solve it efficiently.

We note here that this MMD formulation is equivalent to Equation 6 in (Zhang et al., 2013) when applied on discrete labels and with a few other minor modifications. However, we are aware of no prior work on the theoretical analysis of this MMD estimate. The analysis of MMD for the covariance shift problem is very different (Yu & Szepesvari, 2012; Gretton et al., 2009; Cortes et al., 2008) from ours because their objectives and assumptions are different.

2.1. Theoretical Analysis

We next discuss results that show the closeness of $\widehat{\theta}(n)$ to the true $\boldsymbol{\theta}^*$ both in the finite sample case and asymptotically. We already assumed that θ^* is unique. For any formal comparison, $\widehat{\theta}(n)$ is also required to be unique. Our proof makes another mild assumption that makes the objective strongly convex:

$$\begin{array}{ll} \bar{A} \text{ has full column rank} & (\mathcal{A}4) \\ \widehat{A}(n) \text{ has full column rank} & (\mathcal{A}5) \end{array}$$

These assumptions imply the uniqueness of $\widehat{\theta}(n)$ and θ^*. For example, $\mathcal{A}5$ holds whenever all labeled and unlabeled data points are distinct and K is universal.

At this point we note that typical learning theory bounds are derived for knowing how close the objectives of (2) and (3) are, while we desire to know how close their solutions are. Also, (2) and (3) do not have a closed form solution[1]. Hence deriving finite sample closeness bounds as well as asymptotic convergence between *solutions* of (2) and (3) is interesting. To this end, we present the following theorem.

Theorem 1. *With the notation and strong convexity assumptions ($\mathcal{A}4$, $\mathcal{A}5$) presented above, the following hold:*
1. With probability at least $1-\delta$*,*

$$\|\widehat{\theta}(n)-\theta^*\|^2 \leq \frac{R^2\left(\frac{c^2+2c+2}{n_u} + \sum_{y=0}^{c}\frac{2}{n_y}\right)\left(1+\sqrt{\log\frac{2}{\delta}}\right)^2}{\mathit{mineig}(\widehat{A}(n)^\top\widehat{A}(n))} \tag{5}$$

2. $\left\{\|\widehat{\theta}(n)-\theta^*\|^2\right\} \xrightarrow{p} 0$ *where* $\mathit{mineig}(M)$ *denotes*[2] *the minimum eigen value of* M *and* R *is the data-spread given by* $R \equiv \max_{x\in\mathcal{X}}\|\Phi(\mathbf{x})\|$.

The proof has two key steps. The first is a result, given below as Lemma 1, that helps to bound the distance between the optimal objectives of (2) and (3). The second is a result, given below as Lemma 2, that bounds the closeness between the solutions of (2) and (3) in terms of closeness between their optimal objectives. We proceed by presenting the two lemmas and defer their proofs to the end of the section.

[1] Problems (2) and (3) have closed form solutions (least squares kind) only if we assume $\theta_y^* \neq 0, \widehat{\theta}_y \neq 0, \forall y$. However, in practice this is not reasonable to assume.

[2] Also, $\{X_n\} \xrightarrow{p} X$ denotes that the sequence of random variables $X_1, \ldots, X_n, \ldots$ converges in probability to X.

Lemma 1.

$$\|\widehat{A}(n)\theta^* - \widehat{a}(n)\|^2 - \|\widehat{A}(n)\widehat{\theta}(n) - \widehat{a}(n)\|^2 \qquad (6)$$
$$\le R^2\left(\frac{c^2+2c+2}{n_u} + \sum_{y=0}^{c}\frac{2}{n_y}\right)\left(1+\sqrt{\log\frac{2}{\delta}}\right)^2$$

with at least probability $1-\delta$.

Lemma 2. *The following two claims hold:*

1. $\|\widehat{\theta}(n) - \theta^*\|^2 \le \frac{\|\widehat{A}(n)\theta^* - \widehat{a}(n)\|^2 - \|\widehat{A}(n)\widehat{\theta}(n) - \widehat{a}(n)\|^2}{\textit{mineig}(\widehat{A}(n)^\top \widehat{A}(n))}$
2. *With probability* $1-\delta$,

$$\textit{mineig}(\bar{A}^\top \bar{A}) - \textit{mineig}(\widehat{A}(n)^\top \widehat{A}(n))$$
$$\le 8R^2 \sum_{y=0}^{c}\frac{1}{n_y}\sqrt{\left(\sum_{y=1}^{c}\frac{1}{n_y} + \frac{c^2}{n_0}\right)\log\frac{1}{\delta}}$$

Now from Lemma 1 and Lemma 2.1, we obtain the first result of the theorem. From Lemma 1, Lemma 2 and union bound, we obtain that with probability $1-\delta$, we have:

$$\|\widehat{\theta}(n) - \theta^*\|^2$$
$$\le \frac{R^2\left(\frac{c^2+2c+2}{n_u} + \sum_{y=0}^{c}\frac{2}{n_y}\right)\left(1+\sqrt{\log\frac{4}{\delta}}\right)^2}{\text{mineig}(\bar{A}^\top \bar{A}) - 8R^2\sum_{y=0}^{c}\frac{1}{n_y}\sqrt{\left(\sum_{y=1}^{c}\frac{1}{n_y}+\frac{c^2}{n_0}\right)\log\frac{2}{\delta}}}$$

Note that the above bound holds as long as $n \equiv (n_0,\ldots,n_c,n_u)$ is high enough such that the denominator in the above expression is positive. Choosing such n is possible because of the assumption $\mathcal{A}4$, $\text{mineig}(\bar{A}^\top \bar{A}) > 0$. From the above bound, we obtain that $\left\{\|\widehat{\theta}(n) - \theta^*\|^2\right\} \xrightarrow{p} 0$ as $n = (n_0,\ldots,n_c,n_u) \to \infty$. In the following we provide a sketch of proof for the lemmas and postpone the details to the supplementary.

