Intelligent Data Analysis 25 (2021) 57–79
DOI 10.3233/IDA-194980
IOS Press

Classification from positive and unlabeled data based on likelihood invariance for measurement

Takeshi Yoshida, Takashi Washio*, Takahito Ohshiro and Masateru Taniguchi
The Institute of Scientific and Industrial Research, Osaka University, Osaka, Japan

Abstract. We propose novel approaches for classification from positive and unlabeled data (PUC) based on maximum likelihood principle. These are particularly suited to measurement tasks in which the class prior of the target object in each measurement is unknown and significantly different from the class prior used for training, while the likelihood function representing the observation process is invariant over the training and measurement stages. Our PUCs effectively work without estimating the class priors of the unlabeled objects. First, we present a PUC approach called Naive Likelihood PUC (NL-PUC) using the maximum likelihood principle in a nontrivial but rather straightforward manner. The extended version called Enhanced Likelihood PUC (EL-PUC) employs an algorithm iteratively improving the likelihood estimation of the positive class. This is advantageous when the availability of the labeled positive data is limited. These characteristics are demonstrated both theoretically and experimentally. Moreover, the practicality of our PUCs is demonstrated in a real application to single molecule measurement.

Keywords: Semi-supervised learning, classification from positive and unlabeled data, maximum likelihood estimation, likelihood invariance, application to measurement

1. Introduction

Classification from positive and unlabeled data (PUC) [7,27] has been actively researched in recent years. PUC is required in many problems where large amounts of unlabeled data are easily acquired but only limited quantities of one-class-labeled data are available. Many measurement tasks in scientific, industrial, and social sensing face such problems.

For example, the reinforced concrete in various real-world environments is classified as damaged or undamaged by ultrasonic testing [12]. The measurements of actually damaged samples are acquired from a limited number of past incidents, while many non-incident samples, which are either damaged or undamaged, are acquired without their labels. Similar situations are common in the field of industrial nondestructive inspection. Another example is the problem of classifying the risks of an individual suffering heart failure within the next 10 years based on cardiac function indices such as stroke volume (SV) and fractional shortening (%FS) which are geometrically measured by ultrasonic imaging [16]. Occurrences of heart failure are only recorded for persons who are actually hospitalized within the 10 years after their heart inspections, whereas most of those who are inspected but did not experience the

*Corresponding author: Takashi Washio, The Institute of Scientific and Industrial Research, Osaka University, 8-1, Mihogaoka, Ibarakishi, Osaka, 567-0047, Japan. Tel.: +81 6 6879 8540; Fax: +81 6 6879 8544; E-mail: washio@ar.sanken.osaka-u.ac.jp.

heart failure remain unlabeled. Similar problems appear in other medical fields. As described later, some advanced scientific measurement such as single molecule measurement also includes the similar problem.

The problem setting underlying many measurement tasks, including these examples, is that we have a large data set D_U of unlabeled samples X and a small data set D_P of other samples X having positive labels $Y = P$ such as “damaged” and “heart failure.” $X \in D_P$ is drawn from the positive marginal density $p(X|Y = P)$, and $X \in D_U$ is from the marginal density:

$$p_D(X) = \pi_D p(X|Y = P) + (1 - \pi_D)p(X|Y = N),$$

where π_D is the positive class prior probability $p_D(Y = P)$. PUC learns a classifier from D_P and D_U, and this classifier can be used to obtain the measurement consequence.

A distinguishing feature of the measurement tasks is that $p(X|Y = P)$ and $p(X|Y = N)$ are defined by the sensing devices as likelihood functions of Y; these functions are mostly invariant because the sensing devices are designed to preserve their characteristics for long periods to ensure high accuracy. In the aforementioned two examples, the ultrasonic sensors provide invariant likelihood functions on the concrete and the heart motion as long as they are well maintained. Another distinguishing feature of the measurement tasks is that the class prior probability of an individual target object or a set of target objects is unknown in each measurement and can be significantly different from the class prior probabilities of the objects targeted in other measurements. For instance, the prior probability of concrete damage deviates according to the real-world environment surrounding the individual concrete object. The prior probability of the heart failure is heavily dependent on the individual’s living environment. Thus, the problem setting underlying the measurement tasks can be more precisely stated as being that a newly measured sample follows the new marginal density

$$p_M(X) = \pi_M p(X|Y = P) + (1 - \pi_M)p(X|Y = N),$$

where an unknown prior probability π_M $(= p_M(Y = P))$ can be very different from π_D of D_U used for the PUC training.

These arguments reveal difficulties to apply conventional PUCs to measurement problems [7,27].

First, all past PUCs learn maximum a posteriori (MAP) classifiers in the standard setting of machine learning. A MAP classifier estimates the label Y assuming that $p(Y|X) \propto p(X|Y)p(Y) = p(X, Y)$ is unchanged over the training and test (measurement) stages, and thus does not anticipate a large difference between π_D and π_M. This nature does not fit the aforementioned problem setting of the measurement tasks.

Second, all past PUCs require the availability of the class prior probabilities π_D and π_M before the training and test (measurement) stages [27] or a special labeling mechanism in the training data acquisition [7]. These requirements also do not meet the conditions of the measurement tasks.

In measurement tasks having the invariant $p(X|Y)$ and the large discrepancies between π_D and unknown π_M, the maximum likelihood estimation of Y from X relying on the invariance of $p(X|Y)$ has been widely employed to avoid the above difficulty of the MAP approach, when both positive and negative labeled data sets: D_P and D_N are available for estimating $p(X|Y)$ [28]. However, to the best of our knowledge, there is no principle for accurately estimating $p(X|Y)$ from positive and unlabeled data sets: D_P and D_U without knowing π_D. If maximum likelihood PUCs based on such a principle could be established, they would enable various new measurement tasks.

The contributions of the present study are as follows.

1. We propose the novel principle of maximum likelihood PUCs to address the aforementioned problems. These are the first PUCs that learn from the independently acquired two sample datasets: D_P and D_U and that do not require the estimation of the class prior probability.

2. The proposed PUCs are accurate even when a small labeled data set D_P and a massive unlabeled data set D_U are given, and thus have a practical advantage.
3. These characteristics and the statistical accuracies of the proposed PUCs are theoretically demonstrated.
4. We further show a feasible performance measure of the PUCs requiring positive and unlabeled data sets only. This measure can be used to select appropriate parameters for the PUCs in cross validation.
5. The performance of our proposed PUCs is confirmed through numerical experiments using artificial and benchmark data and a real-world application to noise reduction in advanced single molecule measurement.

2. Problem setting

Here, we define our problem setting more precisely. Let $X \in \mathcal{X} \subseteq \mathbb{R}^d$ $(d \in \mathbb{N})$ and $Y \in \{P, N\}$ be a sample in its domain $\mathcal{X}$ and a positive (P) or negative (N) label, respectively. Given a training data set D_U of unlabeled samples X and a training data set D_P of other samples X having positive labels $Y = P$, let M_U be an unlabeled test data set given by measuring new objects. M_U may include only one sample, when the target object is unique. Samples in D_P are i.i.d. sampled from the positive marginal density $p_D(X|Y = P)$, and samples in D_U and M_U are i.i.d. sampled from marginal densities $p_D(X)$ and $p_M(X)$, respectively. No negative data are given.

We now introduce two assumptions on $p(X|Y)$.

Assumption 1. $p_D(X|Y = P)$, $p_D(X)$ and $p_M(X)$ share an invariant $p(X|Y)$. □

Assumption 2. $p(X|Y)$ is positive and has bounded and continuous second derivatives on $\mathcal{X}$. □

These are not special assumptions. All past PUCs assumed a common $p(X|Y)$ over all data sets [7,27]. All past classifiers using a non-parametric estimation of the probability densities in a continuous space rely on Assumption 2. This assumption is used in a later section. From the first assumption, $p_D(X|Y = P) = p(X|Y = P)$ holds, and $p_D(X)$, $p_M(X)$ consist of the common $p(X|Y)$ and the positive class prior probabilities $\pi_D = p_D(Y = P)$ and $\pi_M = p_M(Y = P)$, respectively, as follows.

$$p_D(X) = \pi_D p(X|Y = P) + (1 - \pi_D)p(X|Y = N), \tag{1}$$

$$p_M(X) = \pi_M p(X|Y = P) + (1 - \pi_M)p(X|Y = N). \tag{2}$$

$\pi_D, \pi_M \in (0, 1)$ are unknown and independently given.

Our problem is to learn an accurate non-parametric classifier of unseen samples in the unlabeled test data set M_U using only D_P and D_U. This is a two sample setting of PUC [27], whereas D_U and M_U follow mutually different unknown class prior probabilities.

3. Related work

There are two issues in tackling this problem. The first relates to the distinguishing feature of the measurement tasks such that no information about M_U is known before classifying a sample in M_U while its class prior π_M can be significantly different from π_D of D_U. In the field of traditional classifiers that learn from both positive and negative labeled data, many approaches for adapting the classifiers to slight or gradual concept drift [8] and larger differences between the training and test data sets [19,20] have been proposed. Classical semi-supervised classifiers that learn from small positive and negative labeled data sets and a large unlabeled data set are based on the transductive learning framework [1,29], and the adaptation of the classifiers to different class prior distributions has been studied. Saerens et al. proposed a covariate shift adaptation for the difference in the class prior in the semi-supervised setting [22]. More recent study has proposed a transductive transfer learning method to estimate the change in the class prior under the same setting [5]. In the PUC setting, most of the conventional PUCs are unsuitable because they assumes an invariant $p(X,Y)$ as mentioned earlier. An exception is the PUC proposed by Li et al. which adapts to gradual changes in micro-cluster labels in a data stream [15]. However, all these approaches rely on the availability of a certain size of M_U before classifying its samples or a gradual change in the class prior distribution of M_U, and they are therefore not applicable to our problem.

