Semi-Supervised Learning of Class Balance under Class-Prior Change by Distribution Matching*

Marthinus Christoffel du Plessis
Tokyo Institute of Technology, Japan.
christo@sg.cs.titech.ac.jp
Masashi Sugiyama
Tokyo Institute of Technology, Japan.
sugi@cs.titech.ac.jp http://sugiyama-www.cs.titech.ac.jp/~sugi

Abstract

In real-world classification problems, the class balance in the training dataset does not necessarily reflect that of the test dataset, which can cause significant estimation bias. If the class ratio of the test dataset is known, instance re-weighting or resampling allows systematical bias correction. However, learning the class ratio of the test dataset is challenging when no labeled data is available from the test domain. In this paper, we propose to estimate the class ratio in the test dataset by matching probability distributions of training and test input data. We demonstrate the utility of the proposed approach through experiments.

Keywords

Class-prior change, density ratio, f-divergence, selection bias.

1 Introduction

Most supervised learning algorithms assume that training and test data follow the same probability distribution (Vapnik, 1998; Hastie et al., 2001; Bishop, 2006). However, this de facto standard assumption is often violated in real-world problems, caused by intrinsic sample selection bias or inevitable non-stationarity (Heckman, 1979; Quiñonero-Candela et al., 2009; Sugiyama and Kawanabe, 2012).

In classification scenarios, changes in class balance are often observed—for example, the male-female ratio is almost fifty-fifty in the real-world (test set), whereas training samples collected in a research laboratory tends to be dominated by male data. Such a

*This paper is an extended version of an earlier conference paper (du Plessis and Sugiyama, 2012).

situation is called a *class-prior change*, and the bias caused by differing class balances can be systematically adjusted by instance re-weighting or resampling if the class balance in the test dataset is known (Elkan, 2001; Lin et al., 2002).

However, the class ratio in the test dataset is often unknown in practice. A possible approach to mitigating this problem is to learn a classifier so that the performance for all possible class balances are improved, e.g., through maximization of the area under the ROC curve (Cortes and Mohri, 2004; Clémençon et al., 2009). Alternatively, in the minimax approach, a classifier is learned so as to minimize the worst-case performance for any change in the class prior (Duda et al., 2001; Van Trees, 1968). The disadvantage of the minimax approach is that it is often overly pessimistic. A more direct approach is to estimate the class ratio in the test dataset and use this estimate for instance reweighting or resampling. We focus on this scenario under a semi-supervised learning setup (Chapelle et al., 2006), where no labeled data is available from the test domain.

Saerens et al. (2001) is a seminal paper on this topic, which proposed to estimate the class ratio by the expectation-maximization (EM) algorithm (Dempster et al., 1977)—alternately updating the test class-prior and class-posterior probabilities from some initial estimates until convergence. This method has been successfully applied to various real-world problems such as word sense disambiguation (Chan and Ng, 2006) and remote sensing (Latinne et al., 2001).

In this paper, we first reformulate the algorithm in Saerens et al. (2001), and show that this actually corresponds to approximating the test input distribution by a linear combination of class-wise training input distributions under the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951). In this procedure, the class-wise input distributions are approximated via class-posterior estimation, for example, by kernel logistic regression (Hastie et al., 2001) or its squared-loss variant (Sugiyama, 2010).

Since indirectly estimating the divergence by estimating the individual class-posterior distributions may not be the best scheme, the above reformulation motivates us to develop a more direct approach: matching the mixture of class-wise training input densities to the test input distribution. Historically, non-parametric estimation of the mixing proportions by matching the empirical distribution functions was investigated in Hall (1981), and its variant based on kernel density estimation has been developed in Titterington (1983). However, these classical approaches do not perform well in high-dimensional problems (Sugiyama et al., 2013). Recently, KL-divergence estimation based on *direct density-ratio estimation* has been shown to be promising (Nguyen et al., 2010; Sugiyama et al., 2008). Furthermore, a squared-loss variant of the KL divergence called the Pearson (PE) divergence (Pearson, 1900) can also be approximated in the same way, with an analytic solution that can be computed efficiently (Kanamori et al., 2009). Note that the PE-divergence and the KL divergence both belong to the f-divergence class (Ali and Silvey, 1966; Csiszár, 1967), which share similar properties. In this paper, with the aid of this density-ratio based PE-divergence estimator, we propose a new semi-supervised method for estimating the class ratio in the test dataset. Through experiments, we demonstrate the usefulness of the proposed method.

2 Problem Formulation and Existing Method

In this section, we formulate the problem of semi-supervised class-prior estimation and review an existing method (Saerens et al., 2001).

2.1 Problem Formulation

Let $\boldsymbol{x} \in \mathbb{R}^d$ be the d-dimensional input data, $y \in \{1, \ldots, c\}$ be the class label, and c be the number of classes. We consider class-prior change, i.e., the class-prior probability for training data $p(y)$ and that for test data $p'(y)$ are different. However, we assume that the class-conditional density for training data $p(\boldsymbol{x}|y)$ and that for test data $p'(\boldsymbol{x}|y)$ are the same:

$$p(\boldsymbol{x}|y) = p'(\boldsymbol{x}|y). \tag{1}$$

Note that training and test joint densities $p(\boldsymbol{x}, y)$ and $p'(\boldsymbol{x}, y)$ as well as training and test input densities $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$ are generally different under this setup.

For the purposes of classification, we are generally interested in selecting a classifier that minimizes the expected loss (or the risk) with respect to the test distribution. We can rewrite the expected loss in terms of the training class-conditional density, $p(\boldsymbol{x}|y)$, as

$$\begin{aligned} R &= \sum_y \int L(f(\boldsymbol{x}), y) p'(\boldsymbol{x}, y) \mathrm{d}\boldsymbol{x} \\ &= \sum_y \int L(f(\boldsymbol{x}), y) p(\boldsymbol{x}|y) p'(y) \mathrm{d}\boldsymbol{x}, \end{aligned} \tag{2}$$

where $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is the loss function. Thus, if an estimate of the test class-priors is known, the expected loss can be calculated from the training class-conditional densities. The goal of this paper is to estimate $p'(y)$ from labeled training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ drawn independently from $p(\boldsymbol{x}, y)$ and unlabeled test samples $\{\boldsymbol{x}'_i\}_{i=1}^{n'}$ drawn independently from $p'(\boldsymbol{x})$[1]. Given test labels $\{y'_i\}_{i=1}^{n'}$, $p'(y)$ can be naively estimated by n'_y/n', where n'_y is the number of test samples in class y. Here, however, we would like to estimate $p'(y)$ *without* $\{y'_i\}_{i=1}^{n'}$.

2.2 Existing Method

We give a brief overview of an existing method for semi-supervised class-prior estimation (Saerens et al., 2001), which is based on the expectation-maximization (EM) algorithm (Dempster et al., 1977).

[1] As we can confirm later, our proposed method does not actually require the independence assumption on $\{y_i\}_{i=1}^{n}$, but is valid for *deterministic* $\{y_i\}_{i=1}^{n}$ as long as $\boldsymbol{x}_i$ $(i = 1, \ldots, n)$ is drawn independently from $p(\boldsymbol{x}|y = y_i)$. However, for being consistent with other methods, we assume the independence condition here.

In the algorithm, test class-prior and class-posterior estimates $\widehat{p}'(y)$ and $\widehat{p}'(y|\boldsymbol{x})$ are iteratively updated as follows:

1. Obtain an estimate of the training class-posterior probability, $\widehat{p}(y|\boldsymbol{x})$, from training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, for example, by kernel logistic regression (Hastie et al., 2001) or its squared-loss variant (Sugiyama, 2010).