Proof of Lemma 1 First note that $\|\widehat{A}(n)\theta^* - \widehat{a}(n)\|^2 - \|\widehat{A}(n)\widehat{\theta}(n) - \widehat{a}(n)\|^2 \le \|\widehat{A}(n)\theta^* - \widehat{a}(n)\|^2$. Now, we upper-bound the RHS. Let $f(X_0,\ldots,X_c,X_u) \equiv \widehat{A}(n)\theta^* - \widehat{a}(n) = \sum_{y=0}^{c}\theta^*_y\widehat{\Phi}_y(n_y) - \widehat{\Phi}_U(n_u)$, where X_y, X_u denote independent samples of size n_y, n_u from $P_D(\mathbf{x}|y)$ and $P_U(\mathbf{x})$ respectively. Note that $||f(X_o,\ldots,X_c,X_u)||$ satisfies the bounded difference property and applying McDiarmid's inequality we get that with probability at least $1-\delta$,

$$\|f\| - \mathbb{E}[\|f\|] \le R\sqrt{2\log\frac{2}{\delta}\left(\frac{1}{n_u} + \sum_{y=0}^{c}\frac{1}{n_y}\right)} \qquad (7)$$

Next, we use $\mathbb{E}(\|f\|)^2 \le \mathbb{E}(\|f\|^2)$ and the claim $\mathbb{E}(\|f\|^2) \le R^2(\frac{c^2+2c+2}{n_u} + \sum_{y=0}^{c}\frac{2}{n_y})$ to get the desired result.

Proof of Lemma 2 Let $h(\boldsymbol{\theta}) \equiv \|\widehat{A}(n)\boldsymbol{\theta} - \widehat{a}(n)\|^2$. Since h is quadratic in $\boldsymbol{\theta}$, we then have:

$$h(\theta^*) - h(\widehat{\theta}(n)) = \nabla h(\widehat{\theta}(n))^\top(\theta^* - \widehat{\theta}(n))$$
$$+(\theta^* - \widehat{\theta}(n))^\top \widehat{A}(n)^\top \widehat{A}(n)(\theta^* - \widehat{\theta}(n))$$

Moreover, $\nabla h(\widehat{\theta}(n))^\top(\boldsymbol{\theta} - \widehat{\theta}(n)) \ge 0$ for any $\boldsymbol{\theta} \in \wedge^c$. This is because, $\widehat{\theta}(n)$ is the optimal solution of 3 and hence the gradient at this point $\nabla h(\widehat{\theta}(n))$ should lie in the normal cone of the feasibility set $\wedge^c$ at $\widehat{\theta}(n)$. Hence,

$$h(\theta^*) - h(\widehat{\theta}(n)) \ge (\theta^* - \widehat{\theta}(n))^\top \widehat{A}(n)^\top \widehat{A}(n)(\theta^* - \widehat{\theta}(n))$$
$$\ge \text{mineig}(\widehat{A}(n)^\top \widehat{A}(n))\|\theta^* - \widehat{\theta}(n)\|^2$$

This, together with assumption $\mathcal{A}5$, gives Lemma 2.1. To prove Lemma 2.2, we first prove in the supplementary that[3]

$$\text{mineig}(\bar{A}^\top \bar{A}) - \text{mineig}(\widehat{A}(n)^\top \widehat{A}(n))$$
$$\le \|\bar{A}^\top \bar{A} - \widehat{A}(n)^\top \widehat{A}(n)\|_F$$

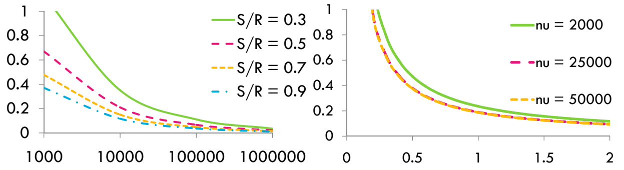
Let $g(X_0,\ldots,X_c) = \|\bar{A}^\top \bar{A} - \widehat{A}(n)^\top \widehat{A}(n)\|_F$. It is easy to verify that $\mathbb{E}g = 0$. Also, g satisfies the bounded difference property, hence by an application of McDiarmid's inequality, we get that with probability $1-\delta$

$$g \le 8R^2\sum_{y=0}^{c}\frac{1}{n_y}\sqrt{\left(\sum_{y=1}^{c}\frac{1}{n_y}+\frac{c^2}{n_0}\right)\log\frac{1}{\delta}}. \qquad (8)$$

This completes the proof for Lemma 2.