The second issue comes from the fact that the values of π_D and π_M are unknown, while the training processes of many PUCs need these values. The past PUCs use one of the following two options to overcome this difficulty. The PUC of Elkan and Noto [7] uses an option called the one sample setting which assumes that π_D and π_M are identical, positive samples drawn from $p_D(X)$ are labeled with a constant probability and included in D_P, and the remaining unlabeled samples are included in D_U [4]. This PUC implicitly estimates the class priors by partially matching the distribution of D_P to that of D_U. However, these assumptions are not always valid in measurement tasks where D_P and D_U are independently given and π_D and π_M are largely different. Some PUCs use the other option called the two sample setting which accepts independently acquired D_P and D_U [6,18,27]. This setting assumes that π_D and π_M are correctly estimated from D_P, D_U and M_U by relying on irreducibility of $p(X|Y)$ that $p(X|Y=P)$ is not represented by a linear combination of $p(X|Y=N)$ and another probability density function of X [2,6,21,23]. However, this irreducibility is not generally guaranteed in measurement tasks. Moreover, M_U is not usually available for estimating π_M before the new measurement, and so cannot be applied to our problem setting.

In the next section, we propose two maximum likelihood PUCs not requiring M_U in advance nor the estimation of π_D and π_M. Thus, they do not suffer the difficulties for the measurement tasks.

4. Proposed method and theoretical analysis

4.1. Naive Likelihood PUC (NL-PUC)

To design a PUC that is robust to differences between π_D and π_M, we employ the following maximum likelihood measure that does not depend on the class prior probabilities. According to Assumption 1, the maximum likelihood estimation of Y under $x \in M_U$ is given as follows:

$$y = \begin{cases} P \text{ if } p(x|Y=P) \geqslant p(x|Y=N), \\ N \text{ if } p(x|Y=P) < p(x|Y=N). \end{cases} \tag{3}$$

Associated with this measure, the following lemma holds.

Lemma 1. Given a marginal density

$$p_\pi(X) = \pi p(X|Y=P) + (1-\pi)p(X|Y=N),$$

the following inequalities are equivalent for any $\pi \in (0,1)$ and X.

$$p(X|Y=P) \geqslant p_\pi(X) \Leftrightarrow p(X|Y=P) \geqslant p(X|Y=N).$$

□

Proof. We obtain the following relation by substituting the definition of $p_\pi(X)$ to $p(X|Y=P) \geqslant p_\pi(X)$.

$$\begin{aligned}&p(X|Y=P) \geqslant \pi p(X|Y=P) + (1-\pi)p(X|Y=N)\\ &\Leftrightarrow (1-\pi)p(X|Y=P) \geqslant (1-\pi)p(X|Y=N)\\ &\Leftrightarrow p(X|Y=P) \geqslant p(X|Y=N) \text{ s.t. } \pi \in (0,1).\end{aligned}$$

□

From this lemma and Assumption 1, we immediately obtain the following, which is our core theorem.

Theorem 1. For any combination of π_D and π_M in $(0,1)$, the following formula is the maximum likelihood measure equivalent to Eq. (3) for all $x \in M_U$ following $p_M(X)$.

$$y = \begin{cases} P \text{ if } p(x|Y=P) \geqslant p_D(x), \\ N \text{ if } p(x|Y=P) < p_D(x). \end{cases} \tag{4}$$

□

Proof. From Eqs (1) and (2) and Lemma 1, the following holds:

$$\begin{aligned}&p(X|Y=P) \geqslant p_D(X) \Leftrightarrow p(X|Y=P) \geqslant p_M(X)\\ &\Leftrightarrow p(X|Y=P) \geqslant p(X|Y=N).\end{aligned}$$

Because the last inequality is independent of π_D and π_M according to Assumption 1, these three inequalities equivalently and invariantly hold for any π_D and π_M in $(0,1)$. Equation (4) is therefore equivalent to Eq. (3), and the theorem holds. □

We construct a maximum likelihood PUC by substituting the non-parametric estimations $\hat{p}(x|Y=P) := \hat{p}_D(x|Y=P)$ and $\hat{p}_D(x)$ derived from D_P and D_U, respectively, into this measure. They are provided by

$$\hat{p}_D(x|Y=P) = \frac{1}{|D_P|}\sum_{x \in D_P} p_K(X|x) \text{ and } \hat{p}_D(x) = \frac{1}{|D_U|}\sum_{x \in D_U} p_K(X|x),$$

where we use kernel densities $p_K(X|x)$ such as Gaussian kernel which well approximates $p(x|Y=P)$ and $p_D(x)$ under the condition of Assumption 2. We refer to this classification as Naive Likelihood PUC (NL-PUC).

4.2. Enhanced Likelihood PUC (EL-PUC)

Although NL-PUC satisfies our requirements, the classification may be degraded by large statistical errors in $\hat{p}_D(X|Y=P)$ induced from the small size of D_P, i.e., $|D_P|$. To alleviate this difficulty, we estimate $\hat{p}(x|Y=P) := \hat{p}^{(k)}(X|Y=P)$ more accurately by iteratively improving $\tilde{p}^{(k-1)}(X|Y=P)$ using its linear combination with $\hat{p}_D(X|Y=P)$.

$$\begin{aligned}&\hat{p}^{(k)}(X|Y=P)\\ &:= (1-r)\hat{p}_D(X|Y=P) + r\tilde{p}^{(k-1)}(X|Y=P).\end{aligned} \tag{5}$$

where $r \in (0,1)$ and $k = 2, 3, \ldots$. $\tilde{p}^{(k-1)}(X|Y=P)$ is given by the following non-parametric approximation of $p(X|Y=P)$ using kernel densities $p_K(X|x)$ and their weights $\tilde{w}^{(k-1)}(x)$. We use $p_K(X|x)$ such as Gaussian kernel as well as in NL-PUC.

$$\tilde{p}^{(k-1)}(X|Y=P) = \sum_{x \in D_U} \tilde{w}^{(k-1)}(x)\, p_K(X|x). \tag{6}$$

$$\tilde{w}^{(k-1)}(x) = \frac{\hat{p}^{(k-1)}(x|Y=P)/\hat{p}_D(x)}{\sum_{x'\in D_U}\hat{p}^{(k-1)}(x'|Y=P)/\hat{p}_D(x')}. \tag{7}$$

Their initial values are set as follows.

$$\begin{aligned} &\tilde{p}^{(1)}(X|Y=P) = \sum_{x\in D_U}\tilde{w}(x)p_K(X|x) \\ &\text{s.t. } \tilde{w}(x) = \frac{\hat{p}_D(x|Y=P)/\hat{p}_D(x)}{\sum_{x'\in D_U}\hat{p}_D(x'|Y=P)/\hat{p}_D(x')}. \end{aligned} \tag{8}$$

Equation (8) is a weighted non-parametric approximation of $p(X|Y=P)$ using the samples in D_U and the normalized weights $\tilde{w}(x)$ where D_U follows $p_D(X)$ and $\tilde{w}(x)$ approximates $w(x)$ defined by $p_D(x|Y=P) \propto w(x)p_D(x)$ [10,26]. Equations (6) and (7) further repeat the weighted non-parametric approximation of $p(X|Y=P)$ using its estimation from the previous step.

In the next subsection, we show that, if D_U is of sufficient size and we choose an appropriate r, the mean square error of $\hat{p}^{(k)}(X|Y=P)$ can be no more than that of the non-parametric approximation $\hat{p}_D(X|Y=P)$ directly derived from D_P. Moreover, the mean square error of $\hat{p}_D(X|Y=P)$ is limited under Assumption 2, because $p(X|Y=P)$ is positive and has bounded and continuous second derivatives on $\mathcal{X}$ [10,24]. Accordingly, $\hat{p}^{(k)}(X|Y=P)$ converges to a value having better accuracy than $\hat{p}_D(X|Y=P)$. In practice, our numerical experiments show that $\hat{p}^{(k)}(X|Y=P)$ using almost any $r \in (0,1)$ widely gives a better estimation of $p(X|Y=P)$ than $\hat{p}_D(X|Y=P)$.

When the rate of change of $\tilde{w}^{(k-1)}(x)$ converges within a small tolerance $\varepsilon = 10^{-4}$ for all $x \in D_U$, we obtain a more accurate $\hat{p}^{(k)}(X|Y=P)$ than $\hat{p}_D(X|Y=P)$ given by NL-PUC, and apply $\hat{p}^{(k)}(X|Y=P)$ to Eq. (4) with $\hat{p}_D(X)$. This is called Enhanced Likelihood PUC (EL-PUC) and expected to have better accuracy particularly when $|D_P|$ is small.

4.3. Theoretical accuracy of EL-PUC

$p_D(X)$ is positive and has bounded and continuous second derivatives on $\mathcal{X}$ from Assumption 2, Eq. (1) and $\pi_D \in (0,1)$. Moreover, the size of D_U is sufficiently large in most practical cases. Accordingly, the non-parametric estimation $\hat{p}_D(X)$ given by D_U and used on the r.h.s. of Eq. (4) has a limited mean square error [10,24]. Thus, we focus on the error of $\hat{p}^{(k)}(X|Y=P)$ which is used on the l.h.s. of Eq. (4).

Let $\hat{q}(X)$ be the estimation of $q(X)$, $\mathrm{B}[\hat{q}(X)]$ be the absolute value of the bias $|\mathrm{E}[\hat{q}(X)-q(X)]|$, $\mathrm{V}[\hat{q}(X)]$ be the variance $\mathrm{E}[\{\hat{q}(X)-\mathrm{E}[\hat{q}(X)]\}^2]$, and $\mathrm{MSE}[\hat{q}(X)]$ be the mean square error $\mathrm{B}[\hat{q}(X)]^2 + \mathrm{V}[\hat{q}(X)]$. Here, $\mathrm{E}[\cdot]$ denotes the expectation over the distribution $p(D_P \cup D_U)$.

