IEICE Transactions on Information and Systems,
vol.E97-D, no.5, pp.1358–1362, 2014.

Class Prior Estimation from Positive and Unlabeled Data

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

We consider the problem of learning a classifier using only positive and unlabeled samples. In this setting, it is known that a classifier can be successfully learned if the class prior is available. However, in practice, the class prior is unknown and thus must be estimated from data. In this paper, we propose a new method to estimate the class prior by partially matching the class-conditional density of the positive class to the input density. By performing this partial matching in terms of the Pearson divergence, which we estimate directly without density estimation via lower-bound maximization, we can obtain an analytical estimator of the class prior. We further show that an existing class prior estimation method can also be interpreted as performing partial matching under the Pearson divergence, but in an indirect manner. The superiority of our *direct* class prior estimation method is illustrated on several benchmark datasets.

Keywords

Class-prior change, outlier detection, positive and unlabeled learning, divergence estimation, Pearson divergence.

1 Introduction

Standard classification problems assume that all training samples are labeled. However, in many practical problems only some of the positive samples may be labeled. This occurs in problems such as outlier/novelty detection [4, 5], where the labels of some inlier samples are known, or one-class land-cover identification [8], where land-cover areas of the same class as the labeled instances must be identified. The goal of this paper is to learn a classifier only from positive and unlabeled data.

We assume that the data is drawn according to

$$(\boldsymbol{x}, y, s) \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}, y, s),$$

where $\boldsymbol{x}(\in \mathbb{R}^d)$ are the unlabeled features, $y(\in \{-1, 1\})$ are the (unknown) class labels, and $s(\in \{0, 1\})$ determines whether the sample is labeled. The assumption is that only positive samples are labeled [2],

$$p(s = 1|\boldsymbol{x}, y = -1) = 0, \tag{1}$$

and that the probability that a sample is labeled depends only on the underlying label:

$$p(s = 1|\boldsymbol{x}, y = 1) = p(s = 1|y = 1).$$

Since we do not observe all class labels, the dataset would typically be

$$\mathcal{X} := \{(\boldsymbol{x}_i, s_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}, s). \tag{2}$$

When $s_i = 1$, sample $\boldsymbol{x}_i$ would have the label $y_i = 1$, according to assumption (1). When $s_i = 0$, the sample is unlabeled and the (unknown) underlying label may be $y_i = 1$ or $y_i = -1$.

We note that this learning setting has an intrinsic problem: unlike the traditional classification setting, we can not trivially estimate the class prior $p(y = 1)$ from the dataset $\mathcal{X}$. The focus of this paper is to present a new method for estimating this class prior.

When the class prior $p(y = 1)$ is estimated or specified by the user, the classifier can be estimated from the dataset. In [2] the following relation was proven,

$$p(y = 1|\boldsymbol{x}) = \frac{1}{c} p(s = 1|\boldsymbol{x}),$$

where

$$c := p(s = 1|y = 1).$$

The posterior $p(s = 1|\boldsymbol{x})$ is referred to in [2] as a 'non-traditional' classifier. This can be estimated from the training set in (2) by a probabilistic classification method such as kernel logistic regression [3] or its squared-loss variant [10].

The constant c that is used to reweight the non-traditional classifier can be expressed as

$$c = \frac{p(y = 1|s = 1)p(s = 1)}{p(y = 1)} = \frac{p(s = 1)}{p(y = 1)}. \tag{3}$$

$p(s = 1)$ can be directly estimated from (2), so the reweighting constant can be calculated if we can obtain an estimate of $p(y = 1)$.

We propose a method in the next section to estimate this class prior from the training data. In Section 3 we show that the existing method [2] can be interpreted as indirectly estimating the same quantity as the proposed method. The superiority of our proposed method is illustrated on benchmark data in Section 4.