Theorem 1 is indeed interesting: first, it shows that our empirical estimate is statistically consistent. Second, it shows that the convergence rate of the squared error is at least $\mathcal{O}\left(\frac{1}{n}\right)$. We observe this graphically in Figure 1(a) as we see the error bound asymptotically going to zero with increasing training and test sizes. More importantly, in the finite regime, the theorem provides an upper-bound on the square error in terms of known and intuitive geometric quantities like $\text{mineig}(\widehat{A}(n)^\top \widehat{A}(n))$ and R. In particular, for the two-class case, this bound says that the estimate is accurate whenever the distance between the sample means of feature vectors of the two classes are far apart and the overall data spread (R) is small. We illustrate the dependence graphically in Figure 1(b), where we plot the bounds for increasing S/R where $S = \sqrt{\text{mineig}(\widehat{A}^\top \widehat{A})}$. When the positive and negative means are sufficiently separated ($S/R \ge 0.7$), the error bounds are quite tight. In the following section, we present a novel formulation for exploiting these bounds for kernel selection.

[3] Here, $\|M\|_F$ denotes the Frobenius norm of M.



(a) (b)

Figure 1. Error bound computed by Theorem 1 against increasing training and test size in log scale (Figure 1(a)), increasing separation ratio S/R (Figure 1(b)).

Comparison with other bounds We are aware of no other work that bounds the error of class-ratio estimates via any other method. For regression and under covariance shift, (Yu & Szepesvari, 2012) bounds the error of the mean y in an unlabeled set U estimated as $\sum_{(\mathbf{x},y)\in D} \hat{\beta}(\mathbf{x})y$ where $\hat{\beta}(\mathbf{x})$ is estimated using MMD for the covariance shift problem. Even though the setting is different, it is interesting to compare their convergence rates (with 0/1 values of y) with ours for two classes. Their convergence rate is $\mathcal{O}(\log^{-s} \frac{nn_u}{n+n_u})$ where $n = \sum_{y=0}^{c} n_y$ and s is a positive constant. Note that these rates are much slower than ours.

3. The Kernel Learning Formulation

Our theoretical analysis shows that the universal kernel used in the MMD estimator needs to be carefully chosen to obtain accurate class-ratios. Here, we present a novel convex formulation for learning such a kernel. We also present an efficient algorithm for solving this formulation.

We assume that we are given a set of base kernels $k_1, \ldots, k_{n_k}$ and the goal is to find the right conic combination, $k_{\mathbf{w}} = \sum_{j=1}^{n_k} w_j k_j, w_j \geq 0 \; \forall \; j$, that makes the estimated class-ratios close to the true class-ratios. Posing the problem of kernel learning as that of optimizing the kernel weights in the conic combination is a popular strategy in the kernel learning community (Lanckriet et al., 2002).

We first use the bounds in Theorem 1 to increase the accuracy of class ratio estimates: we choose the kernel weights such that the upper-bound (5) is minimized. There are two quantities in this upper-bound that depend on the kernel weights: i) $\text{mineig}(\widehat{A}(n)^\top \widehat{A}(n)) = \text{mineig}(\sum_{j=1}^{n_k} w_j \widehat{A}_j(n)^\top \widehat{A}_j(n))$, where $\widehat{A}_j(n)$ is the $\widehat{A}$ term computed using the j^{th} base kernel, and ii) $R = \|w\|_2$. The first term needs to be maximized and the second needs to be minimized. Since we wish to obtain a sparse set of kernel weights, leading to kernel selection, we minimize $\|w\|_1$ instead of $\|w\|_2$.

Our goal of reducing error in the MMD-estimate may not be adequately served by minimizing the upper bound alone because the bound is derived without making any distributional assumptions (other than $\mathcal{A}1$–$\mathcal{A}4$) and may be overly pessimistic for the given data distribution. We therefore include an empirical term in the objective that reduces the deviation between the estimated and true class-ratios over several datasets. We assume we have a set of datasets $\{U_i\}_{i=1}^{m}$ (each U_i is a set $\{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_u}\}$) with known true class-ratios $\boldsymbol{\theta}_i^*$. We discuss later one way of obtaining such a dataset from the given labeled set D. Let $\text{mmd}(U_i, \boldsymbol{\theta}, \mathbf{w})$ denote the value of the MMD objective (4) at $\boldsymbol{\theta}$ when applied to U_i with kernel $k_{\mathbf{w}}$. One way of minimizing the deviation between the estimate and the true class-ratios is by simply minimizing the deviation between the corresponding $\text{mmd}()s$. We cast this goal in the max-margin framework for structured learning (Tsochantaridis et al., 2005): we want $\mathbf{w}$ to be such that for each $(U_i, \boldsymbol{\theta}_i^*)$, $\text{mmd}(U_i, \boldsymbol{\theta}_i^*, \mathbf{w}) \leq \text{mmd}(U_i, \boldsymbol{\theta}, \mathbf{w})$ for all $\boldsymbol{\theta}$ far from $\boldsymbol{\theta}_i^*$. We rewrite $\text{mmd}(U_i, \boldsymbol{\theta}, \mathbf{w})$ as a linear function of $\mathbf{w}$ as:

$$\text{mmd}(U_i, \boldsymbol{\theta}, \mathbf{w}) = -\mathbf{w}^\top \mathbf{F}(U_i, \boldsymbol{\theta}) \;\; s.t. \tag{9}$$

$$F_j(U_i, \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \widehat{A}_j^\top \widehat{A}_j \boldsymbol{\theta} + 2\boldsymbol{\theta}^\top \widehat{A}_j^\top \widehat{a}_j(U_i) \tag{10}$$

where $\widehat{a}_j(U_i) = \widehat{\Phi}^j(U_i) - \widehat{\Phi}_0^j$ with $\widehat{\Phi}^j(U_i)$ being the mean feature map over the jth kernel calculated on sample U_i and $\widehat{\Phi}_y^j$ is also specialized to kernel k_j. Let $\text{E}(\boldsymbol{\theta}_i^*, \boldsymbol{\theta})$ be a measure of the distance between $\boldsymbol{\theta}_i^*$ and $\boldsymbol{\theta}$. We want the margin $\mathbf{w}^\top \mathbf{F}(U_i, \boldsymbol{\theta}_i^*) - \mathbf{w}^\top \mathbf{F}(U_i, \boldsymbol{\theta}) \equiv \mathbf{w}^\top \delta \mathbf{F}_i(U_i, \boldsymbol{\theta})$ to be large when $\text{E}(\boldsymbol{\theta}_i^*, \boldsymbol{\theta})$ is large.

Combining the two bound-related objectives and the empirical term we obtain the following convex formulation for learning the kernel weights

$$\min_{\mathbf{w}\in\mathbf{R}_+^{n_K}, \boldsymbol{\xi}\in\mathbf{R}_+^m} \|\mathbf{w}\|_1 + C\|\boldsymbol{\xi}\|_1 + B \;\text{maxeig}(\sum_{j=1}^{n_k} -w_j \widehat{A}_j^\top \widehat{A}_j)$$
$$\text{s.t. } \mathbf{w}^\top \delta \mathbf{F}_i(U_i, \boldsymbol{\theta}) \geq \text{E}(\boldsymbol{\theta}_i^*, \boldsymbol{\theta}) - \xi_i \;\; : \forall \|\boldsymbol{\theta} - \boldsymbol{\theta}_i^*\| \geq \epsilon \; \forall i \tag{11}$$

where $\boldsymbol{\xi} = \{\xi_i\}_{i=1}^m$ are the slack variables, C, B are the parameters of the optimization problem and $\epsilon > 0$ is a user-given tolerance. The above is an instance of a convex program with infinitely many constraints and hence we resort to solving it using the cutting-plane algorithm (Tsochantaridis et al., 2005). In Figure 2 we present an overview. The input to the algorithm is a labeled dataset D and candidate kernels $k_1, \ldots, k_{n_K}$. We create the $(U_i, \boldsymbol{\theta}_i^*)$ dataset pairs required for our training by resampling from available labeled set D. To create m datasets we repeat this process m times: first, sample a value of $\boldsymbol{\theta}_i^*$ uniformly randomly from the $c+1$ dimensional simplex. If a prior distribution on the test set $\boldsymbol{\theta}$s is known, one can use it in place of the uniform distribution. We did not assume any such priors in our experiments. Second, form U_i as follows. Let T be the expected size of the test set. For each class y, use sampling with replacement to select $T\theta_{iy}^*$ instances from

```
1: **Input:** $D, k_1, \ldots, k_{n_k}$
2: $(U_i, \boldsymbol{\theta}_i^*)$ = sampled sets from $D$ with varying ratios $\boldsymbol{\theta}_i^*$
3: $\mathbf{w}$ = Initial parameter $w_j = 1$
4: Initial constraint set $\mathcal{S} = \{\}$
5: **while** no convergence **do**
6:     $\bar{i}, \bar{\boldsymbol{\theta}} = \text{argmin}_{i,\boldsymbol{\theta}} \text{mmd}(U_i, \boldsymbol{\theta}, \mathbf{w}) - \text{E}(\boldsymbol{\theta}_i^*, \boldsymbol{\theta})$ (Sec 3.1)
7:     If $\mathbf{w}^\top \delta \mathbf{F}_{\bar{i}}(U_{\bar{i}}, \bar{\boldsymbol{\theta}}) \geq \text{E}(\boldsymbol{\theta}_{\bar{i}}^*, \bar{\boldsymbol{\theta}}) - \xi_{\bar{i}}$, then exit; else
       add $(\bar{i}, \bar{\boldsymbol{\theta}})$ to S.
8:     Solve (11) restricted to $\mathcal{S}$ and obtain $\mathbf{w}, \boldsymbol{\xi}$ (Sec 3.2).
9: **end while**
```

Figure 2. Kernel selection algorithm for the MMD estimator.

$D_y = \{(\mathbf{x}, y) \in D\}$. We now discuss the two crucial optimization problems solved within the loop of the algorithm: (i) the selection of the most violating constraint (step 6), (ii) solving the MKL objective with the finite set of constraints (step 8). We elaborate on how each is solved.