2. Obtain an estimate of the training class-prior probability, $\widehat{p}(y)$, from the labeled training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ as $\widehat{p}(y) = n_y/n$, where n_y is the number of training samples in class y. Set the initial estimate of the test class-prior probability equal to it: $\widehat{p}'_0(y) = \widehat{p}(y)$.

3. Repeat until convergence: $t = 1, 2, \ldots$

   (a) Compute a new test class-posterior estimate $\widehat{p}'_t(y|\boldsymbol{x})$ based on the current test class-prior estimate $\widehat{p}'_{t-1}(y)$ as

   $$\widehat{p}'_t(y|\boldsymbol{x}) = \frac{\widehat{p}'_{t-1}(y)\widehat{p}(y|\boldsymbol{x})/\widehat{p}(y)}{\sum_{y'=1}^{c} \widehat{p}'_{t-1}(y')\widehat{p}(y'|\boldsymbol{x})/\widehat{p}(y')}. \tag{3}$$

   (b) Compute a new test class-prior estimate $\widehat{p}'_t(y)$ based on the current test class-posterior estimate $\widehat{p}'_t(y|\boldsymbol{x})$ as

   $$\widehat{p}'_t(y) = \frac{1}{n'} \sum_{i=1}^{n'} \widehat{p}'_t(y|\boldsymbol{x}'_i). \tag{4}$$

Note that Eq.(3) comes from the Bayes formulae,

$$p(\boldsymbol{x}|y) = \frac{p(y|\boldsymbol{x})p(\boldsymbol{x})}{p(y)} \quad \text{and} \quad p'(\boldsymbol{x}|y) = \frac{p'(y|\boldsymbol{x})p'(\boldsymbol{x})}{p'(y)},$$

combined with Eq.(1):

$$p'(y|\boldsymbol{x}) \propto \frac{p'(y)}{p(y)} p(y|\boldsymbol{x}).$$

Eq.(4) comes from empirical marginalization of

$$p'(y) = \int p'(y|\boldsymbol{x})p'(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

It was suggested that this procedure may converge to a local optimal solution (Saerens et al., 2001). In the following section, we will show that the objective function is actually convex, but that the method suggested in Saerens et al. (2001) may fail to converge to the unique optimal value.

3 Reformulation of the EM Algorithm as Distribution Matching

In this section, we show that the class priors can be estimated by matching the test input density to a linear combination of class-wise training input distributions under the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951). We show that the existing EM method performs this matching via an estimation of the class posterior. Furthermore, we show that this results in a convex problem, but that the existing EM method may not obtain the optimal result.

3.1 Class-Prior Estimation as Distribution Matching

Based on the assumption that the class-conditional densities for training and test data are unchanged (see Eq.(1)), let us model the test input density $p'(\boldsymbol{x})$ by

$$q'(\boldsymbol{x}) = \sum_{y=1}^{c} \theta_y p(\boldsymbol{x}|y), \tag{5}$$

where θ_y is a coefficient corresponding to $p'(y)$:

$$\sum_{y=1}^{c} \theta_y = 1. \tag{6}$$

We match the model $q'(\boldsymbol{x})$ with the test input density $p'(\boldsymbol{x})$ under the KL divergence:

$$\begin{aligned} \mathrm{KL}(q'\|p') &:= \int p'(\boldsymbol{x}) \log \frac{p'(\boldsymbol{x})}{q'(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}, \\ &= \int p'(\boldsymbol{x}) \log p'(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int p'(\boldsymbol{x}) \log \left(\sum_{y=1}^{c} \theta_y p(\boldsymbol{x}|y) \right) \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{7}$$

We wish to select the class prior, under the constraint Eq.(6), that minimizes this Kullback-Leibler divergence.

3.2 Equivalence of the EM Method to Divergence Matching

When the KL divergence is minimized in Eq.(7), we can omit the term that is constant with respect to the class prior. This results in an optimization problem of

$$\begin{aligned}
\underset{\{\theta_y\}_{y=1}^c}{\operatorname{argmin}} \mathrm{KL}(q' \| p') &= \underset{\{\theta_y\}_{y=1}^c}{\operatorname{argmax}} \int p'(\boldsymbol{x}) \log \left(\sum_{y=1}^{c} \theta_y p(\boldsymbol{x}|y) \right) \mathrm{d}\boldsymbol{x}, \\
&= \underset{\{\theta_y\}_{y=1}^c}{\operatorname{argmax}} \int p'(\boldsymbol{x}) \log \left(p(\boldsymbol{x}) \sum_{y=1}^{c} \theta_y \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x}) p(y)} \right) \mathrm{d}\boldsymbol{x}, \\
&= \underset{\{\theta_y\}_{y=1}^c}{\operatorname{argmax}} \int p'(\boldsymbol{x}) \log p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int p'(\boldsymbol{x}) \log \left(\sum_{y=1}^{c} \theta_y \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x}) p(y)} \right) \mathrm{d}\boldsymbol{x}, \\
&= \underset{\{\theta_y\}_{y=1}^c}{\operatorname{argmax}} \int p'(\boldsymbol{x}) \log \left(\sum_{y=1}^{c} \theta_y \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x}) p(y)} \right) \mathrm{d}\boldsymbol{x}.
\end{aligned}$$

Approximating the expectation with its empirical average gives the following optimization problem:

$$\max_{\{\theta_y\}} \frac{1}{n'} \sum_{i=1}^{n'} \log \left(\sum_{y=1}^{c} \theta_y p(y|\boldsymbol{x}'_i)/p(y) \right), \tag{8}$$

subject to Eq.(6).

The above can be viewed as a convex problem since the concave log function is maximized and the constraints in Eq.(6) is linear. Therefore, the optimal points must satisfy the Karush-Kuhn-Tucker (KKT) conditions (Boyd and Vandenberghe, 2004). The KKT conditions for the above problem is given by Eq.(6) and

$$\frac{1}{n'} \sum_{i=1}^{n'} \frac{p(y|\boldsymbol{x}'_i)/p(y)}{\sum_{y'=1}^{c} \theta_{y'} p(y'|\boldsymbol{x}'_i)/p(y')} = \nu, \quad \forall y = 1, \ldots, c, \tag{9}$$

where ν is a Lagrange multiplier. From these equations, we can determine ν as

$$\begin{aligned}
\nu = 1 \cdot \nu &= \left(\sum_{y=1}^{c} \theta_y \right) \cdot \left(\frac{1}{n'} \sum_{i=1}^{n'} \frac{p(y|\boldsymbol{x}'_i)/p(y)}{\sum_{y'=1}^{c} \theta_{y'} p(y'|\boldsymbol{x}'_i)/p(y')} \right) \\
&= \frac{1}{n'} \sum_{i=1}^{n'} \frac{\sum_{y=1}^{c} \theta_y p(y|\boldsymbol{x}'_i)/p(y)}{\sum_{y'=1}^{c} \theta_{y'} p(y'|\boldsymbol{x}'_i)/p(y')} = 1.
\end{aligned}$$

Then the solution $\{\theta_y\}_{y=1}^c$ can be calculated by fixed-point iteration as follows (McLachlan and Krishnan, 1997):

$$\theta_y \longleftarrow \theta_y \left(\frac{1}{n'} \sum_{i=1}^{n'} \frac{p(y|\boldsymbol{x}'_i)/p(y)}{\sum_{y'=1}^{c} \theta_{y'} p(y'|\boldsymbol{x}'_i)/p(y')} \right). \tag{10}$$

By using an estimator of the class-posterior, $\widehat{p}(y|\boldsymbol{x})$, in the above expression, we obtain an estimator for the test class-prior $\widehat{p}'(y)$. The above is actually the same as Eq.(4) with Eq.(3) substituted.