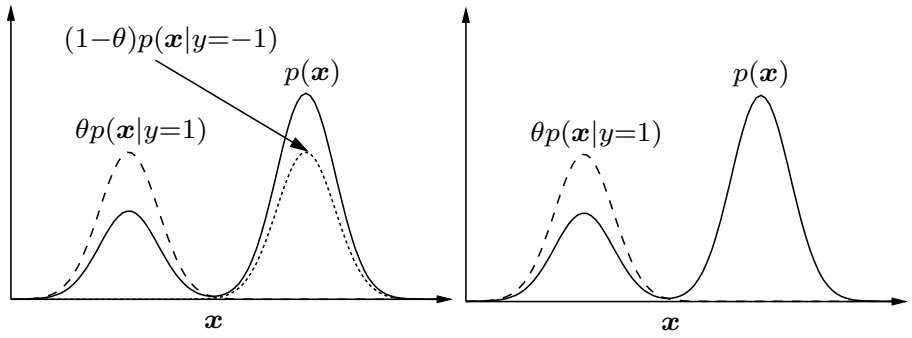

Figure 1: Estimating the class prior via *full matching* (left-hand side) and *partial matching* (right-hand side).

2 Prior Estimation via Partial Matching

In this section we will propose a new method to estimate the class prior by partial matching.

2.1 Basic Idea

We denote the subset of samples in (2) for which $s = 1$ as

$$\mathcal{X}' := \{\boldsymbol{x}'_i\}_{i=1}^{n'}.$$

From the assumptions, the above samples are drawn according to

$$p(\boldsymbol{x}|s = 1) = p(\boldsymbol{x}|y = 1). \tag{4}$$

We model the input density as

$$q(\boldsymbol{x}; \theta) := \theta p(\boldsymbol{x}|y = 1) + (1 - \theta)p(\boldsymbol{x}|y = -1),$$

where $\theta (\in [0, 1])$ is a scalar value that represents the unknown class prior $p(y = 1)$. The above model $q(\boldsymbol{x}; \theta)$ would equal $p(\boldsymbol{x})$ if θ is the unknown class prior $p(y = 1)$. Therefore, by selecting θ so that the two distributions are equal (illustrated in the left graph of Fig. 1), the class prior can be estimated [1]. This setup will however not work in our current context, since we do not have samples drawn from $p(\boldsymbol{x}|y = -1)$ and consequently $q(\boldsymbol{x}; \theta)$ can not be estimated.

Nevertheless, if the class-conditional densities $p(\boldsymbol{x}|y = 1)$ and $p(\boldsymbol{x}|y = -1)$ are not strongly overlapping, we may estimate θ so that $\theta p(\boldsymbol{x}|y = -1)$ is as similar to $p(\boldsymbol{x})$ as possible (this is illustrated in the right graph of Fig. 1). Here we propose to use the Pearson (PE) divergence[1] for matching $\theta p(\boldsymbol{x}|y = -1)$ to $p(\boldsymbol{x})$:

$$\theta^* = \operatorname*{argmin}_{\theta} \mathrm{PE}(\boldsymbol{\theta}),$$

[1]Note that $\theta p(\boldsymbol{x})$ is not a density unless $\theta = 1$.

where $\mathrm{PE}(\boldsymbol{\theta})$ denotes the PE divergence from $\theta p(\boldsymbol{x}|y=1)$ to $p(\boldsymbol{x})$:

$$\begin{aligned}\mathrm{PE} &= \frac{1}{2}\int\left(\frac{\theta p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x})}-1\right)^2 p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &= \frac{1}{2}\int\left(\frac{\theta p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x})}\right)^2 p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}-\theta+\frac{1}{2}. \end{aligned} \tag{5}$$

The above PE divergence is defined in terms of unknown densities, but only samples drawn from these densities are available. A possible approach is to first estimate $p(\boldsymbol{x}|y=1)$ and $p(\boldsymbol{x})$ from the samples using, e.g., kernel density estimation and then plug these estimators into the above expression. This however does not work well since high-dimensional density estimation is a difficult problem [11]. Furthermore, the division by an estimated density may exacerbate the estimation error.

2.2 Estimation Algorithm

Here we show how we can avoid density estimation and directly minimize the PE divergence.