3.1. Finding the most violating constraint

For a given value of $\mathbf{w}$ we have to find the $\boldsymbol{\theta}$ corresponding to the most violating constraint for $(U_i, \boldsymbol{\theta}_i^*)$ by solving

$$\min_{\boldsymbol{\theta} \in \wedge^c} \; \text{mmd}(U_i, \boldsymbol{\theta}, \mathbf{w}) - \text{E}(\boldsymbol{\theta}_i^*, \boldsymbol{\theta}) \tag{12}$$

This optimization problem differs from our original convex objective in only the additional $\text{E}(\boldsymbol{\theta}_i^*, \boldsymbol{\theta})$ term. Since the $\boldsymbol{\theta}$ corresponds to parameters of a multinomial distribution, suitable choices for $\text{E}(\boldsymbol{\theta}_i^*, \boldsymbol{\theta})$ are the L_γ distance ($\gamma \geq 1$) and KL-divergence both of which are convex in $\boldsymbol{\theta}$. This makes objective (12) non-convex but since it is the difference of two convex functions, algorithms like CCCP (Yuille & Rangarajan, 2003) can give a local optimum. However, for two very apt error measures: the L_∞ distance and L_1 we are able to provide an optimal answer. Both of these can be expressed as a max over a small number of linear functions of $\boldsymbol{\theta}$. For example when $\text{E}(\boldsymbol{\theta}_i^*, \boldsymbol{\theta})$ is the L_∞ distance $\max_{y=0}^{c} |\theta_{iy}^* - \theta_y|$, we can rewrite it as $\max_{\boldsymbol{\lambda} \in E_\infty^{c+1}} \boldsymbol{\lambda}^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_i^*)$, where we use E_∞^{c+1} to denote all vectors with exactly one of the $c+1$ positions either +1 or -1 and zero for the rest. There are $2c+2$ such vectors. Now, we can find the most violating constraint by solving these $2c+2$ MMD-like convex objectives:

$$\min_{\boldsymbol{\lambda} \in E_\infty^{c+1}} \min_{\boldsymbol{\theta} \in \wedge^c} \; \text{mmd}(U_i, \boldsymbol{\theta}) - \boldsymbol{\lambda}^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_i^*) \tag{13}$$

Similarly, the L_1 distance can be expressed as $\max_{\boldsymbol{\lambda} \in E_1^{c+1}} \boldsymbol{\lambda}^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_i^*)$ where E_1^c consists of vectors with +1 and -1 over any of the c positions.

3.2. Solving the MKL objective with finite constraints

This section focuses on solving (11) with the constraint-set restricted to S. We begin by noting that the formulation in this case can be written as a Semi-Definite Program (SDP):

$$\begin{aligned} \min_{\mathbf{w} \in \mathbf{R}_+^{n_K}, \boldsymbol{\xi} \in \mathbf{R}_+^m, t \in \mathbf{R}} \quad & \|\mathbf{w}\|_1 + C \, \|\boldsymbol{\xi}\|_1 + B \, t \\ \text{s.t. } \mathbf{w}^\top \delta \mathbf{F}_i(U_i, \boldsymbol{\theta}) & \geq \text{E}(\boldsymbol{\theta}_i^*, \boldsymbol{\theta}) - \xi_i \; : \forall (i, \boldsymbol{\theta}) \in \mathcal{S}, \\ t\mathbb{I} + \sum_{j=1}^{n_k} & w_j \widehat{A}_j^\top \widehat{A}_j \succeq 0, \end{aligned} \tag{14}$$

where $\mathbb{I}$ is the identity matrix of size $c \times c$. As long as c and n_k are not high, standard SDP solvers like `Mosek` or `SeDuMi` can solve (14). Otherwise, we re-write the above as the following non-differentiable convex problem:

$$\min_{\mathbf{w} \in \mathbf{R}_+^{n_K}} \|\mathbf{w}\|_1 + B \, \text{maxeig}\Big(\sum_{j=1}^{n_k} -w_j \widehat{A}_j^\top \widehat{A}_j\Big)$$
$$+ C \sum_{\forall (i, \boldsymbol{\theta}) \in \mathcal{S}} \max\big(0, \text{E}(\boldsymbol{\theta}_i^*, \boldsymbol{\theta}) - \mathbf{w}^\top \delta \mathbf{F}_i(U_i, \boldsymbol{\theta})\big)$$

The above program can be solved using projected subgradient descent[4]. The sub-gradient expression for the first and third terms in the objective is easy, for the second term $\text{maxeig}(\sum_{j=1}^{n_k} -w_j \widehat{A}_j^\top \widehat{A}_j)$, the subgradient is[5] $\Big[-e_{\mathbf{w}}^\top \widehat{A}_1^\top \widehat{A}_1 e_{\mathbf{w}} \; \ldots \; - e_{\mathbf{w}}^\top \widehat{A}_{n_k}^\top \widehat{A}_{n_k} e_{\mathbf{w}}\Big]^\top$ where $e_{\mathbf{w}}$ is any eigenvector corresponding to the maximum eigenvalue of the matrix $\sum_{j=1}^{n_k} -w_j \widehat{A}_j^\top \widehat{A}_j$.