3.3 Fixed-Point Iteration

The unknown class-priors can therefore be obtained as the solution to the non-linear equation given by Eq.(9). A simply way to construct a solution to a non-linear equation is via a fixed-point iteration (as in Eq.(10)). For conciseness, we rewrite the fixed-point iteration as a mapping $T : \mathbb{R}^c \to \mathbb{R}^c$:

$$[T(\boldsymbol{\theta})]_y = \frac{1}{n'} \sum_{i=1}^{n'} \frac{\theta_y p(y|\boldsymbol{x}'_i)/p(y)}{\sum_{y'=1}^{c} \theta_{y'} p(y'|\boldsymbol{x}'_i)/p(y')}, \tag{11}$$

where $\boldsymbol{\theta} = [\theta_1 \ \theta_2 \ \cdots \ \theta_c]^\top$ and $[\quad]_y$ denotes the yth component of a vector. The solution is then iteratively calculated as

$$\boldsymbol{\theta} \leftarrow T(\boldsymbol{\theta}),$$

until a fixed point $\boldsymbol{\theta} = T(\boldsymbol{\theta})$ is reached. Since the problem is convex, we would expect that there is a single unique fixed point. The *Banach fixed-point theorem* (also known as the *contraction mapping theorem*) (Hunter and Nachtergaele, 2001, p.62) guarantees a unique solution if T is a contraction mapping. T is a contraction mapping if

$$d\left(T(\boldsymbol{\theta}^j), T(\boldsymbol{\theta}^k)\right) < d\left(\boldsymbol{\theta}^j, \boldsymbol{\theta}^k\right), \quad \forall\, \boldsymbol{\theta}^j, \boldsymbol{\theta}^k \in \mathbb{R}^c, \tag{12}$$

where $d : \mathbb{R}^c \times \mathbb{R}^c \to \mathbb{R}$ is a metric.

However, we can actually show that Eq.(11) is *not* a contraction mapping. To explain this, we consider the counter example with vectors defined as

$$\left[\boldsymbol{\theta}^j\right]_i = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases}$$

where $1 \leq j \leq c$. By substituting this into Eq.(11), we obtain

$$T(\boldsymbol{\theta}^j) = \boldsymbol{\theta}^j, \qquad \forall j = 1, \ldots, c.$$

Therefore, for any two vectors $\boldsymbol{\theta}^j$ and $\boldsymbol{\theta}^k$, $j, k \in 1 \ldots c$, selected as above, $d\left(T(\boldsymbol{\theta}^j), T(\boldsymbol{\theta}^k)\right) = d\left(\boldsymbol{\theta}^j, \boldsymbol{\theta}^k\right)$. The condition in Eq.(12) is therefore violated, which means that T is not a contraction mapping. From this example, it is also immediately obvious that any $\boldsymbol{\theta}^j$ is a fixed point, but not necessarily the optimal value.

As shown above, the method of Saerens et al. (2001) can be regarded as solving a convex problem via fixed point iteration, but it may not result in the unique optimal value. These spurious optimal values is not a characteristic of the problem itself (which is convex), but due to solving the KKT conditions with a fixed-point iteration.

Spurious fixed points may be avoided by using several different initial values and then selecting the optimal value according to Eq.(8). Alternatively, the objective function Eq.(8) can be directly solved, e.g. through gradient descent and projection (Boyd and Vandenberghe, 2004). However, indirectly estimating the KL divergence via class-posterior estimation may not be the best scheme in practice.

4 Class-Prior Estimation by Direct Divergence Minimization

The analysis in the previous section motivates us to explore a more direct way to learn coefficients $\{\theta_y\}_{y=1}^c$. That is, given an estimator of a divergence from p' to q', coefficients $\{\theta_y\}_{y=1}^c$ are learned so that the divergence is minimized.

In this section, we first review a general framework of approximating the *f-divergences* (Ali and Silvey, 1966; Csiszár, 1967) via *Legendre-Fenchel convex duality* (Keziou, 2003; Nguyen et al., 2010). Then we review two specific methods of divergence estimation for the KL divergence and the Pearson (PE) divergence (Pearson, 1900). Finally, we propose to use the PE-divergence estimator for determining the coefficients $\{\theta_y\}_{y=1}^c$.

4.1 Framework of f-Divergence Approximation

An f-divergence (Ali and Silvey, 1966; Csiszár, 1967) from p to q' is a general divergence measure defined as

$$D_f(p'\|q') := \int q'(\boldsymbol{x}) f\left(\frac{p'(\boldsymbol{x})}{q'(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x}, \tag{13}$$

where f is a convex function such that $f(1) = 0$.

The naive way to approximate the f-divergence is to separately estimate the densities $p'(\boldsymbol{x})$ and $q'(\boldsymbol{x})$ and then plug these estimators into Eq.(13). Alternatively, the divergence can be approximated by estimating the class posterior (as was done in the method of Saerens et al. (2001)). Computing the divergence this way is however disadvantageous since this is a two-step approach. The first step (estimation of the densities or the posterior) is done without regard to the second step (estimating the divergence) and a small estimation error in the first step may cause a big error in the second stage.

Therefore we will review in this section a method to estimate the f-divergence in a single shot approach via convex duality (Keziou, 2003; Nguyen et al., 2010). The following inequality is trivially obvious

$$tv - f(t) \leq \sup_{u \in \mathrm{dom}\, f} uv - f(u), \tag{14}$$

where $\mathrm{dom}\, f$ denotes the domain of the function f and $t \in \mathrm{dom}\ f$. The value on the right-hand side is known as the *Legendre-Fenchel dual* or *convex conjugate* (Rockafellar, 1970; Boyd and Vandenberghe, 2004), denoted as

$$f^*(v) := \sup_{u \in \operatorname{dom} f} uv - f(u). \tag{15}$$

By re-arranging Eq.(14), we may obtain the following useful inequality,

$$f(t) \geq tv - f^*(v),$$

which provides a lower bound that is linear with respect to t. If the function f is convex, then there is $v \in \operatorname{dom} f^*$ for which the above bound will be met with equality (since the conjugate of the conjugate is the function itself) (Boyd and Vandenberghe, 2004, p.94). We can apply this inequality in a pointwise manner to obtain,

$$\begin{aligned} f\left(\frac{p'(\boldsymbol{x})}{q'(\boldsymbol{x})}\right) &\geq \frac{p'(\boldsymbol{x})}{q'(\boldsymbol{x})} r(\boldsymbol{x}) - f^*(r(\boldsymbol{x})), \\ q'(\boldsymbol{x}) f\left(\frac{p'(\boldsymbol{x})}{q'(\boldsymbol{x})}\right) &\geq p'(\boldsymbol{x}) r(\boldsymbol{x}) - q'(\boldsymbol{x}) f^*(r(\boldsymbol{x})), \end{aligned}$$

where $r(\boldsymbol{x})$ plays the role of v. Integrating this and selecting r so as to maximize the lower bound gives the following estimator for Eq.(13) :

$$D_f(p' \| q') \geq \max_r \int p'(\boldsymbol{x}) r(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int q'(\boldsymbol{x}) f^*(r(\boldsymbol{x})) \mathrm{d}\boldsymbol{x}. \tag{16}$$

The above is a useful expression because the right-hand side only contains expectations of r and $f^*(r(\boldsymbol{x}))$, which can be approximated by sample averages. For a continuous f, the maximum is attained for a function r such that $p'(x)/q'(x) = \partial f^*(r(x))$ (where ∂f^* is the derivative of f^*) (Nguyen et al., 2010). Therefore, in contrast to the plug-in approach, the f divergence is directly estimated in terms of the density ratio. This is intuitively advantageous since the estimation of densities is a more general problem than the estimation of a density ratio (Sugiyama et al., 2012a).