Lemma 2. Let $\Delta^{(k)}(X) = |\mathrm{E}[\tilde{p}^{(k)}(X|Y=P)] - \mathrm{E}[\hat{p}^{(k)}(X|Y=P)]|$, then $\Delta_{sup}(X) = \sup_{k\geqslant 1}\Delta^{(k)}(X)$ is bounded, and the following holds for $k \to \infty$.

$$\mathrm{B}[\hat{p}^{(k)}(X|Y=P)] \leqslant \frac{r}{1-r}\Delta_{sup}(X) + \mathrm{B}[\hat{p}_D(X|Y=P)].$$

□

Proof. From Eqs (6) and (7), $\tilde{p}^{(k)}(X|Y=P)$ is a weighted non-parametric approximation of $\hat{p}^{(k)}(X|Y=P)$. As $\hat{p}^{(k)}(X|Y=P)$ in Eq. (5) is a linear combination of $\hat{p}_D(X|Y=P)$ and $\tilde{p}^{(k-1)}(X|Y=P)$ which are positive and have bounded and continuous second derivatives by the use of the kernel density $p_K(X|x)$ satisfying the condition of Assumption 2, the absolute expected error of $\tilde{p}^{(k)}(X|Y=P)$ against $\hat{p}^{(k)}(X|Y=P)$ is bounded [10,24]. Hence, $\Delta^{(k)}(X)$ is bounded for any k, i.e., $\Delta_{sup}(X)$ is bounded. Accordingly, by subtracting $\hat{p}_D(X|Y = P)$ from both sides of Eq. (5) and taking their absolute expectations, the following holds.

$$\begin{aligned}&|\mathrm{E}[\hat{p}^{(k)}(X|Y = P) - \hat{p}_D(X|Y = P)]|\\&= r|\mathrm{E}[\tilde{p}^{(k-1)}(X|Y = P) - \hat{p}_D(X|Y = P)]|\\&= r|\mathrm{E}[\hat{p}^{(k-1)}(X|Y = P)] \pm \Delta^{(k-1)}(X) - \mathrm{E}[\hat{p}_D(X|Y = P)]|\\&\leqslant r|\mathrm{E}[\hat{p}^{(k-1)}(X|Y = P) - \hat{p}_D(X|Y = P)]| + r\Delta^{(k-1)}(X).\end{aligned}$$

By repeatedly applying this formula and applying $\hat{p}^{(1)}(x|Y = P) := \hat{p}_D(x|Y = P)$ in Eq. (8) when $i = k - 1$, we derive:

$$\begin{aligned}&|\mathrm{E}[\hat{p}^{(k)}(X|Y = P) - \hat{p}_D(X|Y = P)]| \leqslant\\&\sum_{i=1}^{k-1} r^i \Delta^{(k-i)}(X) \leqslant \Delta_{sup}(X) \sum_{i=1}^{k-1} r^i \xrightarrow[k\to\infty]{} \frac{r}{1-r}\Delta_{sup}(X).\end{aligned}$$

Moreover, the following holds.

$$\begin{aligned}&\mathrm{B}[\hat{p}^{(k)}(X|Y = P)] = |\mathrm{E}[\hat{p}^{(k)}(X|Y = P) - p(X|Y = P)]|\\&\leqslant |\mathrm{E}[\hat{p}^{(k)}(X|Y = P) - \hat{p}_D(X|Y = P)]|\\&\quad+|\mathrm{E}[\hat{p}_D(X|Y = P) - p(X|Y = P)]|\\&= |\mathrm{E}[\hat{p}^{(k)}(X|Y = P) - \hat{p}_D(X|Y = P)]| + \mathrm{B}[\hat{p}_D(X|Y = P)].\end{aligned}$$

By substituting the former inequality into this last line, we obtain our following result for $k \to \infty$.

$$\mathrm{B}[\hat{p}^{(k)}(X|Y = P)] \leqslant \frac{r}{1-r}\Delta_{\mathrm{sup}}(X) + \mathrm{B}[\hat{p}_D(X|Y = P)].$$

□

This lemma shows that $\mathrm{B}[\hat{p}^{(k)}(X|Y = P)]$ can be less than $\mathrm{B}[\hat{p}_D(X|Y = P)]$, if r is sufficiently small.

Lemma 3. The following holds for all $k \geqslant 2$.

$$\begin{aligned}&\mathrm{V}[\hat{p}^{(k)}(X|Y = P)] \leqslant \{\mathrm{V}[\hat{p}_D(X|Y = P)]^{1/2}\\&+r(\mathrm{V}[\tilde{p}^{(k-1)}(X|Y = P)]^{1/2} - \mathrm{V}[\hat{p}_D(X|Y = P)]^{1/2})\}^2.\end{aligned}$$

□

Proof. By taking the expectation of Eq. (5) and subtracting it from Eq. (5) itself, we obtain the following for $k \geqslant 2$.

$$\begin{aligned}&\hat{p}^{(k)}(X|Y = P) - \mathrm{E}[\hat{p}^{(k)}(X|Y = P)]\\&= (1 - r)\{\hat{p}_D(X|Y = P) - \mathrm{E}[\hat{p}_D(X|Y = P)]\}\\&\quad+r\{\tilde{p}^{(k-1)}(X|Y = P) - \mathrm{E}[\tilde{p}^{(k-1)}(X|Y = P)]\}.\end{aligned}$$

By taking the mean square of both sides, the following formula is derived.

$$\begin{aligned}
&\mathrm{V}[\hat{p}^{(k)}(X|Y=P)] = (1-r)^2\mathrm{V}[\hat{p}_D(X|Y=P)] \\
&+2(1-r)r\underline{\mathrm{E}[(\hat{p}_D(X|Y=P) - \mathrm{E}[\hat{p}_D(X|Y=P)])} \\
&\qquad\qquad\quad \underline{\times(\tilde{p}^{(k-1)}(X|Y=P) - \mathrm{E}[\tilde{p}^{(k-1)}(X|Y=P)])]} \\
&+r^2\mathrm{V}[\tilde{p}^{(k-1)}(X|Y=P)].
\end{aligned}$$

According to the Cauchy-Schwarz inequality and the positivity of the variances, the underlined term is no more than $\mathrm{V}[\hat{p}_D(X|Y=P)]^{1/2}\mathrm{V}[\tilde{p}^{(k-1)}(X|Y=P)]^{1/2}$. Thus the following inequality is obtained, and the lemma holds.

$$\begin{aligned}
&\mathrm{V}[\hat{p}^{(k)}(X|Y=P)] \leqslant (1-r)^2\mathrm{V}[\hat{p}_D(X|Y=P)] \\
&\quad +2(1-r)r\mathrm{V}[\hat{p}_D(X|Y=P)]^{1/2}\mathrm{V}[\tilde{p}^{(k-1)}(X|Y=P)]^{1/2} \\
&\quad +r^2\mathrm{V}[\tilde{p}^{(k-1)}(X|Y=P)] \\
&= \{(1-r)\mathrm{V}[\hat{p}_D(X|Y=P)]^{1/2} + r\mathrm{V}[\tilde{p}^{(k-1)}(X|Y=P)]^{1/2}\}^2 \\
&= \{\mathrm{V}[\hat{p}_D(X|Y=P)]^{1/2} \\
&\quad +r(\mathrm{V}[\tilde{p}^{(k-1)}(X|Y=P)]^{1/2} - \mathrm{V}[\hat{p}_D(X|Y=P)]^{1/2})\}^2.
\end{aligned}$$

□

This lemma indicates that $\mathrm{V}[\hat{p}^{(k)}(X|Y=P)]$ can be less than $\mathrm{V}[\hat{p}_D(X|Y=P)]$, if r is sufficiently small. The following theorem is easily derived from these two lemmas.

Theorem 2. If the following holds, there exists some $r \in (0,1)$ such that $\mathrm{MSE}[\hat{p}^{(k)}(X|Y=P)] < \mathrm{MSE}[\hat{p}_D(X|Y=P)]$.

$$\begin{aligned}
&\mathrm{B}[\hat{p}_D(X|Y=P)]\Delta_{sup}(X) < \mathrm{V}[\hat{p}_D(X|Y=P)] \\
&\qquad -\mathrm{V}[\tilde{p}^{(k-1)}(X|Y=P)]^{1/2}\mathrm{V}[\hat{p}_D(X|Y=P)]^{1/2}.
\end{aligned} \tag{9}$$

□

Proof. $\mathrm{MSE}[\hat{p}^{(k)}(X|Y=P)]$ is bounded above by

$$\begin{aligned}
&\mathrm{MSE}_{sup}[\hat{p}^{(k)}(X|Y=P)] \\
&= \left\{\mathrm{B}[\hat{p}_D(X|Y=P)] + \frac{r}{1-r}\Delta_{sup}(X)\right\}^2 \\
&\quad + \Big\{\mathrm{V}[\hat{p}_D(X|Y=P)]^{1/2} \\
&\quad +r(\mathrm{V}[\tilde{p}^{(k-1)}(X|Y=P)]^{1/2} - \mathrm{V}[\hat{p}_D(X|Y=P)]^{1/2})\Big\}^2
\end{aligned}$$

according to Lemmas 2 and 3. Therefore,

$$\mathrm{MSE}_{sup}[\hat{p}^{(k)}(X|Y=P)] \xrightarrow[r\to 0^+]{} \mathrm{MSE}[\hat{p}_D(X|Y=P)],$$

$$\text{and } \mathrm{MSE}_{sup}[\hat{p}^{(k)}(X|Y=P)] \xrightarrow[r\to 1^-]{} \infty$$

hold. These facts imply that there exists some $r \in (0,1)$ such that MSE$[\hat{p}^{(k)}(X|Y = P)] <$ MSE$[\hat{p}_D(X|Y = P)]$, if the differential of MSE$_{sup}[\hat{p}^{(k)}(X|Y = P)]$ for r is negative in a vicinity of 0^+, i.e.,

$$\begin{aligned}
&\forall\epsilon > 0, \exists\delta \in (0,1), c < 0, \text{ s.t. } \forall r \in (0,\delta),\\
&\Rightarrow \left| \frac{d\text{MSE}_{sup}[\hat{p}^{(k)}(X|Y = P)]}{dr} - c \right| \leqslant \varepsilon.
\end{aligned}$$

The above condition is equivalent to the following, which proves the lemma.

$$\begin{aligned}
&\lim_{r\to 0^+} \frac{d\text{MSE}_{sup}[\hat{p}^{(k)}(X|Y = P)]}{dr}\\
&= 2\left\{\text{B}[\hat{p}_D(X|Y = P)] + \frac{r}{1-r}\Delta_{sup}(X)\right\}\frac{1}{(1-r)^2}\Delta_{sup}(X)\\
&\quad +2\left\{\text{V}[\hat{p}_D(X|Y = P)]^{1/2}\right.\\
&\quad \left.+r(\text{V}[\tilde{p}^{(k-1)}(X|Y = P)]^{1/2} - \text{V}[\hat{p}_D(X|Y = P)]^{1/2})\right\}\\
&\quad \times(\text{V}[\tilde{p}^{(k-1)}(X|Y = P)]^{1/2} - \text{V}[\hat{p}_D(X|Y = P)]^{1/2}) \text{ s.t. } r \to 0^+\\
&= 2\text{B}[\hat{p}_D(X|Y = P)]\Delta_{sup}(X)\\
&\quad +2\text{V}[\tilde{p}^{(k-1)}(X|Y = P)]^{1/2}\text{V}[\hat{p}_D(X|Y = P)]^{1/2}\\
&\quad -2\text{V}[\hat{p}_D(X|Y = P)] < 0.\\
&\Leftrightarrow \text{B}[\hat{p}_D(X|Y = P)]\Delta_{sup}(X) < \text{V}[\hat{p}_D(X|Y = P)]\\
&\quad -\text{V}[\tilde{p}^{(k-1)}(X|Y = P)]^{1/2}\text{V}[\hat{p}_D(X|Y = P)]^{1/2}.
\end{aligned}$$