Related work on kernel selection for MMD We are aware of no other work on learning kernels in the context of MMD-based class ratio estimation. For the two-sample test problem, (Gretton et al., 2012b) proposed a kernel learning formulation that minimizes the asymptotic probability of hypothesizing two distributions as same, when they are different (Type II error) for a given bound of the Type I error. In contrast, the novelty of our formulation is that it includes both asymptotic terms and an empirical term that minimizes error under finite data settings.

4. Experiments

We compare our proposed estimator with several existing methods under various settings. First, we compare different methods under varying true class ratios. Second, we compare them under varying training sizes. Finally, we compare different kernel selection methods.

Our chosen kernel family $\mathcal{K}_\mathcal{F}$ is a conic combination of univariate and multivariate Gaussian (RBF) kernels, a popular family in various kernel learning literature (Gretton et al., 2012b; Vishwanathan et al., 2010). We first fix a kernel width (σ) based on training data as per (Zhang et al.,

[4] The feasibility set in this case is the first orthant and hence it is simple to compute the projection onto this.

[5] Please refer to the supplementary for the derivation.

Dataset	Number Features	Number Instances	$\lvert\mathcal{Y}\rvert$ c+1	n_y	n_u
Australian	14	690	2	200	100
Diabetes	8	768	2	200	100
German	24	1000	2	200	100
Ionosphere	34	351	2	200	100
SAHeart	9	462	2	200	100
Youtube	1000	6,431,471	2	250	1000
Acoustic	50	78823	3	varying	10000
Botswana	145	3248	14	varying	10000
Shuttle	9	43500	7	varying	10000

Table 1. Summary of Datasets

2013). For each feature, we have one univariate kernel with this width. We create several multivariate kernels based on bandwidths from this set: $[2^{-6}\ 2^{-5}\ \ldots\ 2^{6}] * \sigma * d^2$ where d is the number of features.

Methods: We compare the following methods.

SMO-MKL**:** As a first baseline, we estimate class ratio by aggregating from per-instance predictions from a classifier. The classifier we used was SMO-MKL[6] (Vishwanathan et al., 2010) which trains a SVM but with the benefit of kernel selection from our kernel family $\mathcal{K}_{\mathcal{F}}$.

PE-DR**:** As a member of the direct method, we use the PE divergence based class ratio estimation method of (Plessis & Sugiyama, 2012). We thank the authors for providing us with the Matlab code for this method[7]. We exclude results for the method in (Saerens et al., 2002), since it offered no benefit over the baseline SMO-MKL.

MMD: This is the MMD based approach (Section 2) with a single best kernel chosen from our kernel family $\mathcal{K}_{\mathcal{F}}$ through cross-validation. Recall that (Zhang et al., 2013)'s proposal is also a MMD method — the only difference is in how kernel parameters are chosen. We found that choosing a single kernel via cross-validation provided much higher accuracy than their formula of kernel width selection.

MMD-MKL**:** Here, we used MMD on a kernel learned as in Section 3. The datasets i.e. $\{(U_i, \boldsymbol{\theta}_i^*)\}$ pairs required for this training were sampled from the training data D using the method of Section 3 with $m = 110$. The class means $\widehat{\Phi}_y$ were estimated from the entire labeled data. The parameters C and B were fixed via cross-validation.

Datasets: Table 1 summarizes the datasets we used. The first six are binary datasets comprising five of the six UCI datasets used in (Plessis & Sugiyama, 2012) and a dataset based on YouTube comments that we created based on this[8] collection. The goal in the YouTube dataset is to estimate the fraction of comments that are spams on a YouTube video. The dataset was crawled by tracking 6407 popular YouTube videos over 77 days and comprises of 6,431,471 comments labeled spam or not. The feature set is a normalized TF-IDF vector over 1000 words + a comment length feature. The next three are multi-class datasets. Acoustic is a dataset about classifying military vehicles from geophone recordings and is used in (Plessis & Sugiyama, 2012). Botswana is a dataset about classifying spectral signatures into different land cover types and is from (Zhang et al., 2013) and Shuttle is a UCI dataset.

We created a training set by sampling n_y samples from each class y, and series of test sets with n_u points each for a given ratio of classes $\boldsymbol{\theta}^*$. The default values of n_y, n_u are in Table 1. For the binary datasets, all experiments are with varying $\boldsymbol{\theta}^*$ and for the multi-class datasets, the default $\boldsymbol{\theta}^*$ is the class prior skew in the entire labeled data. All numbers are averaged over 10 random seeds. We measure error as the L_1 distance between the true and estimated class ratios normalized by the number of classes.

Varying class ratios: We perform these experiments on the six binary datasets by varying θ_0^* as per the set $\{0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.99\}$. In Figure 4, we plot the estimation error ($|\theta_0^* - \widehat{\theta}_0|$) against true fractions (θ_0^*) for the six binary datasets. The following conclusions can be drawn from the plots: SMO-MKL, the baseline that aggregates per-instance predictions, is indeed very sensitive to the changes in test prior distribution. For many datasets, we see a "bowl" shaped graph and usually the minimum is when test prior is close to 0.5 which is equal to the training prior. The curves for the direct methods (PE-DR, MMD, MMD-MKL) are much flatter showing that they are much less sensitive to the training class ratios. Except for YouTube, both the MMD-based methods provide lower error than PE-DR. MMD-MKL is more accurate than MMD in most cases, and on YouTube we get upto a 33% drop in error.