Below, we show specific methods of divergence approximation for the KL and PE divergences under the model (5) and the following parametric expression of the density ratio $r(\boldsymbol{x})$:

$$r(\boldsymbol{x}) = \sum_{\ell=0}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}), \tag{17}$$

where $\{\alpha_\ell\}_{\ell=0}^{b}$ are parameters and $\{\varphi_\ell(\boldsymbol{x})\}_{\ell=0}^{b}$ are basis functions. In practice, we use a constant basis and Gaussian kernels centered at the training data points, i.e., for $b = n$ and $\ell = 1, 2, \ldots, n$,

$$\varphi_0(\boldsymbol{x}) = 1 \quad \text{and} \quad \varphi_\ell(\boldsymbol{x}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}_\ell\|^2}{2\sigma^2}\right).$$

The constant basis function is included since, if two distributions are equal, the density ratio would be $r(\boldsymbol{x}) = 1$. To prevent overfitting, we add a regularizer of the form $\lambda \boldsymbol{\alpha}^\top \boldsymbol{R} \boldsymbol{\alpha}$ to the objective function, where λ is a small constant, $\boldsymbol{R}$ is defined as

$$\boldsymbol{R} = \begin{bmatrix} 0 & \mathbf{0}_{1\times b} \\ \mathbf{0}_{b\times 1} & \boldsymbol{I}_{b\times b} \end{bmatrix}, \tag{18}$$

$\mathbf{0}_{a\times b}$ denotes the zero matrix of size $a \times b$, and $\boldsymbol{I}_{b\times b}$ denotes a $b \times b$ identity matrix. Since the regularizer should penalize non-smoothness, the constant basis function was not regularized. The model for the density ratio is then learned by the following regularized empirical maximization problem:

$$\max_{\{\alpha_\ell\}_{\ell=0}^b} \sum_{\ell=0}^{b} \frac{\alpha_\ell}{n'} \sum_{i=1}^{n'} \varphi_\ell(\boldsymbol{x}_i) - \sum_{y=1}^{c} \frac{\theta_y}{n_y} \sum_{i:y_i=y} f^*\left(\sum_{\ell=0}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i)\right) \\ -\lambda \sum_{\ell=0}^{b} \sum_{\ell'=0}^{b} \alpha_\ell \alpha_{\ell'} R_{\ell,\ell'}. \tag{19}$$

4.2 KL-Divergence Approximation

The KL divergence is an f-divergence with the function f chosen as

$$f(u) = \begin{cases} -\log(u) & u > 0, \\ +\infty & u \leq 0. \end{cases}$$

By substituting the above into Eq.(15), the convex conjugate can be calculated as

$$f^*(v) = \begin{cases} -1 - \log(-v) & v < 0, \\ +\infty & v \geq 0. \end{cases}$$

For the sake of convenience, we regard $-r(\boldsymbol{x})$ as $r(\boldsymbol{x})$ (Nguyen et al., 2010). We can then write the empirical approximation of Eq.(16) under Eqs.(5) and (17) as follows (Nguyen et al., 2010):

$$\mathrm{KL}\left(p' \| q'\right) \approx \max_{\{\alpha_\ell\}_{\ell=0}^b} -\frac{1}{n'} \sum_{i=1}^{n'} \sum_{\ell=0}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i') + \sum_{y=1}^{c} \frac{\theta_y}{n_y} \sum_{i:y_i=y} \log\left(\sum_{\ell=0}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i)\right) + 1.$$

We note that when the above is used to estimate $\mathrm{KL}(p' \| q')$, the function $r(\boldsymbol{x})$ will be an estimate of the density ratio $q'(\boldsymbol{x})/p'(\boldsymbol{x})$. An alternative choice would be to estimate $\mathrm{KL}(q' \| p')$ which would lead to an estimate of $r(\boldsymbol{x}) = p'(\boldsymbol{x})/q'(\boldsymbol{x})$. The density $q'(\boldsymbol{x})$ may however have a small support for certain values of $\{\theta_y\}_{y=1}^c$, causing the density ratio $p'(\boldsymbol{x})/q'(\boldsymbol{x})$ to diverge. For this reason, the estimator $\mathrm{KL}(p' \| q')$ which estimates the density ratio $q'(\boldsymbol{x})/p'(\boldsymbol{x})$ is preferred.

The resulting regularized optimization problem,

$$\max_{\{\alpha_\ell\}_{\ell=0}^{b}} -\frac{1}{n'}\sum_{i=1}^{n'}\sum_{\ell=0}^{b}\alpha_\ell\varphi_\ell(\boldsymbol{x}'_i)+\sum_{y=1}^{c}\frac{\theta_y}{n_y}\sum_{i:y_i=y}\log\left(\sum_{\ell=0}^{b}\alpha_\ell\varphi_\ell(\boldsymbol{x}_i)\right)-\lambda\sum_{\ell=0}^{b}\sum_{\ell'=0}^{b}\alpha_\ell\alpha_{\ell'}R_{\ell,\ell'},$$

is convex and the solution can be obtained by naive optimization. The Gaussian width and regularization constant can be systematically optimized by cross-validation. The KL-divergence estimator obtained above was proved to possess superior convergence properties both in parametric and non-parametric setups (Nguyen et al., 2010; Sugiyama et al., 2008).

However, in the current context of estimating the test class-priors, computing the KL-divergence estimator is rather time-consuming because optimization of $\{\alpha_\ell\}_{\ell=0}^{b}$ needs to be carried out for each $\{\theta_y\}_{y=1}^{c}$.

4.3 PE-Divergence Approximation

As an alternative to the KL divergence, let us consider the PE divergence defined by

$$\begin{aligned}\mathrm{PE}(q'\|p') &:= \frac{1}{2}\int\left(\frac{q'(\boldsymbol{x})}{p'(\boldsymbol{x})}-1\right)^2 p'(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &=\frac{1}{2}\int\left(\frac{q'(\boldsymbol{x})}{p'(\boldsymbol{x})}\right)^2 p'(\boldsymbol{x})\mathrm{d}\boldsymbol{x}-\frac{1}{2},\end{aligned}$$

which is an f-divergence with

$$f(u)=\frac{u^2}{2}-\frac{1}{2}.$$

For this f, the convex conjugate is given by

$$f^*(v)=\frac{v^2}{2}+\frac{1}{2}.$$

The function $r(\boldsymbol{x})$ will again be an estimate of the ratio $q'(\boldsymbol{x})/p'(\boldsymbol{x})$. The empirical approximation of Eq.(16) under Eqs.(5) and (17) is given as follows (Kanamori et al., 2009):

$$\mathrm{PE}(q'\|p')\approx\max_{\boldsymbol{\alpha}}\left[-\frac{1}{2}\boldsymbol{\alpha}^\top\widehat{\boldsymbol{G}}\boldsymbol{\alpha}+\boldsymbol{\alpha}^\top\widehat{\boldsymbol{H}}\boldsymbol{\theta}-\frac{1}{2}\right],$$

where

$$\begin{aligned}&\boldsymbol{\alpha}=[\alpha_0\ \alpha_1\ \cdots\ \alpha_b]^\top,\quad \widehat{\boldsymbol{G}}=\frac{1}{n'}\sum_{i=1}^{n'}\boldsymbol{\varphi}(\boldsymbol{x}'_i)\boldsymbol{\varphi}(\boldsymbol{x}'_i)^\top,\\ &\boldsymbol{\varphi}(\boldsymbol{x})=[\varphi_0(\boldsymbol{x})\ \varphi_1(\boldsymbol{x})\ \cdots\ \varphi_b(\boldsymbol{x})],\ \widehat{\boldsymbol{H}}=\left[\widehat{\boldsymbol{h}}_1\ \cdots\ \widehat{\boldsymbol{h}}_c\right],\\ &\widehat{\boldsymbol{h}}_y=\frac{1}{n_y}\sum_{i:y_i=y}\boldsymbol{\varphi}(\boldsymbol{x}_i),\quad \boldsymbol{\theta}=[\theta_1\ \theta_2\ \cdots\ \theta_c]^\top.\end{aligned}$$

