Clustering Unclustered Data: Unsupervised Binary Labeling of Two Datasets Having Different Class Balances

Marthinus Christoffel du Plessis, Gang Niu, Masashi Sugiyama
Dept. of Computer Science
Tokyo Institute of Technology, Japan
{christo@sg., gang@sg., sugi@}cs.titech.ac.jp

***Abstract*—We consider the unsupervised learning problem of assigning labels to unlabeled data. A naive approach is to use clustering methods, but this works well only when data is properly clustered and each cluster corresponds to an underlying class. In this paper, we first show that this unsupervised labeling problem in balanced binary cases can be solved if two unlabeled datasets having different class balances are available. More specifically, estimation of the *sign* of the difference between probability densities of two unlabeled datasets gives the solution. We then introduce a new method to directly estimate the sign of the density difference without density estimation. Finally, we demonstrate the usefulness of the proposed method against several clustering methods on various toy problems and real-world datasets.**

***Index Terms*—clustering; class-balance change.**

I. Introduction

Gathering labeled data is expensive and time consuming in many practical machine learning problems, and therefore class labels are often absent. In this paper, we consider the problem of *labeling*, which is aimed at giving a label to each sample. Labeling is similar to classification, but it is slightly simpler than classification because classes do not have to be specified. That is, labeling just tries to split unlabeled samples into disjoint subsets, and class labels such as male/female or positive/negative are not assigned to samples.

A naive approach to the labeling problem is to use a clustering technique which is aimed at assigning a label to each sample of the dataset to divide the dataset into disjoint clusters. The tacit assumption in clustering is that the clusters correspond to the underlying classes. However, this assumption is often violated in practical datasets, for example, when clusters are not well separated or a dataset exhibits within-class multimodality. An example of the labeling problem is illustrated in Figure 1. Figure 1(a) shows the samples drawn from a mixture of two normal distributions (differing only in the mean). Because the two clusters are highly overlapping, it may not be possible to properly label them by a clustering method.

In this paper we show that if one more dataset with a different class balance is available (Figure 1(b)), the labeling problem can be solved (Figures 1(c) and 1(d)). More specifically, we show that a labeling for the samples can be obtained by estimating *the sign of the difference between probability densities of two unlabeled datasets*. Thus, now our challenge is to estimate the sign of the density difference as accurate as possible.

A naive way to estimate the sign of the density difference is to first separately estimate two densities from two sets of samples and then take the sign of their difference to obtain a labeling. However, this naive procedure violates *Vapnik's principle* [1]:

> If you possess a restricted amount of information for solving some problem, try to solve the problem directly and never solve a more general problem as an intermediate step. It is possible that the available information is sufficient for a direct solution but is insufficient for solving a more general intermediate problem.

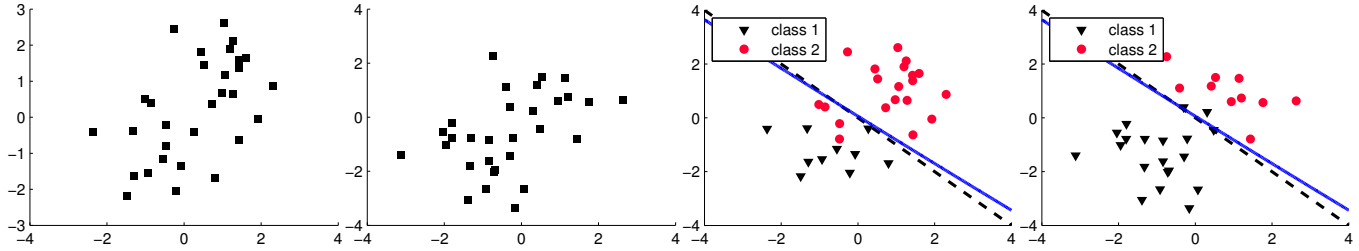
This principle was successfully used in the development of *support vector machines* (SVMs): Rather than modeling two classes of samples, SVM directly learns a decision boundary that is sufficient for performing pattern recognition.