Our idea is to consider a lower bound which is linear in the unknown densities and can then be estimated from sample averages. Using the inequality $y^2/2 \geq ty - t^2/2$ which can be obtained from *Fenchel duality* [7, 9], we can lower bound (5) in a pointwise manner as follows:

$$\frac{1}{2}\left(\frac{\theta p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x})}\right)^2 \geq \left(\frac{\theta p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x})}\right) r(\boldsymbol{x}) - \frac{1}{2}r(\boldsymbol{x})^2,$$

where $r(\boldsymbol{x})$ fulfills the role of t. This yields

$$\frac{1}{2}\left(\frac{\theta p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x})}\right)^2 p(\boldsymbol{x}) \geq \theta p(\boldsymbol{x}|y=1) r(\boldsymbol{x}) - \frac{1}{2}r(\boldsymbol{x})^2 p(\boldsymbol{x}).$$

Therefore the PE divergence is lower bounded as

$$\mathrm{PE} \geq \theta\int r(\boldsymbol{x}) p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x} - \frac{1}{2}\int r(\boldsymbol{x})^2 p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \theta + \frac{1}{2}.$$

The above lower bound can be turned into a practical estimator by using a parametric model for $r(\boldsymbol{x})$, replacing the integrals with sample averages, and selecting the tightest bound via maximization of the right-hand side.

We approximate the function $r(\boldsymbol{x})$ by a linear-in-parameter model $\widehat{r}(\boldsymbol{x}) = \boldsymbol{\alpha}^\top \boldsymbol{\varphi}(\boldsymbol{x})$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^\top$ are the parameters and $\boldsymbol{\varphi}(\boldsymbol{x}) = (\varphi_1(\boldsymbol{x}), \ldots, \varphi_n(\boldsymbol{x}))^\top$ are the basis functions. In practice, we use Gaussian basis functions centered at the training points:

$$\varphi_i(\boldsymbol{x}) = \exp\left(-\frac{1}{2\sigma^2}\left\|\boldsymbol{x}-\boldsymbol{x}_i\right\|^2\right), \quad i = 1, \ldots, n.$$

Using this model, objective function can be written as

$$\begin{aligned}\widehat{\boldsymbol{\alpha}} &:= \operatorname*{argmax}_{\boldsymbol{\alpha}} \theta\boldsymbol{\alpha}^\top\boldsymbol{h} - \frac{1}{2}\boldsymbol{\alpha}^\top\boldsymbol{H}\boldsymbol{\alpha} - \theta + \frac{1}{2},\\ &= \operatorname*{argmax}_{\boldsymbol{\alpha}} \theta\boldsymbol{\alpha}^\top\boldsymbol{h} - \frac{1}{2}\boldsymbol{\alpha}^\top\boldsymbol{H}\boldsymbol{\alpha},\end{aligned}$$

where

$$\boldsymbol{H} = \int \boldsymbol{\varphi}(\boldsymbol{x})\boldsymbol{\varphi}(\boldsymbol{x})^\top p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \quad \boldsymbol{h} = \int \boldsymbol{\varphi}(\boldsymbol{x})p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x}.$$

Estimating the integrals by their sample averages gives

$$\widehat{\boldsymbol{H}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\varphi}(\boldsymbol{x}_i)\boldsymbol{\varphi}(\boldsymbol{x}_i)^\top, \quad \widehat{\boldsymbol{h}} = \frac{1}{n'}\sum_{i=1}^{n'} \boldsymbol{\varphi}(\boldsymbol{x}'_i).$$