A regularized solution to the above maximization problem can be obtained analytically as

$$\widehat{\boldsymbol{\alpha}} = \left(\widehat{\boldsymbol{G}} + \lambda \boldsymbol{R}\right)^{-1} \widehat{\boldsymbol{H}} \boldsymbol{\theta}, \tag{20}$$

where the regularization matrix is defined in Eq.(18). The PE-divergence estimator obtained above was proved to have superior convergence properties both in parametric and non-parametric setups (Kanamori et al., 2009, 2012b). The kernel width and regularization parameter can be systematically optimized by cross-validation (Kanamori et al., 2009, 2012b).

4.4 Learning Class Ratios by PE Divergence Matching

As shown above, the KL and PE divergences can be systematically estimated without density estimation via Legendre-Fenchel convex duality. Among them, the PE-divergence estimator is more useful for our purpose of learning class ratios, because of the following reasons: The PE-divergence was shown to be more robust against outliers than the KL divergence, based on power divergence analysis (Basu et al., 1998; Sugiyama et al., 2012b). This is a useful property in practical data analysis suffering high noise and outliers. Furthermore, the above PE-divergence estimator was shown to possess the minimum condition number among a general class of estimators, meaning that it is the most stable estimator (Kanamori et al., 2012a).

Another, practically more important advantage of the PE-divergence estimator is that it can be computed efficiently and analytically. This analytical solution allows us to express the PE divergence directly in terms of the class priors:

$$\begin{aligned}\widehat{\mathrm{PE}}(\boldsymbol{\theta}) := &-\frac{1}{2}\boldsymbol{\theta}^\top \widehat{\boldsymbol{H}}^\top \left(\widehat{\boldsymbol{G}} + \lambda \boldsymbol{R}\right)^{-1} \widehat{\boldsymbol{G}} \left(\widehat{\boldsymbol{G}} + \lambda \boldsymbol{R}\right)^{-1} \widehat{\boldsymbol{H}} \boldsymbol{\theta} \\ &+ \boldsymbol{\theta}^\top \widehat{\boldsymbol{H}}^\top \left(\widehat{\boldsymbol{G}} + \lambda \boldsymbol{R}\right)^{-1} \widehat{\boldsymbol{H}} \boldsymbol{\theta} - \frac{1}{2}.\end{aligned}$$

The solution can then be obtained by minimizing the above expression with respect to $\boldsymbol{\theta}$.

5 Experiments

In this section, we report experimental results.

5.1 Benchmark Datasets

The following five methods are compared:

- **EM-KLR**: The method of Saerens et al. (2001) (see Section 2.2). The class-posterior probability of the training dataset is estimated using ℓ_2-penalized kernel logistic regression with Gaussian kernels. The L-BFGS quasi-Newton implementation included in the 'minFunc' package is used for logistic regression training (Schmidt, 2005).

Table 1: Datasets used in the experiments. The SAHeart dataset was taken from Hastie et al. (2001). All other datasets were taken from the *LIBSVM* webpage (Chang and Lin, 2011).

Dataset	d	# samples	# positives	# negatives
Adult	123	32561	7841	24720
Australian	14	690	307	383
Diabetes	8	768	500	268
German	24	1000	300	700
Ionosphere	34	351	225	126
Ringnorm	20	7400	3664	3736
SAHeart	9	462	302	160
Statlog heart	13	270	120	150

- **KL-KDE**: The estimator of the KL divergence $\mathrm{KL}(p'\|q')$ using kernel density estimation (KDE). The class-wise input densities are estimated by KDE with Gaussian kernels. The kernel widths are estimated using likelihood cross-validation (Silverman, 1986).

- **PE-KDE**: The estimate of the Pearson divergence $\mathrm{PE}(q'\|p')$ using KDE. The class-wise input densities are estimated by KDE with Gaussian kernels. The kernel widths are estimated using least-squares cross-validation (Silverman, 1986).

- **KL-DR**: The proposed method (see Section 4.2) using a KL-divergence estimator based on the density ratio (DR). For the optimization, the L-BFGS implementation 'minFunc' is used (Schmidt, 2005).

- **PE-DR**: The proposed method (see Section 4.4) using the PE-divergence estimator based on DR.

Here, we use binary-classification benchmark datasets listed in Table 1. We select 10 samples from each of the two classes for the training dataset and 50 samples for the test dataset. The samples in the test set are selected with probability θ^* from the first class and with probability $(1-\theta^*)$ from the second class. The experiments are performed for several class-priors, selected as $\theta^* \in [0.1\, 0.2\, \ldots\, 0.8\, 0.9]$.

The squared error of the estimated class-priors averaged over 1000 runs are given in Figure 1. This shows that methods based on the KL and PE divergences overall outperform EM-KLR, implying that our reformulation of the EM algorithm as distribution matching (see Section 3) contributes to obtaining accurate class-ratio estimates. Among the divergence-based methods, PE-DR and KL-DR outperforms PE-KDE and KL-KDE, showing that directly estimating density ratios without density estimation is more promising as divergence estimators. Overall, PE-DR and KL-DR are shown to be the most accurate.

The average calculation time for the estimation of the class priors is given in Figure 2. From this, it can be seen that the speed of the PE-DR method is similar to the EM-KLR method and two orders of magnitude faster than the KL-DR method.