Running time: In this paper we skip a detailed comparison on running time, instead we report some specific timings: On YouTube, our largest dataset, MKL training via Matlab takes 20 minutes whereas deployment takes 5 minutes on a desktop class machine. On Botswana, the dataset with 14 classes, MKL training takes 6 minutes whereas deployment takes 0.3 minutes. In comparison PE-DR took 12 minutes on YouTube and 5 minutes on Botswana.

Increasing training size: For these experiments we vary the value of n_y (number of instances per class) in the range [10 30 50 70 90] keeping all other values fixed to their default in Table 1. We selected the three multi-class datasets for these experiments. We observe smooth error reduction in MMD-MKL with increasing training size and consistent

[6] Code taken from http://research.microsoft.com/en-us/um/people/manik/code/SMO-MKL/download.html

[7] http://sugiyama-www.cs.titech.ac.jp/∼christo/classpriorchange.html

[8] http://mlg.ucd.ie/yt

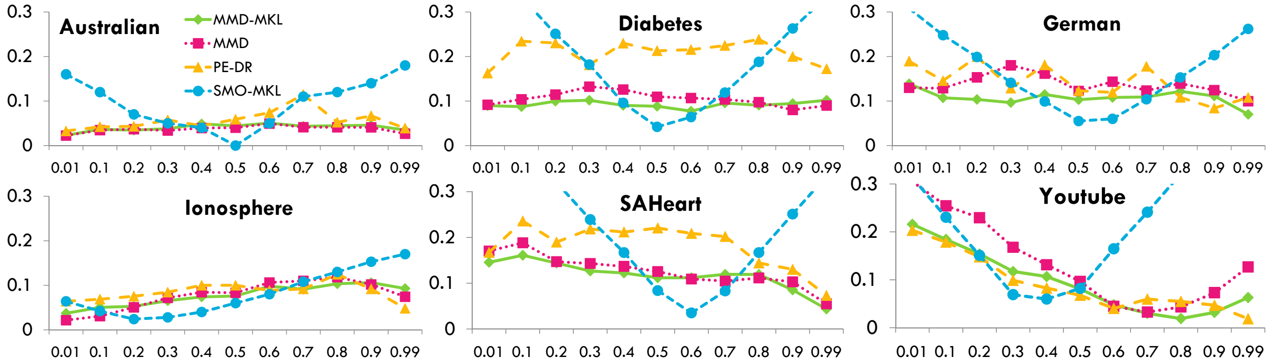


Figure 3. Class ratio estimation error ($|\theta_0^* - \widehat{\theta}_0|$) on Y-axis against varying true fractions (θ_0^*) for the six binary datasets of Table 1. The methods compared are same in all six datasets; the legend is present in only one of them to reduce clutter.

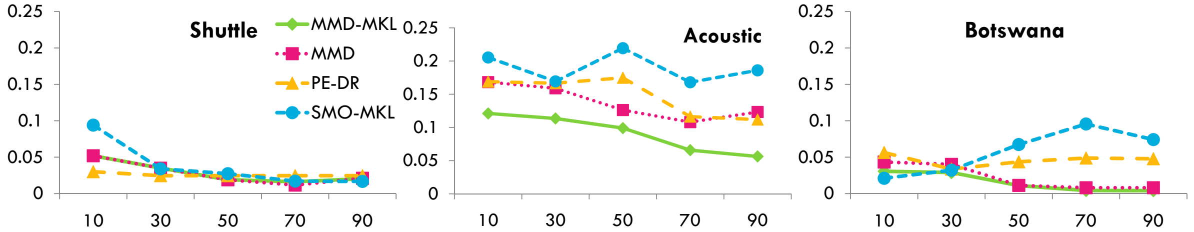


Figure 4. Class ratio estimation error ($|\theta^* - \widehat{\theta}|/c + 1$) on Y-axis against increasing per-class training size n_y for the three multi-class datasets of Table 1. The methods compared are the same in all datasets; the legend is present in only one of them to reduce clutter.

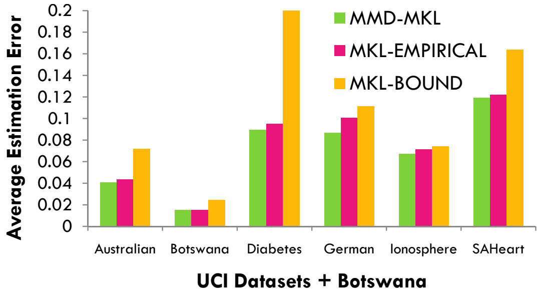


Figure 5. Comparing the kernel selection methods MMD-MKL, MKL-BOUND and MKL-EMPIRICAL on various datasets

improvement over other methods. In contrast, SMO-MKL has inconsistent behavior. The best improvement we get is for the Acoustic dataset where MMD-MKL reduces error from 0.12 to 0.05 with 90 instances per class.

Comparison of Kernel Selection: Our MKL attempts to jointly minimize the theoretical error bounds and empirical error. We evaluate if any one of them would be adequate by comparing our joint model (MMD-MKL) with (1) MKL-BOUND that minimizes only the Eigen and $\|\mathbf{w}\|_1$ terms in (11), and (2) MKL-EMPIRICAL that drops the Eigen term. In Figure 5 we plot error of these methods averaged over various class ratios. We observe that the joint model provides the highest overall accuracy.