Using these empirical estimates and adding an ℓ_2 regularizer leads to the following optimization problem:

$$\widehat{\boldsymbol{\alpha}} := \operatorname*{argmax}_{\boldsymbol{\alpha}} \theta\boldsymbol{\alpha}^\top\widehat{\boldsymbol{h}} - \frac{1}{2}\boldsymbol{\alpha}^\top\widehat{\boldsymbol{H}}\boldsymbol{\alpha} - \frac{\lambda}{2}\boldsymbol{\alpha}^\top\boldsymbol{\alpha},$$

where $\lambda(\geq 0)$ is the regularization parameter. This can be analytically solved as

$$\widehat{\boldsymbol{\alpha}} = \theta\widehat{\boldsymbol{G}}^{-1}\widehat{\boldsymbol{h}}, \quad \widehat{\boldsymbol{G}} = \widehat{\boldsymbol{H}} + \lambda\boldsymbol{I},$$

where $\boldsymbol{I}$ denotes the identity matrix. Note that the above objective function is essentially the same as the least-squares objective function in [6].

Substituting the analytical solution into the lower bound yields the following PE divergence estimator:

$$\widehat{\mathrm{PE}} = \theta^2\widehat{\boldsymbol{h}}^\top\widehat{\boldsymbol{G}}^{-1}\widehat{\boldsymbol{h}} - \theta^2\frac{1}{2}\widehat{\boldsymbol{h}}^\top\widehat{\boldsymbol{G}}^{-1}\widehat{\boldsymbol{H}}\widehat{\boldsymbol{G}}^{-1}\widehat{\boldsymbol{h}} - \theta + \frac{1}{2}.$$

This can be analytically minimized with respect to θ to yield the following estimator of the class prior:

$$\widehat{\theta} = \left[2\widehat{\boldsymbol{h}}^\top\widehat{\boldsymbol{G}}^{-1}\widehat{\boldsymbol{h}} - \widehat{\boldsymbol{h}}^\top\widehat{\boldsymbol{G}}^{-1}\widehat{\boldsymbol{H}}\widehat{\boldsymbol{G}}^{-1}\widehat{\boldsymbol{h}}\right]^{-1}.$$

2.3 Theoretical Analysis

Here, we theoretically investigate the bias of our algorithm when the assumption that class-conditional densities are non-overlapping is violated.

Assuming that the densities $p(\boldsymbol{x}|y=1)$ and $p(\boldsymbol{x})$ are known, we can analytically find the minimizer of (5) with respect to θ as

$$\theta = \left[\int \frac{p(\boldsymbol{x}|y=1)^2}{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}\right]^{-1}. \tag{6}$$

Substituting the identity

$$p(\boldsymbol{x}|y=1) = \left[p(\boldsymbol{x}) - (1-p(y=1))p(\boldsymbol{x}|y=-1)\right]/p(y=1)$$

into the above gives

$$\theta = \frac{p(y=1)}{1-[1-p(y=1)]\int \frac{p(\boldsymbol{x}|y=1)p(\boldsymbol{x}|y=-1)}{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}}.$$

If the class-conditional densities are completely non-overlapping, then $p(\boldsymbol{x}|y=1)p(\boldsymbol{x}|y=-1) = 0$ and the estimator will be unbiased. Otherwise, we see that the value in the denominator is always smaller than 1, which means that the estimator will have a positive bias.

3 Analysis of Existing Method

In this section we analyze the method of estimating the class prior introduced in [2]. The paper proposed that a non-traditional classifier $g(\boldsymbol{x}) \approx p(s=1|\boldsymbol{x})$ is obtained from the training data. Using this classifier and a hold out set of positive samples P of size $|P|$, the constant c given by (3) is estimated as

$$c \approx \frac{1}{|P|}\sum_{x\in P} g(\boldsymbol{x}). \tag{7}$$

Since samples in P are drawn from $p(\boldsymbol{x}|y=1)$, (7) is essentially an estimate of

$$c = \int p(s=1|\boldsymbol{x})p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x},$$

where $p(s=1|\boldsymbol{x})$ is estimated by a non-traditional classifier and the summation in (7) is due to estimation via an empirical average. Using (4) the above can be expressed as

$$c = \int \frac{p(\boldsymbol{x}|y=1)p(s=1)}{p(\boldsymbol{x})}p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x}.$$

Following from (3), the class prior is expressed as

$$p(y=1) = \frac{p(s=1)}{c} = \left[\int \frac{p(\boldsymbol{x}|y=1)^2}{p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}\right]^{-1},$$

which corresponds to (6).