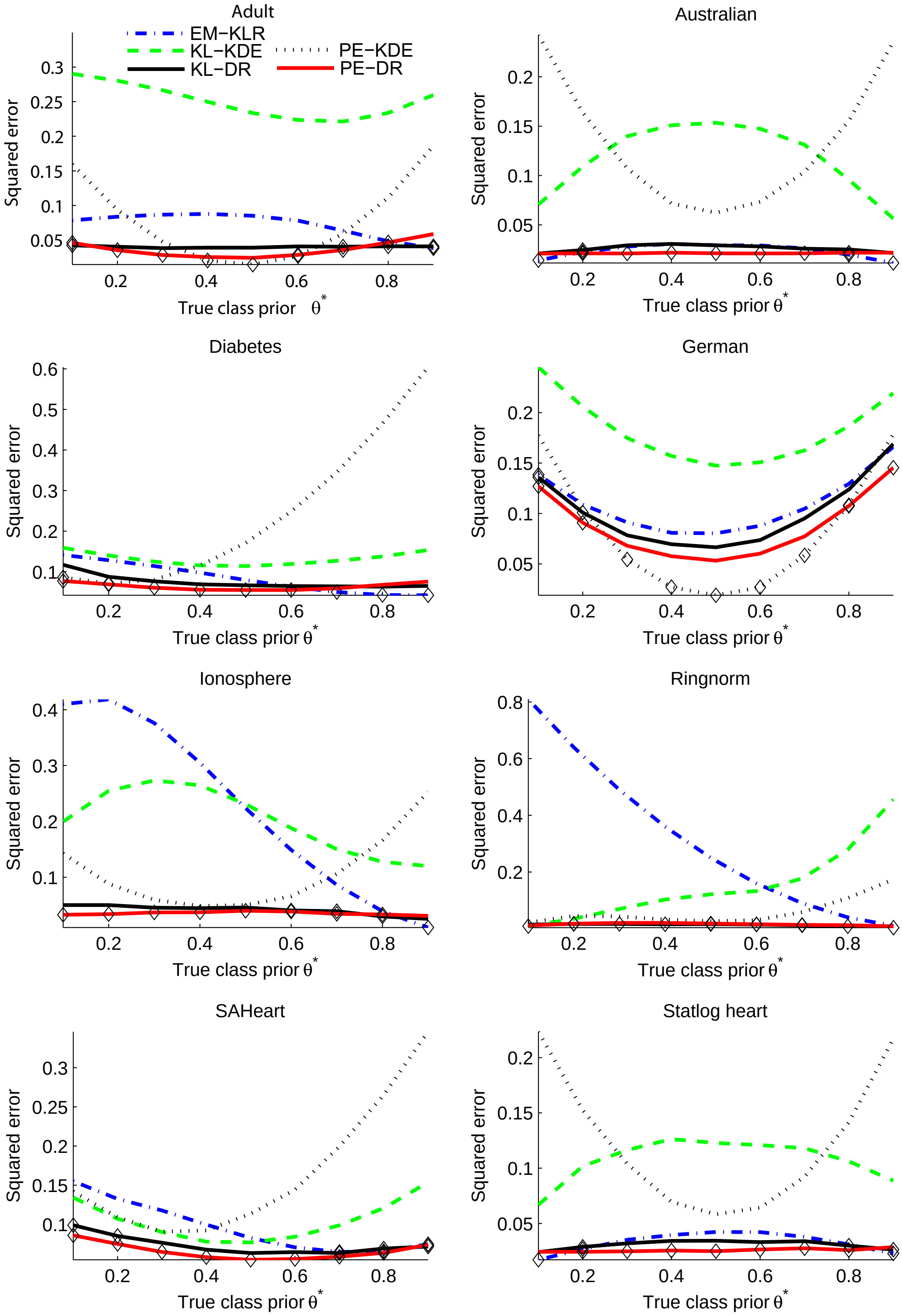


Figure 1: Average squared error between the true class-prior θ^* and estimated class-prior $\widehat{\theta}$ for the benchmark datasets listed in Table 1. The best method and comparable methods according to the t-test at significance level of 5% are indicated with '$\diamond$'

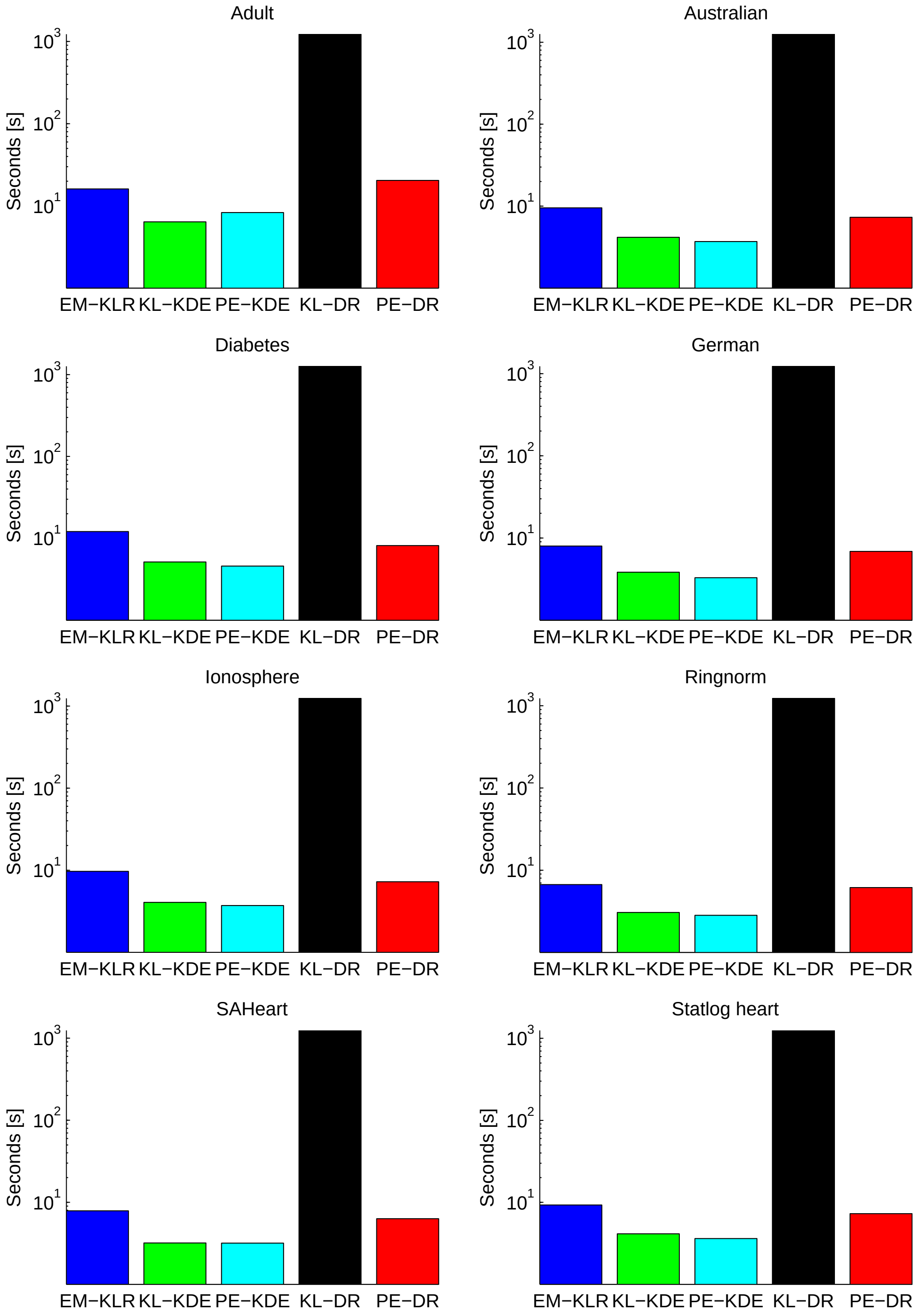


Figure 2: Average calculation time for the estimation of the class priors for the datasets listed in Table 1.

To illustrate how more accurate estimates of the class prior translate into higher classification accuracies, we train a classifier with the estimated prior. For the binary benchmark experiments, a weighted variant of the ℓ_2-regularized kernel logistic regression classifier (Hastie et al., 2001) was used.