□

The variances of $\hat{p}_D(X|Y = P)$ and $\tilde{p}^{(k-1)}(X|Y = P)$ obtained by non-parametric estimation using $|D_P|$ i.i.d. samples and $|D_U|$ weighted i.i.d. samples are known to be

$$\text{V}[\hat{p}_D(X|Y = P)] \simeq \frac{C}{|D_P|}p(X|Y = P), \text{and} \tag{10}$$

$$\text{V}[\tilde{p}^{(k-1)}(X|Y = P)] \simeq \frac{C}{|D_U|_{\text{eff}}^{(k-1)}}p(X|Y = P), \tag{11}$$

respectively [10,14,26]. C is a positive constant given by the integrated square of $p_K(X|x)$, and $|D_U|_{\text{eff}}^{(k-1)}$ is the effective size of D_U with weights $\tilde{w}(x)^{(k-1)}$ defined as

$$|D_U|_{\text{eff}}^{(k-1)} = \frac{1}{\sum_{x\in D_U}(\tilde{w}(x)^{(k-1)})^2} \text{ s.t. } \sum_{x\in D_U}\tilde{w}(x)^{(k-1)} = 1.$$

From Eqs (1) and (7), x in a certain fraction α of D_U have $\tilde{w}(x)^{(k-1)} > 1/|D_U|$, and thus $\sum_{x\in D_U}(\tilde{w}(x)^{(k-1)})^2 \geqslant \alpha|D_U|(1/|D_U|)^2 = \alpha/|D_U| \Rightarrow |D_U|_{\text{eff}}^{(k-1)} = O(|D_U|)$ holds. According to this fact, Eqs (10) and (11), the r.h.s. of Eq. (9) is large, i.e., Eq. (9) may hold, when $|D_P|$ is small and $|D_U|$ is large. Therefore, when there are few samples in D_P and many samples in D_U, MSE$[\hat{p}^{(k)}(X|Y = P)]$

is generally less than $\mathrm{MSE}[\hat{p}_D(X|Y = P)]$ if using an appropriate $r \in (0, 1)$, and EL-PUC is expected to attain higher accuracy than NL-PUC.

This theoretical result is more intuitively explained by the nature of statistical bias end variance. The expectations of $\tilde{p}^{(k)}(X|Y = P)$ and $\hat{p}^{(k)}(X|Y = P)$ do not mutually deviate significantly, since they are kernel based nonparametric estimations of $p(X|Y = P)$. Thus, $\Delta_{sup}(X)$ is small, and this ensures the small difference of the expected biases of $\hat{p}_D(X|Y = P)$ and $\hat{p}^{(k)}(X|Y = P)$ over the wide range of r in Lemma 2. Whereas, the variance of $\hat{p}_D(X|Y = P)$ is expected to be significantly larger than that of $\hat{p}^{(k)}(X|Y = P)$ under the condition of $|D_P| \ll |D_U|$, because they are dominated by $|D_P|$ and $|D_U|$, respectively. In total, the mean square error of $\hat{p}^{(k)}(X|Y = P)$ is less than that of $\hat{p}_D(X|Y = P)$.

4.4. *Performance measure and parameter selection*

When a certain amount of test data is available, the performance of our PUCs can be empirically evaluated for the purpose of comparison with other candidate PUCs and to tune the PUC parameters. However, standard performance indices such as AUC and F-measure are not applicable, because the test data sets in practical problems targeted by PUCs usually do not include labeled negative samples.

For such circumstances, a number of studies have proposed performance indices similar to the traditional AUC that assess the goodness of the classifier classes, but not particular parameter selections of the classifiers [9,11,17]. Only a few studies have proposed indices that are directly applicable to parameter selection. Lee and Liu proposed an index equivalent to the geometric mean of precision and recall in the one sample setting [13]. Calvo et al. proposed a generic pseudo F-measure that matches the traditional F-measure if we substitute the correct π_M [3]. We employ the latter index because it is applicable to the two sample setting.

The unavailability of π_M in our problems presents an issue. Given a labeled positive test data set M_P and a unlabeled test data set M_U with Assumption 1, assuming that both M_P and M_U have certain sizes for statistically stable evaluations, we derive the following lemma for our likelihood based PUC which uses Eq. (3) or equivalent Eq. (4). It ensures that the pseudo F-measure using an arbitrary $\tilde{\pi}$ in place of π_M is a maximum if and only if the PUC correctly classifies all samples in M_U. In the case where M_P is not given, we use a subset $D'_P \subset D_P$ as M_P under Assumption 1 and use $D_P \setminus D'_P$ to train the PUC. In this study, we use a moderate $\tilde{\pi} = 0.5$.

Lemma 4. Let M^P_U be the set of all positive samples in M_U, and let $\hat{M}_P$ and $\hat{M}^P_U$ be the sets of all samples classified as positive in M_P and M_U, respectively, by using Eq. (3) or equivalent Eq. (4). The following pseudo F-measure [3] is expected to be a maximum for any $\tilde{\pi} > 0$, if and only if $\hat{M}^P_U = M^P_U$.[1]

$$\tilde{F} = \frac{2\tilde{\pi}|\hat{M}_P|/|M_P|}{|\hat{M}^P_U|/|M_U| + \tilde{\pi}}$$

□

Proof. By Assumption 1, the positive samples in M^P_U and M_P have identical $p(X|Y = P)$. Thus, the probability of M^P_U that a positive sample in M^P_U is correctly classified as positive by applying Eq. (3) or equivalent Eq. (4) is identical with that of M_P. In addition, $\hat{M}_P \subseteq M_P$ always holds. Accordingly, $E[|M^P_U \cap \hat{M}^P_U|/|M^P_U|] = E[|M_P \cap \hat{M}_P|/|M_P|] = E[|\hat{M}_P|/|M_P|]$, and the following holds.

$$\tilde{F} \simeq \frac{2\tilde{\pi}|M^P_U \cap \hat{M}^P_U|/|M^P_U|}{|\hat{M}^P_U|/|M_U| + \tilde{\pi}}.$$

[1] M^P_U is unknown since M_U is unlabeled.

When $|\hat{M}_U^P| \geqslant |M_U^P|$, the denominator is positive under $\tilde{\pi} > 0$ and attains a minimum when $|\hat{M}_U^P| = |M_U^P|$, because the denominator monotonically increases on $|\hat{M}_U^P| \in [|M_U^P|, |M_U|]$. The numerator with $\tilde{\pi} > 0$ attains a maximum when $M_U^P \subseteq \hat{M}_U^P$ where $|M_U^P \cap \hat{M}_U^P|$ is a maximum for any $|\hat{M}_U^P|$ ($\geqslant |M_U^P|$). Thus, $\tilde{F}$ is a maximum when $\hat{M}_U^P = M_U^P$.

Similarly, when $|\hat{M}_U^P| < |M_U^P|$, the numerator with $\tilde{\pi} > 0$ is a maximum for any $|\hat{M}_U^P|$ ($< |M_U^P|$) if $\hat{M}_U^P \subseteq M_U^P$ where $\tilde{F}$ is represented as

$$\tilde{F} \simeq \frac{2\tilde{\pi}/|M_U^P|}{1/|M_U| + \tilde{\pi}/|\hat{M}_U^P|},$$

which is monotonically increasing on $|\hat{M}_U^P| \in [0, |M_U^P|]$ under $\tilde{\pi} > 0$. Thus, $\tilde{F}$ is maximized when $\hat{M}_U^P = M_U^P$.

By combining these two cases, we conclude that, for any $\tilde{\pi} > 0$, $\tilde{F}$ is maximized if and only if $\hat{M}_U^P = M_U^P$. □

The analysis in Subsection 4.3 indicated that the parameter r in Eq. (5) affects the accuracy of the EL-PUC. We set up EL-PUC using two candidate values of r. EL-PUCdf uses arbitrary default $r = 0.5$ which is a moderate value in the interval (0,1). Another EL-PUCcv uses r selected by cross validation using $\tilde{F}$. In the first stage of EL-PUCcv, we apply 10CV to select the best r by randomly splitting D_P and D_U into 10% fractions: D'_P and D'_U for testing, i.e., M_P and M_U, and remaining 90% fractions: $D_P \setminus D'_P$ and $D_U \setminus D'_U$ for training in each fold, respectively.[2] Finally, EL-PUCcv is trained using the entire D_P and D_U and the selected parameter.

We use a Gaussian kernel density with kernel width h for the non-parametric estimation of the probability density functions required in NL-PUC and EL-PUC. Wang empirically showed that the theoretical value of h that minimizes the mean integrated square error of the un-weighted non-parametric estimator is also optimal in case of a weighted non-parametric estimator, and called this a plug-in estimator [26]. We use h of this plug-in estimator.

5. Experimental evaluation

Using both artificial and UCI data sets, we evaluated the accuracy of NL-PUC, EL-PUCdf and EL-PUCcv and their robustness to differences between π_M and π_D. The results were compared with those given by Elkan & Noto's PUCs [7] using Gaussian Naïve Bayes based (NB-E&N) and Gaussian kernel density based (KD-E&N) estimators of $\hat{p}_D(X|Y = P)$ and $\hat{p}_D(X)$. As with our proposed PUCs, we applied the kernel width h of the plug-in estimator to KD-E&N. We did not include other PUCs for the two sample setting in comparison [6,15,18,27], because they require M_U to estimate π_M, and this is not available in advance in our problem setting.

The codes of the five PUCs were implemented using MATLAB and run in Windows 10 PC equipped with Intel Core i7-4790-3.6GHz-8 cores CPU and 32GB RAM.