Therefore, both methods can be viewed as estimating the class prior via PE divergence estimation. The important difference is how this estimation is performed. The existing method first learns a function g to estimate the posterior $p(s=1|\boldsymbol{x})$ using a method such as kernel logistic regression. The PE divergence is then estimated using this function.

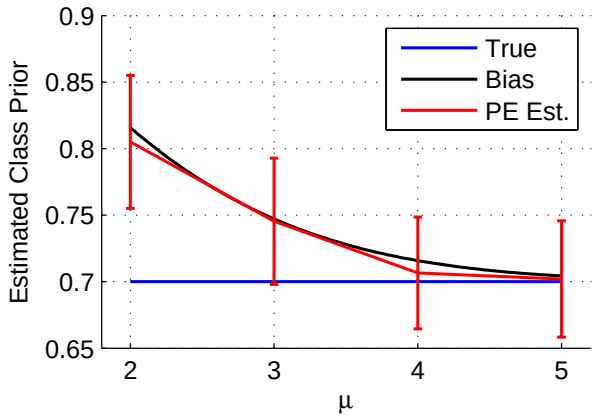

Figure 2: The effect of overlapping class-conditional densities illustrated with by Gaussian class-conditional densities. The true class prior is 0.7. If the class-conditional densities are highly overlapping (when μ is small), the estimate is positively biased.

This two-step approach may however not be optimal, since the best function estimated in the first step may not be the best for PE divergence estimation.

Our proposed method follows a single step approach: we directly learn a function based on how well the PE divergence is estimated. Therefore, our method is expected to perform better. We will experimentally investigate the superiority of our proposed approach in the next section.

4 Experiments

In this section, we illustrate the effect of bias due to overlapping class-conditional densities and then experimentally evaluate the performance of the proposed method on benchmark datasets.

4.1 Numerical Illustration

First, we show the effect of bias caused by highly overlapping class-conditional densities with the aid of a toy example. The class-conditional densities were selected as two univariate Gaussians differing in their means:

$$p(x|y=1) = \mathcal{N}\left(x; 0, 1^2\right), \quad p(x|y=-1) = \mathcal{N}\left(x; \mu, 1^2\right),$$

where $\mathcal{N}\left(x; \mu, \sigma^2\right)$ denotes the Gaussian density with mean μ and variance σ^2 with respect to x. Varying μ controls the overlap between the class conditional densities: a small μ implies a high overlap; conversely, a large μ implies a low overlap. The result of varying μ, when the true class prior is 0.7 is given in Figure 2. From this we see that the bias decreases as the overlap between class-conditional densities decreases.

4.2 Benchmark datasets

We compared the accuracy of the estimate of the class prior on several UCI benchmark datasets[2]. The following methods were compared:

- **PE** (proposed): The method described in Section 2 that directly estimates the PE divergence[3]. All hyper-parameters was set using 5-fold cross validation.
- **EN**: The method of [2] discussed in Section 3. Data was split into 5 folds $\{\mathcal{X}_t\}_{t=1}^5$ and the posterior $p(s=1|\boldsymbol{x})$ was estimated from $\mathcal{X} \setminus \mathcal{X}_t$ (i.e., all samples except $\mathcal{X}_t$). The score in (7) was computed with $P = \mathcal{X}_t$. This was repeated for $t = 1, \ldots, 5$ and the average was used as the estimate of c.

The posterior $p(s=1|\boldsymbol{x})$ was estimated using kernel logistic regression [3]. The true class prior was varied as $0.2, 0.5, 0.8$ and positive samples were labeled with a probability of $p(s=1|y=1) = 0.5$. The resulting squared error and classification accuracy is given in Fig. 3. The classification accuracy was measured on an independent dataset with the same class balance.

As can be seen from the results, our proposed method gave a more accurate estimate of the class prior. Furthermore, the more accurate estimate of the class prior translated into a higher classification accuracy.