We minimize the prior-corrected expected loss of Eq.(2), where the expectation is approximated by its empirical average and the class priors are replaced by the estimated class-priors. Using the logistic loss, a classifier can be learned as,

$$\left(\widehat{\beta}_1, \ldots, \widehat{\beta}_n\right) := \underset{\beta_1,\ldots,\beta_n}{\operatorname{argmin}} \left[\sum_{y=1}^{2} \frac{\widehat{\theta}_y}{n_y} \sum_{i:y_i=y} L\left(z_i, \sum_{\ell=1}^{n} \beta_\ell K(\boldsymbol{x}_i, \boldsymbol{x}_\ell)\right) + \delta \sum_{\ell=1}^{n} \beta_\ell^2 \right],$$

where $L(z, f(\boldsymbol{x}))$ is the logistic loss defined as

$$L(z, f(\boldsymbol{x})) = \log\left(1 + \exp\left(-zf(\boldsymbol{x})\right)\right),$$

and the class labels $y \in \{1, 2\}$ are encoded as $z \in \{-1, 1\}$. The width of the Gaussian kernel $K(\boldsymbol{x}, \boldsymbol{x}')$ and the regularization parameter $\delta(\geq 0)$ are chosen by 5-fold weighted cross-validation (Sugiyama et al., 2007) in terms of the misclassification error. The class label $\widehat{y}$ for the test input $\boldsymbol{x}$ is then estimated by

$$\widehat{y} = \begin{cases} 1 & \sum_{i=1}^{n} \widehat{\beta}_i K(\boldsymbol{x}, \boldsymbol{x}_\ell) < 0, \\ 2 & \text{otherwise.} \end{cases}$$

The results in Figure 3 show that, as expected, a more accurate estimate of the class prior tends to give a lower misclassification rate. Taking into account both the computation time and accuracy, the PE-DR method is overall the most promising method.

5.2 Real-World Application

Finally, we demonstrate the usefulness of the proposed approach in a real-world problem of military vehicle classification from geophone recordings (Duarte and Hu, 2004). This is a three-class problem: two vehicle classes and a class of recorded noise. The features are 50-dimensional. In this vehicle classification task, class-prior change is inevitable because the type of vehicles passing through differs depending on time (e.g., day and night).

n samples are drawn from each of the classes for the training set, whereas 100 samples are drawn with probabilities $p = [0.6\ 0.1\ 0.3]$ from each of the classes for the test set. Due to the prohibitive computational cost, KL-DR was not included in this experiment.

In Figure 4, we plot the ℓ_2-distance between the true and estimated class-priors and the misclassification rate based on instance-weighted kernel logistic regression (Hastie et al., 2001) averaged over 1000 runs as functions of the number of training samples. As can be seen from the graphs, the performance of all methods improves as the number of training samples increases. Among the compared methods, PE-DR provides the most accurate estimates of the class prior and thus yields the lowest classification error.

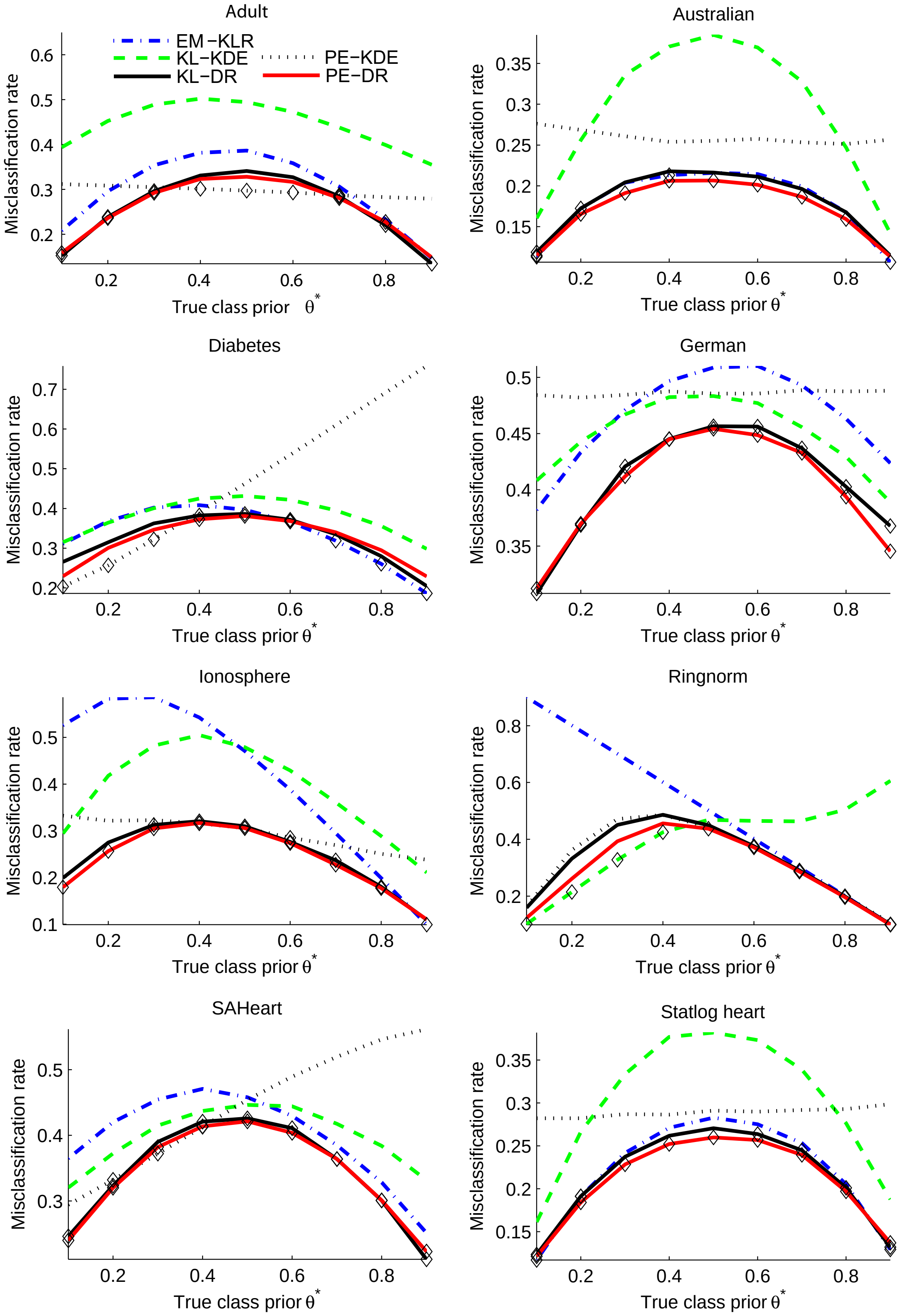
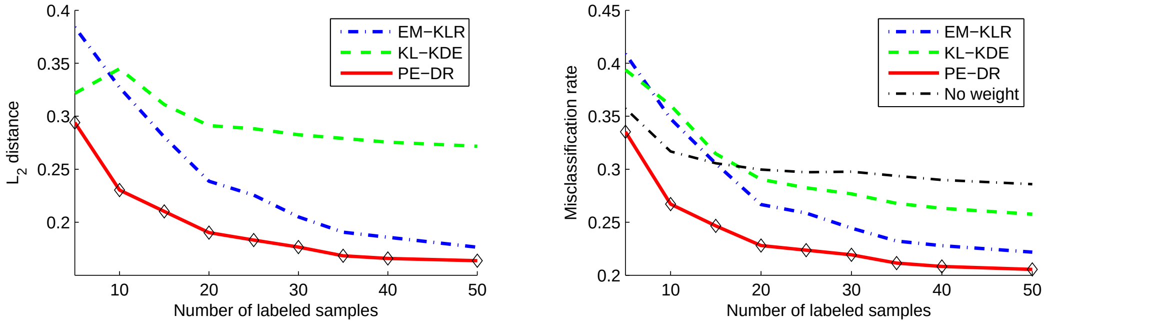


Figure 3: Average misclassification rates for the datasets listed in Table 1. Classification is performed using a regularized kernel logistic regression classifier with instance weighting. The best method and comparable methods according to the t-test at significance level of 5% are indicated with '◇'.