In the current context, estimating two densities is more general than labeling samples. Thus, the above naive scheme may be improved by estimating the density difference directly and then taking its sign to obtain the class labels. Recently, a method was introduced to directly estimate the density difference, called the *least-squares density difference (LSDD)* estimator [2]. Thus, the use of LSDD for labeling is expected to improve the performance.

However, the LSDD-based procedure is still indirect; directly estimating the sign of the density difference would be the most suitable approach to labeling. In this paper, we show that the sign of the density difference can be directly estimated by lower-bounding the L_1-distance between probability densities. Based on this, we give a practical algorithm for labeling and illustrate its usefulness through experiments on various real-world datasets.

II. Problem Formulation and Fundamental Approaches

In this section, we formulate the problem of labeling, give our fundamental strategy, and consider two naive approaches.



(a) Samples in $\mathcal{X}_p$ (b) Samples in $\mathcal{X}_{p'}$ (c) Labeling of $\mathcal{X}_p$ (d) Labeling of $\mathcal{X}_{p'}$

Fig. 1: Illustrative example of labeling samples from unbalanced datasets. Figures 1a and 1b show the samples of the two datasets which differ only by class balance (the datasets are denoted as $\mathcal{X}_p$ and $\mathcal{X}_{p'}$). The discriminant estimated by the method that we propose in this paper is plotted by the blue solid line and the optimal discriminant is plotted by the black dashed line. The true underlying class labels (which are unknown) are illustrated with red and black points.

A. Problem Formulation

Suppose that there are two joint probability distributions on $\boldsymbol{x} \in \mathbb{R}^d$ and $y \in \{1, -1\}$ with densities $p(\boldsymbol{x}, y)$ and $p'(\boldsymbol{x}, y)$, which are different only in class balances:

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

Here $p(y)$ and $p'(y)$ denote the marginal probabilities of y and $p(\boldsymbol{x}|y)$ and $p'(\boldsymbol{x}|y)$ denotes the conditional densities of $\boldsymbol{x}$ given y, respectively. From these distributions, we are given two sets of unlabeled samples:

$$\mathcal{X}_p = \{\boldsymbol{x}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}) \text{ and } \mathcal{X}_{p'} = \{\boldsymbol{x}'_j\}_{j=1}^{n'} \overset{\text{i.i.d.}}{\sim} p'(\boldsymbol{x}),$$

where $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$ denote the marginal densities of $\boldsymbol{x}$. The goal of labeling is to obtain a labeling for the two sets of samples, $\mathcal{X}_p$ and $\mathcal{X}_{p'}$, that corresponds to the underlying class labels $\{y_i\}_{i=1}^n$ and $\{y'_j\}_{j=1}^{n'}$. However, different from classification, we do not obtain correct class labels, but we obtain correct class separation up to label commutation.

B. Fundamental strategy

Here we show that, for the case where the class priors are equal, we can obtain a labeling for samples in $\mathcal{X}_p$ and $\mathcal{X}_{p'}$.

We may write the class-posterior probability for the equal prior case as

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

where $q(y = 1) = q(y = -1) = \frac{1}{2}$. A class label can then be assigned to the most likely class by

$$d(\boldsymbol{x}) = \operatorname{sign}\left[q(y = 1|\boldsymbol{x}) - q(y = -1|\boldsymbol{x})\right],$$

where $d(\boldsymbol{x})$ denotes the criterion for labeling. Below we explain how $d(\boldsymbol{x})$ can be estimated from two sets of unlabeled samples $\mathcal{X}_p$ and $\mathcal{X}_{p'}$.

We can write the difference between class-posteriors as

$$\begin{aligned} &q(y = 1|\boldsymbol{x}) - q(y = -1|\boldsymbol{x}) \\ &= \frac{p(\boldsymbol{x}|y = 1)\frac{1}{2}}{q(\boldsymbol{x})} - \frac{p(\boldsymbol{x}|y = -1)\frac{1}{2}}{q(\boldsymbol{x})} \\ &= \frac{1}{2q(\boldsymbol{x})}\left(p(\boldsymbol{x}|y = 1) - p(\boldsymbol{x}|y = -1)\right). \end{aligned}$$