4.3 Extreme Class Imbalance

Experiments were also performed to illustrate class prior estimation when an extreme class imbalance occurs. Since the number of samples of the minority class is extremely low in such a case, a large number of samples is needed. The squared error and resulting classification accuracy for an experiment with 200 samples is given in Tables 1 and 2.

Since a positive sample is labeled with probability $p(s=1|y=1)$, the number of labeled samples depends on the number of positive samples. When the number of labeled samples is low (i.e., when the class prior is 0.05 and 0.1), our method gives a much more accurate estimate of the class prior. When the number of labeled samples is large (i.e., when the class prior is 0.95) both methods give highly accurate estimates of the class prior which lead to similar classification accuracy.

5 Conclusion

We proposed a new method to estimate the class prior from positive and unlabeled samples by partial matching under the PE divergence. By obtaining a lower bound for the PE divergence, we can directly get an analytical divergence estimator and estimate the class prior in a single step.

[2] '`http://archive.ics.uci.edu/ml/`'.

[3] An implementation of this method is available from '`http://sugiyama-www.cs.titech.ac.jp/~christo/`'.

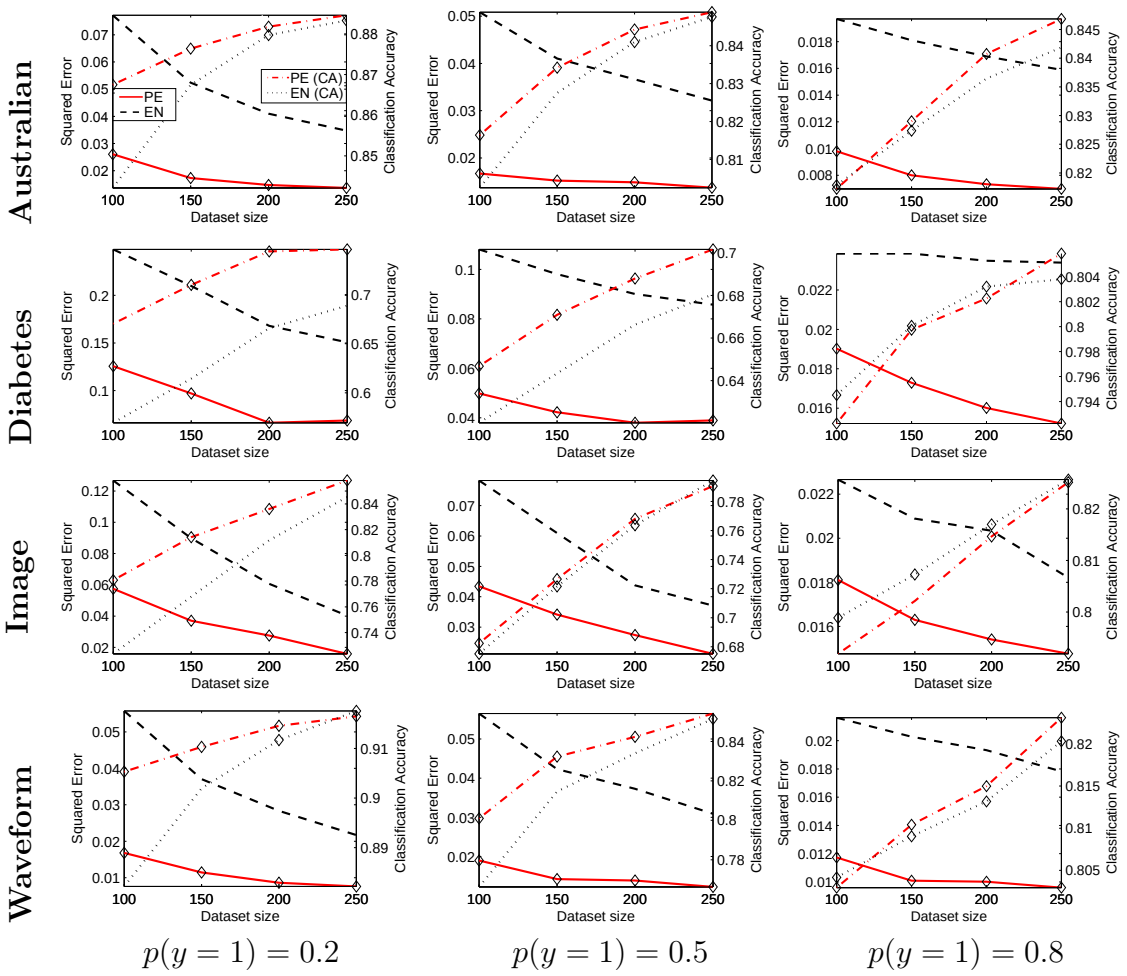

Figure 3: Experimental results on several UCI benchmark datasets. 'PE' and 'PE (CA)' indicate the squared error and classification accuracy for class-prior estimation via direct PE divergence estimation. 'EN' and 'EN (CA)' indicate the squared error and classification accuracy for class-prior estimation using the method of [2]. The diamond symbol means that the method is the best or comparable in terms of the mean performance by t-test with significance level 5%.

As was shown, the existing method of [2] can also be interpreted as matching using the PE divergence. However, in that work, the estimation was indirectly performed using a two-step approach.

Experiments on benchmark data showed that our single-step approach gave a more accurate estimate of the class prior, which in turn resulted in a higher classification accuracy.

Table 1: Accuracy in terms of squared error for class-prior estimation on extremely imbalanced datasets. The dataset size was 200 and bold text indicates the best or comparable method under t-test with significance level of 5%.