5. Conclusion

In this paper we address a real-world motivated problem of estimating the ratio of classes in an unlabeled set. We investigated the use of the maximum mean discrepancy (MMD) measure as a basis for estimating the class ratios. We present the first ever theoretical analysis of the estimator and show that the MMD estimator is consistent under mild conditions. We provide empirical error bounds in terms of intuitive quantities like class-separation and data-spread. Combining these bounds and empirical error we propose a novel convex formulation for kernel learning and also design an efficient cutting plane algorithm for solving it. We empirically compare our estimator with many existing methods and obtain up to 60% reduction in error over the best existing method. Further, our method of kernel learning reduces plain MMD error by up to 40%.

As part of future work, we wish to explore other families of kernel selection, for example directly optimizing the width of the RBF kernel as in (Gehler & Nowozin, 2008; Argyriou et al., 2006).

Acknowledgments

We thank the reviewers for their useful comments. This work was partly supported by research grants from Yahoo! Research.

References

Argyriou, Andreas, Hauser, Raphael, Micchelli, Charles A., and Pontil, Massimiliano. A DC-programming algorithm for kernel selection. In *ICML*, pp. 41–48, 2006.

Beck, Amir and Teboulle, Marc. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167 – 175, 2003.

Clémençon, Stéphan, Vayatis, Nicolas, and Depecker, Marine. AUC optimization and the two-sample problem. In *NIPS*, 2009.

Cortes, Corinna and Mohri, Mehryar. AUC optimization vs. error rate minimization. In *NIPS*, 2004.

Cortes, Corinna, Mohri, Mehryar, Riley, Michael, and Rostamizadeh, Afshin. Sample selection bias correction theory. In *ALT*, 2008.

Elkan, Charles. The foundations of cost-sensitive learning. In *IJCAI*, 2001.

Gehler, Peter Vincent and Nowozin, Sebastian. Infinite kernel learning. Technical report, Max Planck Institute for Biological Cybernetics, 2008.

Gretton, Arthur, Smola, Alexander J., Huang, Jiayuan, Schmittfull, Marcel, Borgwardt, Karsten M., and Schölkopf, Bernhard. *Covariate shift and local learning by distribution matching*, pp. 131–160. MIT Press, Cambridge, MA, USA, 2 2009.

Gretton, Arthur, Borgwardt, Karsten M., Rasch, Malte J., Schölkopf, Bernhard, and Smola, Alexander J. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012a.

Gretton, Arthur, Sriperumbudur, Bharath K., Sejdinovic, Dino, Strathmann, Heiko, Balakrishnan, Sivaraman, Pontil, Massimiliano, and Fukumizu, Kenji. Optimal kernel choice for large-scale two-sample tests. In *NIPS*, 2012b.

Lanckriet, Gert, Cristianini, Nello, Bartlett, Peter, and Ghaoui, Laurent El. Learning the kernel matrix with semi-definite programming. *JMLR*, 2002.

Lin, Yi, Lee, Yoonkyung, and Wahba, Grace. Support vector machines for classification in nonstandard situations. *Machine Learning Journal*, 2002.

Plessis, Marthinus D. and Sugiyama, Masashi. Semi-supervised learning of class balance under class-prior change by distribution matching. In *ICML*, 2012.

Saerens, Marco, Latinne, Patrice, and Decaestecker, Christine. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural Computation*, 2002.

Selvaraj, Sathiya Keerthi, Bhar, Bigyan, Sellamanickam, Sundararajan, and Shevade, Shirish. Semi-supervised SVMs for classification with unknown class proportions and a small labeled dataset. In *CIKM*, 2011.

Sriperumbudur, Bharath K., Fukumizu, Kenji, Gretton, Arthur, Lanckriet, Gert R. G., and Schölkopf, Bernhard. Kernel choice and classifiability for RKHS embeddings of probability distributions. In *NIPS*, pp. 1750–1758, 2009.

Titterington, D. M. Minimum distance non-parametric estimation of mixture proportions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(1):pp. 37–46, 1983.

Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. Large margin methods for structured and interdependent output variables. *JMLR*, 6:1453–1484, 2005.

Vishwanathan, S. V. N., Sun, Zhaonan, Theera-Ampornpunt, Nawanol, and Varma, Manik. Multiple kernel learning and the SMO algorithm. In *Advances in Neural Information Processing Systems*, December 2010.

Woodward, Wayne A., Parr, William C., Schucany, William R., and Lindsey, Hildegard. A comparison of minimum distance and maximum likelihood estimation of a mixture proportion. *Journal of the American Statistical Association*, 79(387):590–598, 1984.

Yu, Yaoliang and Szepesvari, Csaba. Analysis of kernel mean matching under covariate shift. In *ICML*, 2012.

Yuille, A. L. and Rangarajan, Anand. The concave-convex procedure. *Neural Computation*, 15(4):915936, 2003.

Zhang, Kun, Schölkopf, Bernhard, Muandet, Krikamol, and Wang, Zhikun. Domain adaptation under target and conditional shift. In *ICML*, 2013.