Since $1/(2q(\boldsymbol{x}))$ is always positive, the criterion becomes

$$d(\boldsymbol{x}) = \operatorname{sgn}\left[p(\boldsymbol{x}|y = 1) - p(\boldsymbol{x}|y = -1)\right].$$

Now the difference between marginal densities can be written as

$$\begin{aligned} &p(\boldsymbol{x}) - p'(\boldsymbol{x}) \\ &\quad = p(y = 1)p(\boldsymbol{x}|y = 1) + [1 - p(y = 1)]\, p(\boldsymbol{x}|y = -1) \\ &\qquad - p'(y = 1)p(\boldsymbol{x}|y = 1) - [1 - p'(y = 1)]\, p(\boldsymbol{x}|y = -1) \\ &\quad = [p(y = 1) - p'(y = 1)]\left[p(\boldsymbol{x}|y = 1) - p(\boldsymbol{x}|y = -1)\right]. \end{aligned}$$

Therefore, the criterion can be expressed as

$$d(\boldsymbol{x}) = A \operatorname{sgn}\left[p(\boldsymbol{x}) - p'(\boldsymbol{x})\right],$$

where $A = \operatorname{sgn}\left[p(y = 1) - p'(y = 1)\right]$.

This expression means that, if we know the class proportions in $\mathcal{X}_p$ and $\mathcal{X}_{p'}$, we can compute A and thus class labels can be obtained only from unlabeled samples. In practice, however, we may not know the class proportions and thus we can only label unlabeled samples (i.e., split unlabeled samples into disjoint subsets which correspond to the original class labels).

Thus, now our challenge is to obtain a good estimator of the sign of density difference, $\operatorname{sign}\left[p(\boldsymbol{x}) - p'(\boldsymbol{x})\right]$.

C. Kernel Density Estimation

A naive approach to estimating the sign of density-difference is to use *kernel density estimation* (KDE) [3]. For Gaussian kernels, the KDE solutions are given by

$$\widehat{p}(\boldsymbol{x}) \propto \sum_{i=1}^{n} \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}_i\|^2}{2\sigma^2}\right),$$
$$\widehat{p}'(\boldsymbol{x}) \propto \sum_{j=1}^{n'} \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'_j\|^2}{2\sigma'^2}\right).$$

The Gaussian widths σ and σ' may be determined based on least-squares cross-validation [4]. Finally, a labeling is obtained as

$$y = \operatorname{sign}\left[\widehat{p}(\boldsymbol{x}) - \widehat{p}'(\boldsymbol{x})\right]. \tag{2}$$

D. Direct Estimation of the Density Difference

KDE is a nice density estimator, but it is not necessarily suitable in density-difference estimation, because small estimation error incurred in each density estimate can cause a big error in the final density-difference estimate. More intuitively, good density estimators tend to be smooth and thus a density-difference estimator obtained from such smooth density estimators tends to be over-smoothed [5], [6].

The density difference can be estimated in a single shot using the *least-squares density difference* (LSDD) approach [2]. In this approach, we directly fit a model $g(\boldsymbol{x})$ to the density difference under the square loss:

$$\widehat{g} = \operatorname*{argmin}_{g} \frac{1}{2} \int \left(g(\boldsymbol{x}) - (p(\boldsymbol{x}) - p'(\boldsymbol{x}))\right)^2 \mathrm{d}\boldsymbol{x},$$

which can be efficiently obtained for a kernel density-difference model. Finally, a labeling is obtained as

$$y = \operatorname{sign}[\widehat{g}(\boldsymbol{x})].$$

III. Direct Estimation of the Sign of the Density Difference

We expect that an improved solution can be obtained by LSDD over KDEs due to more direct nature of LSDD. However, LSDD is still indirect because the sign of density difference is inspected after the density difference is estimated. In this section, we show how to directly estimate the sign of the density difference.