5.1. Performance evaluation using artificial data

We used artificial data containing $p(X|Y = P)$ and $p(X|Y = N)$ which are two Gaussians having a common covariance $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and distinct means $\mu_P = (0, 0)$, $\mu_N = (2, 0)$. Given π_D and π_M,

[2] We use a line search to seek the best r.

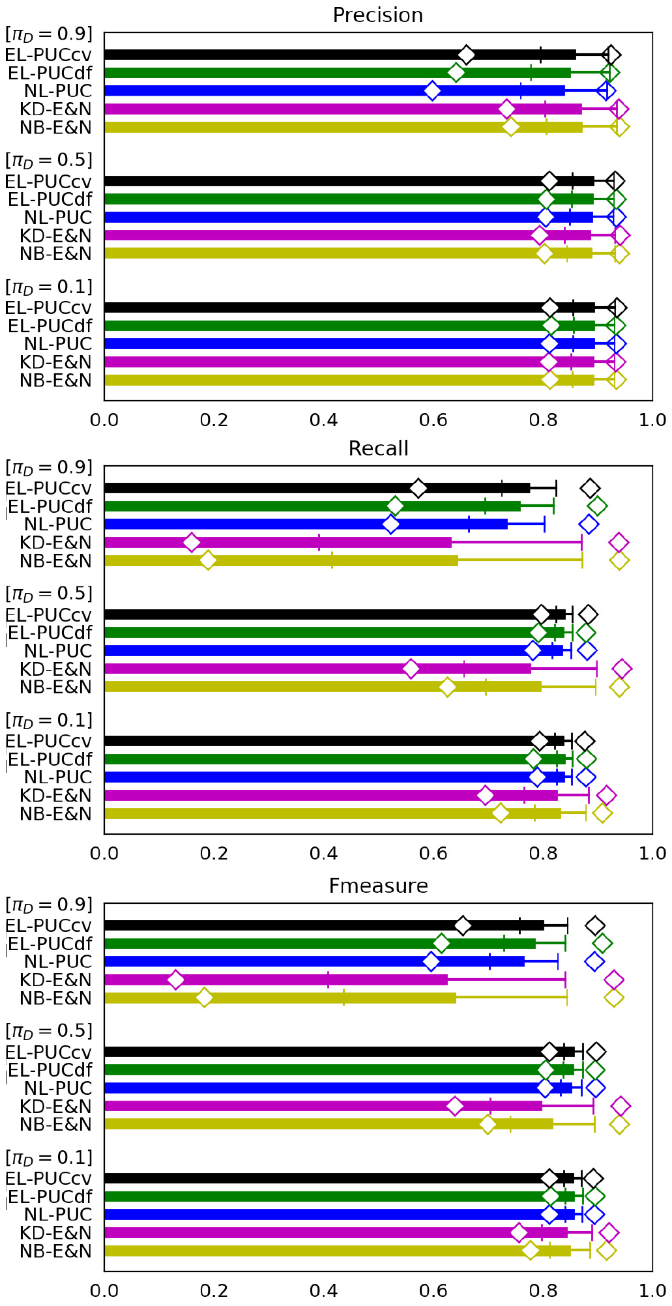


Fig. 1. Performance for artificial data having three different $\pi_D = 0.1$, 0.5 and 0.9 ($|D_P| = 1000$, $|D_U| = 1000$, $|M_U| = 1000$ and $\pi_M \in \{0.1, 0.5, 0.9\}$).

the three data sets D_P, D_U and M_U were generated by i.i.d. sampling from $p_D(X|Y = P)$, $p_D(X)$ in Eq. (1) and $p_M(X)$ in Eq. (2), respectively.

Figure 1 shows comparisons of precision, recall and F-measure between the five PUCs in terms of the three class prior probabilities of D_U: $\pi_D = 0.1$, 0.5 and 0.9. Under given π_D with fixed data sizes

$|D_P| = 1000$, $|D_U| = 1000$ and $|M_U| = 1000$, we generated 100 cases of $\{D_P, D_U, M_U\}$ for each class prior probability of M_U: $\pi_M \in \{0.1, 0.5, 0.9\}$,i.e., 300 cases in total. The respective PUC trained using D_P and D_U is tested by using M_U to measure its performance indecies in each case. A bar, a whisker and a diamond shape of a PUC in the figures represent the mean, the standard deviation and the minimum/maximum of the performance index over the 300 cases covering the three $\pi_M \in \{0.1, 0.5, 0.9\}$, respectively. The mean represents the accuracy of a PUC over the difference of the class prior probabilities between D_U and M_U. The less standard deviation and the less interval between the minimum and the maximum show the higher robustness of a PUC to the difference.

In these figures, the accuracy and robustness of all PUCs mostly decrease with increasing π_D, because the negative samples in D_U become less while they are the unique information source on $p(X|Y = N)$. NB-E&N and KD-E&N particularly lose their recall in this condition, since they use classifiers of labeled and unlabeled samples which do not capture the difference between the labeled positive samples and the unlabeled negative samples when almost all samples are positive. Nevertheless, all the three indecies show high accuracy and higher robustness of our three PUCs than NB-E&N and KD-E&N in almost all conditions. This comes from the fact that the classification measures Eqs (3) and (4) used in our PUCs are independent of the class prior probabilities. The accuracy of EL-PUCdf and EL-PUCcv is particularly higher than that of NL-PUC under $\pi_D = 0.9$. This may be because $\mathrm{V}[\tilde{p}^{(k-1)}(X|Y = P)]$ is kept small by estimating $\tilde{p}^{(k-1)}(X|Y = P)$ from both D_P and D_U, while $\mathrm{B}[\hat{p}_D(X|Y = P)]$ is small as long as Assumption 2 holds on a continuous space $\mathcal{X}$ [10,24]. These effects make Eq. (9) in Theorem 2 to widely hold even under the large π_D where the performance of NL-PUC is degraded by the shortage of the information on $p(X|Y = N)$.

Figure 2 shows comparison of the three indices between the five PUCs over $|D_P| = 10$, 100, 1000 and 10000. As in Fig. 1, we generated 300 cases over $\pi_M = 0.1$, 0.5 and 0.9 under given $|D_P|$, $|D_U| = 1000$ with fixed $\pi_D = 0.5$ and $|M_U| = 1000$ in an experiment, and measured the mean, the standard deviation and the minimum/maximum of the three indices of each PUC over the 300 cases covering the three $\pi_M \in \{0.1, 0.5, 0.9\}$, respectively. Figure 3 was created for the comparison over $|D_U| = 100$, 1000 and 10000 under fixed $\pi_D = 0.5$, $|D_P| = 1000$ and $|M_U| = 1000$ in similar manner. In these figures, our three PUCs again show higher accuracy and robustness than NB-E&N and KD-E&N in the most conditions including the case of $D_P = 10$ in Fig. 2. Elkan & Noto's PUCs show larger standard deviations and intervals between the minimum and the maximum of the accuracy. Particularly, they show significant decreases of the accuracy in the imbalanced conditions between $|D_P|$ and $|D_U|$ such as the cases of $|D_P| = 10$ and 10000 in Fig. 2, and $D_U = 100$ in Fig. 3. This is because they use classifiers of labeled and unlabeled samples which are statistically degraded under the highly imbalanced conditions. The accuracy of EL-PUCdf and EL-PUCcv is particularly higher than that of NL-PUC when a small size $|D_P| = 10$ and a large size $|D_U| = 1000$ are provided in Fig. 2. This character is consistent with the suggestion of Theorem 2 and (10) and (11).

Figures 4a and b represent the dependency of the computation times required to train the five PUCs using the artificial data in log-log plots. The computation times under given $|D_P|$ were averaged over all combinations of $|D_U|$ and π_D in the former, and those under given $|D_U|$ were averaged over all combinations of $|D_P|$ and π_D in the latter. The line of "EL-PUCcv" shows its computation times after the tuning of the parameter r, and the line of "r tuning time" indicates its times required for the tuning by the 10 fold cross validation. Since NL-PUC simply learns Eq. (4), its complexity for training is $O(|D_P| + |D_U|)$. Both EL-PUCdf and EL-PUCcv have their complexity of $O(K(r)|D_U|(|D_P| + |D_U|))$ where $K(r)$ is the number of the iteration of Eqs (5)–(7). In each iteration step, the computation of $\tilde{w}^{(k-1)}(x)$ is repeated $|D_U|$ times in Eq. (6) where each $\tilde{w}^{(k-1)}(x)$ is computed by Eqs (5) and (7)

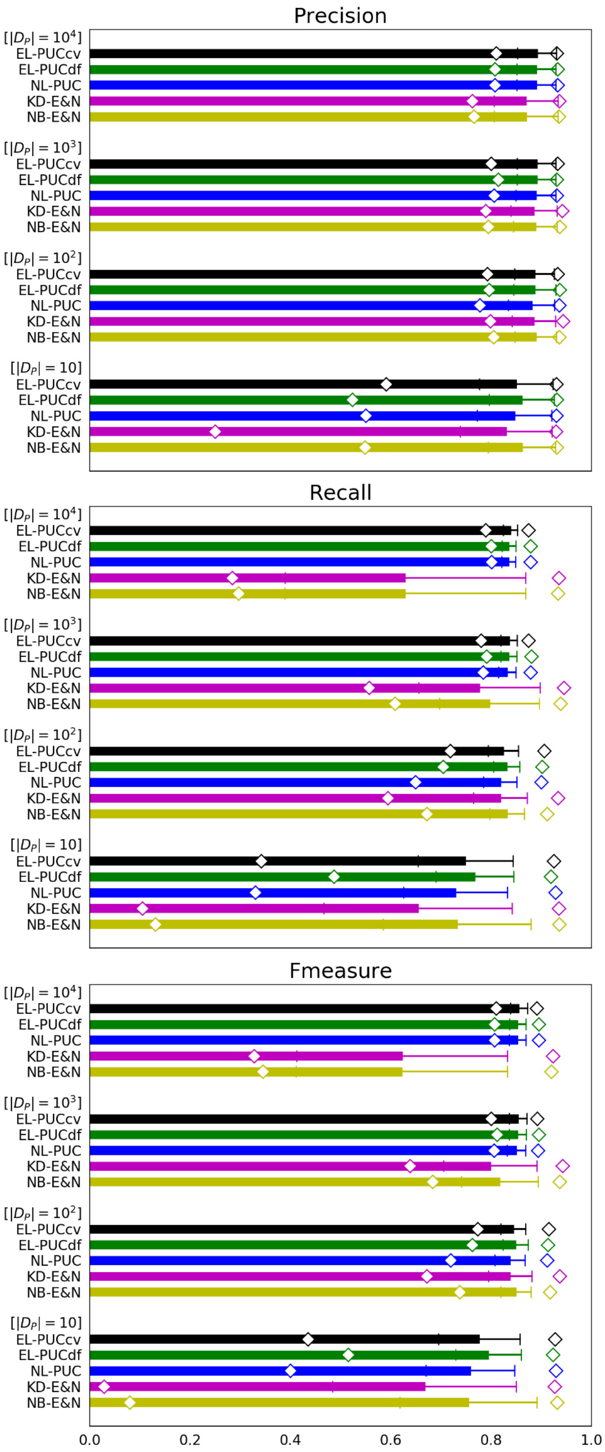


Fig. 2. Performance for artificial data having various $|D_P| = 10$, 100, 1000 and 10000 ($\pi_D = 0.5$, $|D_U| = 1000$, $|M_U| = 1000$ and $\pi_M \in \{0.1, 0.5, 0.9\}$).