(a) ℓ_2-distance between true and estimated class-priors.

(b) Misclassification rate with instance-weighted kernel logistic regression.

Figure 4: Experimental results for the vehicle classification problem. The best method and comparable methods according to the t-test at significance level of 5% are indicated with a '$\diamond$'.

6 Conclusion

Class-prior change is a problem that is conceivable in many real-world datasets, and it can be systematically corrected for if the class prior of the test dataset is known. In this paper, we discussed the problem of estimating the test class-priors under a semi-supervised learning setup.

We first showed that the EM-based estimator introduced in Saerens et al. (2001) can be regarded as indirectly approximating the test input distribution by a linear combination of class-wise input distributions. Based on this view, we proposed to use an explicit and possibly more accurate divergence estimator based on density-ratio estimation (Kanamori et al., 2009) for learning test class-priors. The proposed method was shown to have various nice properties such as high robustness to noise and outliers, superior numerical stability, and excellent computational efficiency. Through experiments, we showed that the class ratios estimated by the proposed method are more accurate than competing methods, which can be translated into better classification accuracy.

References

Ali, S. M., Silvey, S. D., 1966. A general class of coefficients of divergence of one distribution from another. Journal of the Royal Statistical Society, Series B 28, 131–142.

Basu, A., Harris, I. R., Hjort, N. L., Jones, M. C., 1998. Robust and efficient estimation by minimising a density power divergence. Biometrika 85 (3), 549–559.

Bishop, C. M., 2006. Pattern Recognition and Machine Learning. Springer, New York, NY, USA.

Boyd, S., Vandenberghe, L., 2004. Convex Optimization. Cambridge University Press, New York, NY, USA.

Chan, Y. S., Ng, H. T., 2006. Estimating class priors in domain adaptation for word sense disambiguation. In: Proceedings of the 21st International Conference on Computational Linguistics. pp. 89–96.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27, software available at `http://www.csie.ntu.edu.tw/ cjlin/libsvm`.

Chapelle, O., Schölkopf, B., Zien, A. (Eds.), 2006. Semi-Supervised Learning. MIT Press, Cambridge, MA, USA.

Clémençon, S., Vayatis, N., Depecker, M., 2009. AUC optimization and the two-sample problem. In: Advances in Neural Information Processing Systems 22. pp. 360–368.

Cortes, C., Mohri, M., 2004. AUC optimization vs. error rate minimization. In: Advances in Neural Information Processing Systems 16. MIT Press, Cambridge, MA, pp. pp. 313–320.

Csiszár, I., 1967. Information-type measures of difference of probability distributions and indirect observation. Studia Scientiarum Mathematicarum Hungarica 2, 229–318.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, series B 39 (1), 1–38.

du Plessis, M. C., Sugiyama, M., 2012. Semi-supervised learning of class balance under class-prior change by distribution matching. In: Langford, J., Pineau, J. (Eds.), Proceedings of 29th International Conference on Machine Learning (ICML2012). Edinburgh, Scotland, pp. 823–830.

Duarte, M. F., Hu, Y. H., 2004. Vehicle classification in distributed sensor networks. Journal of Parallel and Distributed Computing 64 (7), 826–838.

Duda, R. O., Hart, P. E., Stork, D. G., 2001. Pattern classification, 2nd Edition. Wiley, New York, NY, USA.

Elkan, C., 2001. The foundations of cost-sensitive learning. In: Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence. pp. 973–978.

Hall, P., 1981. On the non-parametric estimation of mixture proportions. Journal of the Royal Statistical Society. Series B (Methodological), 147–156.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, NY, USA.

Heckman, J. J., 1979. Sample selection bias as a specification error. Econometrica 47 (1), 153–161.

Hunter, J., Nachtergaele, B., 2001. Applied Analysis. World Scientific Inc. Co., River Edge, NY, USA.

Kanamori, T., Hido, S., Sugiyama, M., 2009. A least-squares approach to direct importance estimation. Journal of Machine Learning Research 10, 1391–1445.

Kanamori, T., Suzuki, T., Sugiyama, M., 2012a. Computational complexity of kernel-based density-ratio estimation: A condition number analysis. Machine Learning, to appear.

Kanamori, T., Suzuki, T., Sugiyama, M., 2012b. Statistical analysis of kernel-based least-squares density-ratio estimation. Machine Learning 86 (3), 335–367.

Keziou, A., 2003. Dual representation of ϕ-divergences and applications. Comptes Rendus Mathématique 336 (10), 857–862.

Kullback, S., Leibler, R. A., 1951. On information and sufficiency. Annals of Mathematical Statistics 22, 79–86.

Latinne, P., Saerens, M., Decaestecker, C., 2001. Adjusting the outputs of a classifier to new a priori probabilities may significantly improve classification accuracy: Evidence from a multi-class problem in remote sensing. In: Proceedings of the 18th International Conference on Machine Learning. pp. 298–305.

Lin, Y., Lee, Y., Wahba, G., 2002. Support vector machines for classification in nonstandard situations. Machine Learning 46 (1), 191–202.

McLachlan, G. J., Krishnan, T., 1997. The EM algorithm and extensions. John Wiley and Sons, New York, NY, USA.

Nguyen, X., Wainwright, M. J., Jordan, M. I., 2010. Estimating divergence functionals and the likelihood ratio by convex risk minimization. IEEE Transactions on Information Theory 56 (11), 5847–5861.

Pearson, K., 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine 50, 157–175.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N. (Eds.), 2009. Dataset Shift in Machine Learning. MIT Press, Cambridge, MA, USA.

Rockafellar, R. T., 1970. Convex Analysis. Princeton University Press, Princeton, NJ, USA.

Saerens, M., Latinne, P., Decaestecker, C., 2001. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. Neural Computation 14, 21–41.

Schmidt, M., 2005. minFunc—Unconstrained differentiable multivariate optimization in MATLAB.

Silverman, B. W., 1986. Density Estimation: For Statistics and Data Analysis. Chapman and Hall, London, UK.

Sugiyama, M., 2010. Superfast-trainable multi-class probabilistic classifier by least-squares posterior fitting. IEICE Transactions on Information and Systems E93-D, 2690–2701.

Sugiyama, M., Kawanabe, M., 2012. Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation. MIT Press, Cambridge, MA, USA.

Sugiyama, M., Krauledat, M., Müller, K.-R., May 2007. Covariate shift adaptation by importance weighted cross validation. Journal of Machine Learning Research 8, 985–1005.

Sugiyama, M., Suzuki, T., Kanamori, T., 2012a. Density Ratio Estimation in Machine Learning. Cambridge University Press, Cambridge, UK.

Sugiyama, M., Suzuki, T., Kanamori, T., 2012b. Density ratio matching under the Bregman divergence: A unified framework of density ratio estimation. Annals of the Institute of Statistical Mathematics, to appear.

Sugiyama, M., Suzuki, T., Kanamori, T., du Plessis, M. C., Liu, S., Takeuchi, I., 2013. Density-difference estimation. Neural Computation, to appear.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., Kawanabe, M., 2008. Direct importance estimation for covariate shift adaptation. Annals of the Institute of Statistical Mathematics 60 (4), 699–746.

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

Van Trees, H., 1968. Detection, Estimation, and Modulation Theory, Part I. Detection, Estimation, and Modulation Theory. John Wiley and Sons, New York, NY, USA.

Vapnik, V. N., 1998. Statistical Learning Theory. Wiley, New York, NY, USA.