A. Derivation of the Objective Function

By lower-bounding the L_1-distance between probability densities, defined as

$$\int |p(\boldsymbol{x}) - p'(\boldsymbol{x})| \, \mathrm{d}\boldsymbol{x}, \tag{3}$$

we can obtain the sign of the density difference. We begin by considering the following self-evident relation:

$$|t| \geq tz, \text{ if } |z| \leq 1.$$

We can apply this relation at each point $\boldsymbol{x}$, to obtain

$$|p(\boldsymbol{x}) - p'(\boldsymbol{x})| \geq g(\boldsymbol{x}) \left[p(\boldsymbol{x}) - p'(\boldsymbol{x})\right] \text{ if } |g(\boldsymbol{x})| \leq 1, \ \forall \boldsymbol{x}.$$

By applying the above inequality to Eq.(3) and maximizing with respect to $g(\boldsymbol{x})$, we can obtain the tightest lower bound as

$$\int |p(\boldsymbol{x}) - p'(\boldsymbol{x})| \, \mathrm{d}\boldsymbol{x} \geq \sup_{g} \int g(\boldsymbol{x}) \left[p(\boldsymbol{x}) - p'(\boldsymbol{x})\right] \mathrm{d}\boldsymbol{x} \quad \text{s.t. } |g(\boldsymbol{x})| \leq 1, \ \forall \boldsymbol{x}. \tag{4}$$

It is straightforward to verify that the above relation will be met with equality when

$$g(\boldsymbol{x}) = \operatorname{sign}\left(p(\boldsymbol{x}) - p'(\boldsymbol{x})\right).$$

What makes the expression in the right-hand side of Eq.(4) especially useful is that the probability densities occur linearly in the integral. By replacing the integrals with sample averages and searching $g(\boldsymbol{x})$ from a parametric family (denoted as $g_{\boldsymbol{\alpha}}(\boldsymbol{x})$), we can write the above as

$$\widehat{\boldsymbol{\alpha}} = \operatorname*{arg\,min}_{\boldsymbol{\alpha}} \frac{1}{n'} \sum_{i=1}^{n'} g_{\boldsymbol{\alpha}}(\boldsymbol{x}'_i) - \frac{1}{n} \sum_{j=1}^{n} g_{\boldsymbol{\alpha}}(\boldsymbol{x}_j) \quad \text{s.t.} \quad |g_{\boldsymbol{\alpha}}(\boldsymbol{x})| \leq 1, \ \forall \boldsymbol{x}. \tag{5}$$

B. Optimization

Here we briefly discuss how to solve the optimization problem (5). The function in Eq. (5) should satisfy the constraint $|g(\boldsymbol{x})| \leq 1, \ \forall \, \boldsymbol{x}$. We can consider a clipped version of the function that always satisfies the constraint:

$$\widetilde{g}(\boldsymbol{x}) = R(g(\boldsymbol{x})), \quad \text{where} \quad R(z) = \min\left(1, \max\left(-1, z\right)\right).$$

We use a linear-in-parameter model,

$$g(\boldsymbol{x}) = \sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}),$$

where $\varphi_\ell(\boldsymbol{x})$ are the basis functions. Using the above definitions and including a regularizer, we arrive at the following objective function to be minimized:

$$J(\boldsymbol{\alpha}) = \frac{1}{n'} \sum_{i=1}^{n'} R\left(\sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}'_i)\right) - \frac{1}{n} \sum_{j=1}^{n} R\left(\sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}_j)\right) + \frac{\lambda}{2} \sum_{\ell=1}^{b} \alpha_\ell^2. \tag{6}$$

Although the above objective function is non-convex, we can efficiently find a local minimizer using the *convex-concave procedure* (CCCP) [7].

CCCP requires the objective function to be split into convex and concave parts:

$$J(\boldsymbol{\alpha}) = J_{\mathrm{vex}}(\boldsymbol{\alpha}) + J_{\mathrm{cave}}(\boldsymbol{\alpha}).$$