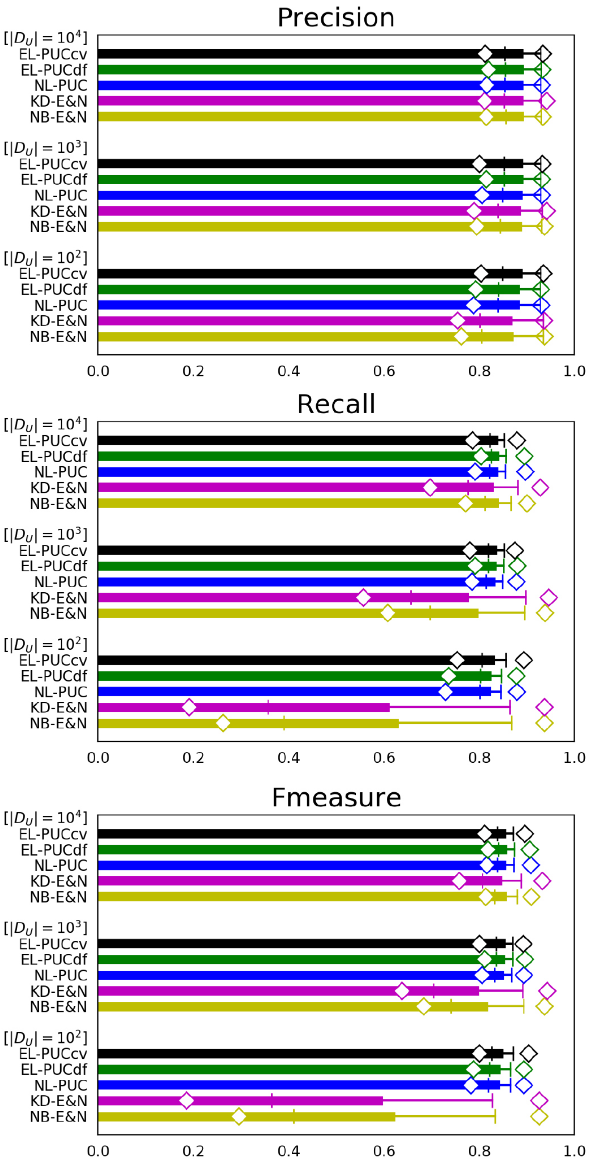
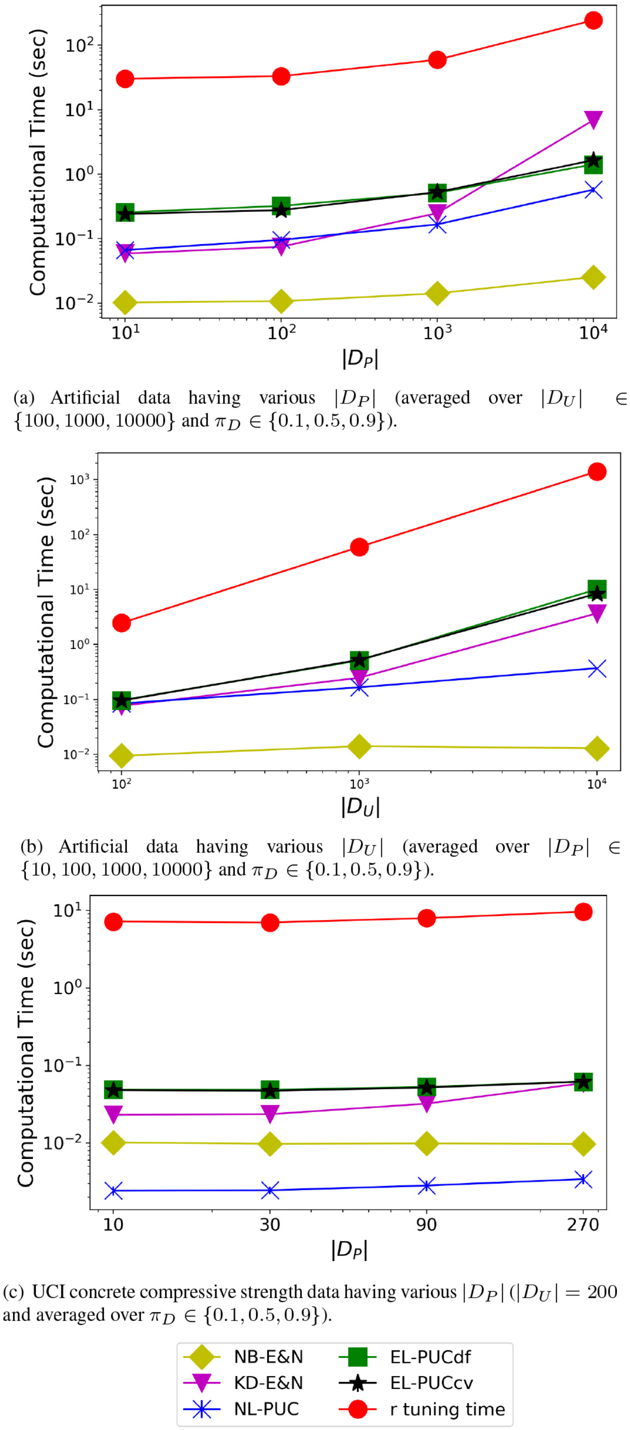


Fig. 3. Performance for artificial data having various $|D_U| = 100$, 1000 and 10000 ($\pi_D = 0.5$, $|D_P| = 1000$, $|M_U| = 1000$ and $\pi_M \in \{0.1, 0.5, 0.9\}$).

using $|D_P| + |D_U|$ samples. In the r tuning, a test computes the pseudo F-measure using the samples proportional to $|D_P| + |D_U|$ where each sample is classified by using the other $|D_P| + |D_U|$ samples. This is repeated 10 times with the training of EL-PUCcv. Thus, the total complexity of the r tuning is $O(10(|D_P| + |D_U|)^2 + 10K(r)|D_U|(|D_P| + |D_U|))$. NB-E&N and KD-E&N process $|D_P| + |D_U|$ samples for classifying a sample if it is labeled or unlabeled. They repeat this process $|D_P|$ times to evaluate the probability that a positive sample is labeled. Thus, their total complexity is $O(|D_P|(|D_P| + |D_U|))$. These facts are well reflected to Fig. 4a and b where the dependency of their computation times is linear or quadratic. Particularly, KD-E&N needs computation times comparable with or more than our



(a) Artificial data having various $|D_P|$ (averaged over $|D_U| \in \{100, 1000, 10000\}$ and $\pi_D \in \{0.1, 0.5, 0.9\}$).

(b) Artificial data having various $|D_U|$ (averaged over $|D_P| \in \{10, 100, 1000, 10000\}$ and $\pi_D \in \{0.1, 0.5, 0.9\}$).

(c) UCI concrete compressive strength data having various $|D_P|$ ($|D_U| = 200$ and averaged over $\pi_D \in \{0.1, 0.5, 0.9\}$).

Fig. 4. Computation times of the five PUCs.

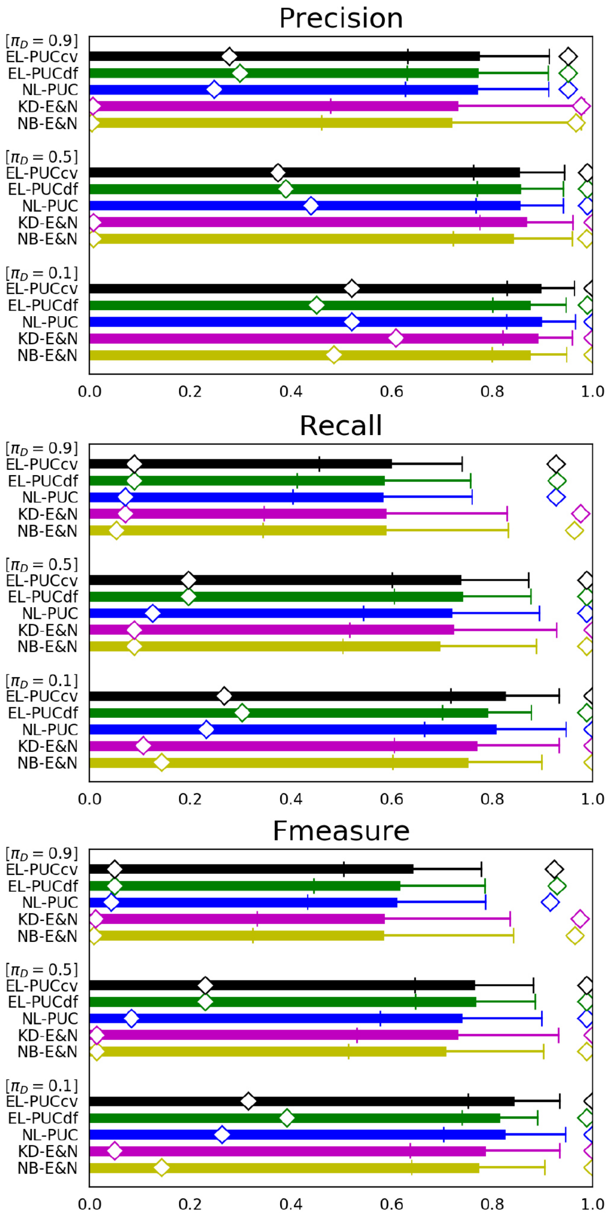


Fig. 5. Performance for UCI concrete compressive strength data having various $\pi_D = 0.1$, 0.5 and 0.9 ($|D_U| = 200$, $|D_P| \in \{10, 30, 90, 270\}$ and $\pi_M \in [0.1, 0.5, 0.9]$).

proposed PUCs. Note that these computation times may be positively biased by some overhead process under small $|D_P| = 10$.