Dataset	**0.05**		**0.1**		**0.9**		**0.95**	
	PE	**EN**	**PE**	**EN**	**PE**	**EN**	**PE**	**EN**
australian	**.037**	.096	**.019**	.048	**.004**	.006	.002	**.002**
diabetes	**.230**	.341	**.119**	.250	**.006**	.007	**.002**	**.002**
image	**.060**	.150	**.031**	.078	**.005**	.006	.002	**.001**
waveform	**.022**	**.027**	**.007**	.023	**.005**	.006	.002	**.001**

Table 2: Resulting classification accuracy for class-prior estimation of extremely imbalanced datasets.

Dataset	**0.05**		**0.1**		**0.9**		**0.95**	
	PE	**EN**	**PE**	**EN**	**PE**	**EN**	**PE**	**EN**
australian	**.900**	.865	**.904**	.893	**.898**	**.899**	**.944**	**.946**
diabetes	**.709**	.588	**.747**	.636	**.895**	**.895**	**.946**	**.946**
image	**.885**	.808	**.887**	.839	**.897**	**.899**	**.947**	**.947**
waveform	**.966**	**.955**	**.952**	.945	**.898**	**.898**	**.948**	**.949**

Acknowledgements

MCdP is supported by the MEXT scholarship and MS is supported by KAKENHI 25700022 and AOARD.

References

[1] M. C. du Plessis and M. Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. In *ICML 2012*, pages 823–830, 2012.

[2] C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *14th ACM SIGKDD*, pages 213–220, 2008.

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

[4] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori. Inlier-based outlier detection via direct density ratio estimation. In *ICDM 2008*, pages 223–232, 2008.

[5] T. Kanamori, S. Hido, and M. Sugiyama. Efficient direct density ratio estimation for non-stationarity adaptation and outlier detection. In *NIPS 21*, pages 809–816, 2009.

[6] T. Kanamori, S. Hido, and M. Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, Jul. 2009.

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

[8] W. Li, Q. Guo, and C. Elkan. A positive and unlabeled learning algorithm for one-class classification of remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 49(2):717–725, 2011.

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

[10] M. Sugiyama. Superfast-trainable multi-class probabilistic classifier by least-squares posterior fitting. *IEICE Transactions on Information and Systems*, E93-D(10):2690–2701, 2010.

[11] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.