5.2. *Performance evaluation using UCI data*

In the UCI repository, there are only a few data sets having known measurement processes. We used “Concrete Compressive Strength" (#samples: 1030, #observed variables [X]: 7, target variable [Y]:

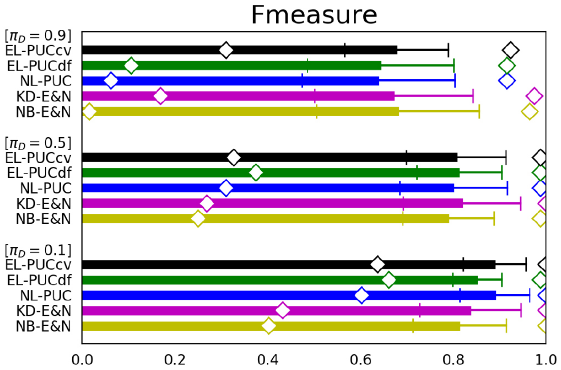


Fig. 6. Performance for UCI airfoil self-noise data having various π_D ($|D_U| = 300$, $|D_P| \in \{15, 45, 135, 405\}$ and $\pi_M \in [0.1, 0.5, 0.9]$).

compressive strength), “Airfoil Self-Noise" (#samples: 1503, #observed variables [X]: 5, target variable [Y]: angle of attack), and “Steel Plates Faults" (#samples: 1941, #observed variables [X]: 26, target variable [Y]: sigmoid of areas) from the physical sciences. We removed their categorical variables, binarized the target variables by applying their median values as a threshold, and chose one of the binary values as positive. Thus, they became balanced data sets in terms of their positive and negative samples. In each data set, we drew 20% of the total to produce three D_U having the probabilities $\pi_D = 0.1$, 0.5 and 0.9 to include positive samples, respectively. We also took extra 1%, 3%, 9% and 27% of the total to create four D_P, respectively, by choosing positive samples only. Furthermore, we drew extra 10% of the total to produce three M_U having $\pi_M = 0.1$, 0.5 and 0.9, respectively. For each of the three D_U, we computed the mean, the standard deviation, and the minimum/maximum of the accuracy indices of the five PUCs by marginalizing over all combinations of D_P and M_U.

Figures 5–7 compare the results using these UCI data sets. Because the figures of precision and recall of the last two data sets show similar behaviors with those of the first one, they are omitted. The figures mostly show a similar tendency with those of the artificial data. Under the most conditions of π_D, our three PUCs are more accurate and robust than NB-E&N and KD-E&N to the varieties of the size $|D_P|$ and the differences of the class prior π_M. Though the difference between our three PUCs are not very significant, EL-PUCcv provides the top average accuracy in many cases.

Figure 4c indicates the computation times to train the five PUCs using Concrete Compressive Strength data set. Because the size of the data set is small, the computation times of their overhead processes became dominant. Nevertheless, NL-PUC having lower complexity in essence was faster than NB-E&N, because the load of the kernel density based estimation was not significant for the small data set. We observed similar behaviors for the other two data sets.

6. Discussion

All results in Section 5 indicate that EL-PUCcv, which tunes r by using the pseudo F-measure: $\tilde{F}$, achieves the highest or comparable accuracy and robustness among our three PUCs. This reflects Lemma 4 that we can appropriately tune the parameters of the classifiers by using $\tilde{F}$. The traditional F-measure: F and $\tilde{F}$ of EL-PUCdf depicted in Fig, 8 more directly represents their consistency. Note that $\tilde{F}$ can be larger than unity according to its definition in Lemma 4. Though the standard deviations and the intervals between the minimum and the maximum of $\tilde{F}$ are larger than those of F, i.e., $\tilde{F}$ is less robust to the

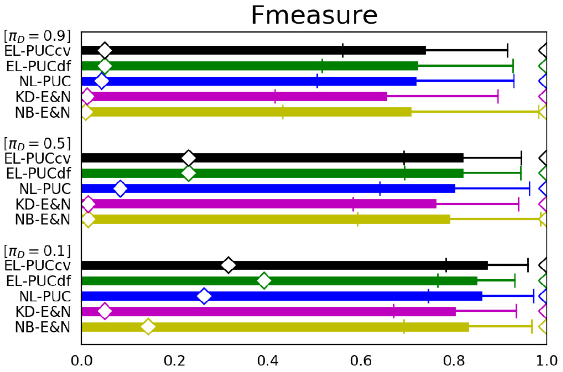


Fig. 7. Performance for UCI steel plates faults data having various π_D ($|D_U| = 400$, $|D_P| \in \{20, 60, 180, 540\}$ and $\pi_M \in$ [0.1, 0.5, 0.9]).

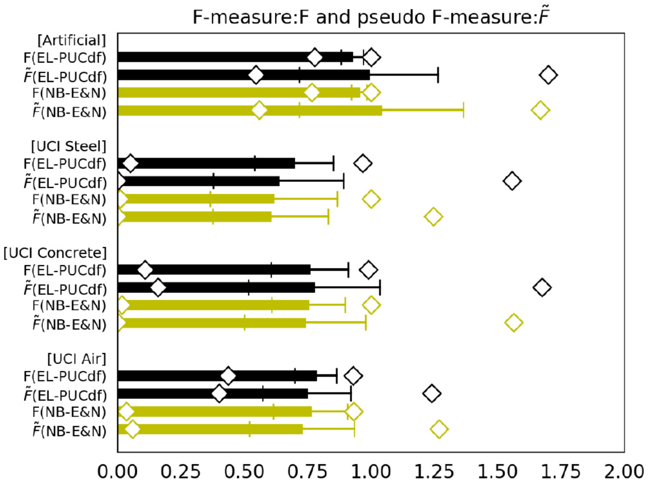


Fig. 8. Comparison between F-measure: F and pseudo F-measure: $\tilde{F}$ of EL-PUCdf and NB-E&N for various data ($|D_P| = 10$, $|D_U| = 1000$ and $\pi_D, \pi_M = \{0.1, 0.5, 0.9\}$ for artificial data, and the conditions of Fig. 5–7 for UCI data).

deviations of π_M from π_D under various conditions, the means of F and $\tilde{F}$ show consistent behaviors over the four data sets. This supports the use of $\tilde{F}$ to evaluate the accuracy of our likelihood based PUCs. Figure 8 also shows the comparisons of F and $\tilde{F}$ for NB-E&N which uses the aforementioned one sample setting. Although this setting does not match with the assumptions required to the consistency of $\tilde{F}$, F and $\tilde{F}$ of NB-E&N indicates their consistency comparable to those of EL-PUCcv. This may be because the probability to correctly classify a positive sample as positive by the Gaussian Naïve Bayes classifier used in NB-E&N is moderatly affected by the variation of the class priors π. This shows that the pseudo F-measure $\tilde{F}$ is widely applicable to qualitatively assess the performance of the PUCs.

According to the dependency of EL-PUC's accuracy on the parameter r as indicated by the lemmas and the theorems in Subsection 4.3, we proposed EL-PUCcv which selects an optimal r using the cross validation to achieve the highest accuracy. Meanwhile, Fig. 9 shows F-measure of EL-PUC over various r for the artificial data having three $\pi_D = 0.1$, 0.5 and 0.9 where all other conditions were marginalized. In the marginalized view, the accuracy of EL-PUC does not have any clear dependency on r. This supports the experimental results in Section 5 indicating the small differences of the accuracy and the robustness

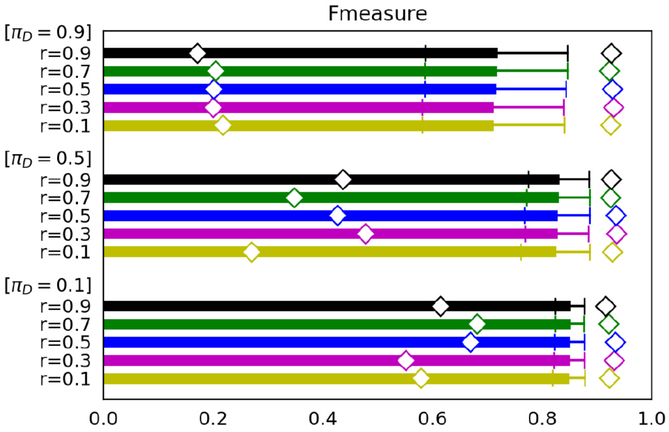


Fig. 9. Dependency of EL-PUC on parameter r for artificial data having various π_D ($|D_P| \in \{15, 45, 135, 405\}$, $|D_U| \in \{100, 1000, 10000\}$ and $\pi_D \in [0.1, 0.5, 0.9]$).

between EL-PUCdf and EL-PUCcv in most cases. On the other hand, EL-PUCcv requires extremely large computation times for the parameter tuning as shown in Section 5. These facts suggest the limited applicability of EL-PUCcv to practical measurement tasks particularly when the available time for training is limited.

Besides, EL-PUCdf shows higher accuracy and robustness than NL-PUC under large π_D and small $|D_P|$ while the computation time of EL-PUCdf is larger than that of NL-PUC. Therefore, we need to consider the trade-off between the classification performance and the computation time for the selection of the two approaches.

7. Real-world application

We applied our PUCs to noise reduction in a real-world single molecule measurement [25]. Figure 10 depicts the core part of the sensor called a “nano-gap.” The passage of an object between two nanoscopic electrodes is observed as an electric pulse induced by the quantum tunneling effect. Our task is to classify each pulse into a target molecule and a contaminant in the solvent based on the pulse outline in real time. A technical difficulty here is that the solvent containing no contaminants is hardly produced. Therefore, we always obtain a data set containing the target molecule’s pulses with the contaminant’s noise pulses, whereas we can get a data set containing the noise pulses of the contaminants only by observing the solvent without the target molecules. Moreover, the nano-gap sensor is disposable in every measurement task. Because characters of the nano-gaps are mutually different depending on the fluctuations of their manufacturing processes while a nano-gap is stable during a measurement task, we need to train the classifier in a short on-line period for the quick start at the beginning of every measurement task. Accordingly, the time for acquiring the noise pulses is very limited, while the target pulses with the noise pulses are acquired for a long period in their measurement. These required specifications meet with the conditions addressed by our proposed PUCs.

In our application, a pulse signal time series between the start and the end of the pulse is partitioned into 10 equi-width time intervals, and the signal points in each interval are averaged. The pulse outline is a 10 dimensional vector consisting of the 10 averages. We acquired D_P consisting of the noise pulse outlines, which were labeled as positive, in a short initial on-line period for the quick start. Successively,

Table 1
Performance comparison of noise reduction

	Time for	$\tilde{F}$		
Method	Training	$\pi_D \simeq \pi_M$	$\pi_D < \pi_M$	$\pi_D \ll \pi_M$
NB-E&N	1.47 msec	0.572	0.582	0.592
KD-E&N	4.90 msec	0.318	0.322	0.282
NL-PUC	7.29 msec	0.715	0.671	0.652
EL-PUCdf	736.19 msec	0.726	0.695	0.672

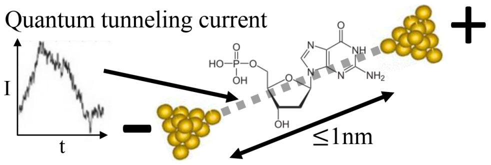


Fig. 10. A Single molecule sensor.

the unlabeled pulse outlines of the target molecules with the contaminants are acquired for a certain limited period to create D_U where their labels are estimated by a PUC during its training and the pulses classified into the target molecules become the initial part of the measurement output. After the PUC is trained by using D_P and D_U, the on-line noise reduction of new incoming pulses starts. The ratio of the targets and the contaminants in the solvent, i.e., the class prior, rapidly varies in the on-line measurement.

We used the pseudo F-measure $\tilde{F}$ to evaluate the performance of the noise reduction, because the ground truth of the classification was unknown. We acquired an extra labeled positive test set M_P together with D_P in the initial short period, and further acquired some unlabeled test sets M_U for computing $\tilde{F}$ of the five PUCs after the on-line measurement started. Their sizes were defined as $|D_P| = |M_P| = 20$, $|D_U| = 800$ and $|M_U| = 100$ by following the aforementioned specifications and the convenience to compute $\tilde{F}$.

Table 1 shows the performances of the four PUCs applied to three M_U. We excluded EL-PUCcv because its large computation time is not suitable for on-line processing. The same codes and the same PC with Section 5 were used. The class prior π_M of the first M_U acquired immediately after the start of the on-line measurement is almost same with π_D. Then, contaminants in the solvent rapidly increased over the periods acquiring the second and third M_U which made the second M_U as $\pi_D < \pi_M$ and the third M_U as $\pi_D \ll \pi_M$. Note that $\tilde{F}$ is not normalized into [0,1], and its value does not represent the absolute accuracy, while the larger $\tilde{F}$ show a better performance. The best numbers are underlined in each column. Particularly, EL-PUCdf showed the best $\tilde{F}$, while it takes relatively large computation time for its training. NL-PUC is better for the instant start of the on-line measurement, and EL-PUCdf is better if the dead time for nearly 1sec is allowed. The entire performance of NB-E&N and KD-E&N were significantly worse than those of NL-PUC and EL-PUCdf becuase of the small $|D_P| = 20$ available for the training. This is consistent with the result for the artificial data depicted in Fig. 2.

8. Conclusion

Our proposed PUCs show high accuracy and robustness under wide conditions. They are particularly advantageous when the numbers of the labeled positive samples and the unlabeled samples are imbalanced and when the positive class prior probability is dominant. Moreover, our proposed EL-PUC provides better accuracy than our proposed simple NL-PUC when the number of the labeled positive samples is limited and when the positive class prior probability is dominant, while EL-PUC needs more computation time.

In various practical problems, the labeled positive dataset D_P may be contaminated by the falsely labeled negative instances. Some instances in D_P and D_U may have missing values. The PUCs including our proposed approaches and the past approaches are not applicable to such incomplete datasets. These issues remain for future work. Moreover, we proposed our PUCs using the nonparametric estimation. The idea may be extended to discriminative frameworks.

References

[1] Y. Bengio, O. Delalleau and N.L. Roux, Efficient non-parametric function induction in semi-supervised learning, in *Proc. AISTATS05: the 10th International Workshop on Artificial Intelligence and Statistics*, 2005, pp. 96–103.

[2] G. Blanchard, G. Lee and C. Scott, Semi-supervised novelty detection, *J. Machine Learning Research* **11** (2010), 2973–3009.

[3] B. Calvo, I. Inza, P. Larranaga and J.A. Lozano, Wrapper positive bayesian network classifiers, *Knowledge and Information Systems* **33** (2010), 631–654.

[4] M.C. du Plessis, G. Niu and M. Sugiyama, Class-prior estimation for learning from positive and unlabeled data, in *Proc. ACML15: the 7th Asian Conf. on Machine Learning*, vol. 45, 2015, pp. 221–236.

[5] M.C. du Plessis and M. Sugiyama, Semi-supervised learning of class balance under class-prior change by distribution matching, *Neural Networks* **50** (2014), 110–119.

[6] M.C. du Plessiss, G. Niu and M. Sugiyama, Analysis of learning from positive and unlabeled data, in *Proc. NIPS14: Advances in Neural Information Processing Systems*, vol. 27, 2014, pp. 703–711.

[7] C. Elkan and K. Noto, Learning classifiers from only positive and unlabeled data, in *Proc. KDD08: the 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2008, pp. 213–220.

[8] J. Gama, I. Žliobait e, A. Bifet, M. Pechenizkiy and A. Bouchachia, A survey on concept drift adaptation, *ACM Computing Surveys (CSUR)* **46** (2014), 44:1–44:37.

[9] S. Hajizadeh, Z. Li, R.P.B.J. Dollevoet and D.M.J. Tax, Evaluating classification performance with only positive and unlabeled samples, in *Proc. S+SSPR14: Structural, Syntactic, and Statistical Pattern Recognition, vol. LNCS 8621*, 2014, pp. 233–242.

[10] N.W. Hengartner and E. Matzner-Lober, Asymptotic unbiased density estimators, *ESAIM: Probability and Statistics* **13** (2009), 1–14.

[11] S. Jain, M. White and P. Radivojac, Recovering true classifier performance in positive-unlabeled learning, in *Proc. AAAI17: the 31st AAAI Conf. on Artificial Intelligence*, 2017, p. 3060.

[12] K. Komlos, S. Popovics, T. Nurnbergerova, B. Babal and J.S. Popovics, Ultrasonic pulse velocity test of concrete properties as specified in various standards, *Cement and Concrete Composites* **18** (1996), 357–364.

[13] W.S. Lee and B. Liu, Learning with positive and unlabeled examples using weighted logistic regression, in *Proc. ICML03: the 20th Int. Conf. on Machine Learning*, 2003.

[14] A. Lewis, Getdist: Kernel density estimation, github: getdist document, University of Sussex, 2015. http://cosmologist.info/notes/GetDist.pdf.

[15] X.-L. Li, P.S. Yu, B. Liu and S.-K. Ng, Positive unlabeled learning for data stream classification, in *Proc. SDM09: the 2009 SIAM Int. Conf. on Data Mining*, 2009, pp. 259–270.

[16] E.A. Marina De Marco, Influence of left ventricular stroke volume on incident heart failure in a population with preserved ejection fraction (from the strong heart study), *American Journal of Cardiology* **119** (2017), 1047–1052.

[17] A. Menon, B.V. Rooyen, C.S. Ong and B. Williamson, Learning from corrupted binary labels via class-probability estimation, in *Proc. ICML15: the 32nd Int. Conf. on Machine Learning*, vol. 37, 2015, 125–134.

[18] G. Niu, M.C. du Plessis, T. Sakai, Y. Ma and M. Sugiyama, Theoretical comparisons of positive-unlabeled learning against positive-negative learning, in *Proc. NIPS16: Advances in Neural Information Processing Systems*, Vol. 29, 2016, pp. 1199–1207.

[19] S.J. Pan and Q. Yang, A survey on transfer learning, *IEEE Trans. on Knowledge and Data Engineering* **22** (2010), 1345–1359.

[20] D. Pfeffermann, C. Skinner, D.J. Holmes, H. Goldstein and J. Rasbash, Weighting for unequal selection probabilities in multilevel models, *J. the Royal Statistical Society. Series B (Statistical Methodology)* **60** (1998), 23–40.

[21] H.G. Ramaswamy, C. Scott and A. Tewari, Mixture proportion estimation via kernel embedding of distributions, in *Proc. ICML16: the 33rd Int. Conf. on Machine Learning*, vol. 5, 2016, pp. 2996–3004.

[22] M. Saerens, P. Latinne and C. Decaestecker, Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure, *Neural Computation* **14** (2002), 21–41.

[23] C. Scott, A rate of convergence for mixture proportion estimation, with application to learning from noisy labels, in *Proc. AISTATS15: the 18th Int. Conf. on Artificial Intelligence and Statistics*, vol. 38, 2015, pp. 838–846.

[24] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall/CRC, 1985, ch. 3.3 and 43.

[25] M. Tsutsui, M. Taniguchi, K. Yokota and T. Kawai, Identifying single nucleotides by tunneling current, *Nature Nanotechnology* **5** (2010), 286–290.

[26] B. Wang and X. Wang, Bandwidth selection for weighted kernel density estimation, *Electronic J. of Statistics* (2007). doi: 10.1214/154957804100000000.

[27] G. Ward, T. Hastie, S. Barry, J. Elith and J.R. Leathwick, Presence-only data and the em algorithm, *Biometrics* **65** (2009), 554–563.

[28] T. Washio, G. Imamura and G. Yoshikawa, Machine learning independent of population distributions for measurement, in *Proc. DSAA17: the 4th IEEE Int. Conf. on Data Science and Advanced Analytics*, 2017, pp. 212–221.

[29] X. Zhu, Z. Ghahramani and J. Laffer, Semisupervised learning using gaussian elds and harmonic functions, in *Proc. ICML03: the 20th Int. Conf. on Machine Learning*, 2003.