This is done by expressing $R(z)$ as

$$R(z) = C_{-1}(z) - C_1(z) - 1,$$

where $C_\epsilon(z) = \max(0, z - \epsilon)$. This results in the following convex and concave functions:

$$J_{\mathrm{vex}}(\boldsymbol{\alpha}) = \frac{1}{n'} \sum_{i=1}^{n'} C_{-1}\left(\sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}'_i)\right) + \frac{1}{n} \sum_{j=1}^{n} C_1\left(\sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}_j)\right) + \frac{\lambda}{2} \sum_{\ell=1}^{b} \alpha_\ell^2,$$

$$J_{\mathrm{cave}}(\boldsymbol{\alpha}) = -\frac{1}{n'} \sum_{i=1}^{n'} C_1\left(\sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}'_i)\right) - \frac{1}{n} \sum_{j=1}^{n} C_{-1}\left(\sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}_j)\right).$$

Using the *Fenchel inequality* [8, p. 94], we can bound the function $C_\epsilon(z)$ as

$$C_\epsilon(z) \geq zt - C^*_\epsilon(t),$$

where $C_\epsilon^*(t)$ is the *Fenchel dual* of $C_\epsilon(z)$,

$$C_\epsilon^*(t) = \begin{cases} \epsilon t & 0 \le t \le 1, \\ \infty & \text{otherwise.} \end{cases}$$

Applying this to the concave part gives

$$J_{\mathrm{cave}}(\boldsymbol{\alpha}) \le \bar{J}_{\mathrm{cave}}(\boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{c}),$$

where the bound is specified by $\boldsymbol{b}$ and $\boldsymbol{c}$:

$$\begin{aligned}\bar{J}_{\mathrm{cave}}(\boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{c}) =& \frac{1}{n'} \sum_{i=1}^{n'} \left(C_1^*(b_i) - b_i \sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}'_i) \right) \\ &+ \frac{1}{n} \sum_{j=1}^{n} \left(C_{-1}^*(c_j) - c_j \sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}_j) \right).\end{aligned}$$

This bound is convex w.r.t. $\boldsymbol{b}$ and $\boldsymbol{c}$ if $\boldsymbol{\alpha}$ is fixed. Using this bound, we have

$$J(\boldsymbol{\alpha}) \le J_{\mathrm{vex}}(\boldsymbol{\alpha}) + \bar{J}_{\mathrm{cave}}(\boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{c}).$$

The strategy to minimize $J(\boldsymbol{\alpha})$ is then to alternately minimize the right-hand side by minimizing w.r.t. $\boldsymbol{\alpha}$ (keeping $\boldsymbol{b}$ and $\boldsymbol{c}$ fixed) and minimize w.r.t. $\boldsymbol{b}$ and $\boldsymbol{c}$ (keeping $\boldsymbol{\alpha}$ fixed). Minimization w.r.t. $\boldsymbol{\alpha}$ minimizes the current upper bound and minimization w.r.t. $\boldsymbol{b}$ and $\boldsymbol{c}$ corresponds to tightening the bound at the current point.

Our final optimization algorithm is summarized below:

1) *Initialize the starting value:*

$$\boldsymbol{\alpha}^1 \leftarrow \arg\min_{\boldsymbol{\alpha}} J_{\mathrm{vex}}(\boldsymbol{\alpha}).$$

2) *For $t = 1, \ldots T$:*

a) *Tighten the upper-bound:* Obtain $\boldsymbol{b}$ and $\boldsymbol{c}$ as

$$\boldsymbol{b}^t, \boldsymbol{c}^t \leftarrow \arg\min_{\boldsymbol{b}, \boldsymbol{c}} \bar{J}_{\mathrm{cave}}(\boldsymbol{\alpha}^t, \boldsymbol{b}, \boldsymbol{c}),$$

which can be analytically performed as

$$b_i^t \leftarrow \begin{cases} 0 & \text{if } \sum_{\ell=1}^{b} \alpha_\ell^t \varphi_\ell(\boldsymbol{x}'_i) < 1, \\ 1 & \text{otherwise,} \end{cases}$$

$$c_j^t \leftarrow \begin{cases} 0 & \text{if } \sum_{\ell=1}^{b} \alpha_\ell^t \varphi(\boldsymbol{x}_j) < -1, \\ 1 & \text{otherwise.} \end{cases}$$

b) *Minimize the upper bound:* Set

$$\boldsymbol{\alpha}^{t+1} \leftarrow \arg\min_{\boldsymbol{\alpha}} J_{\mathrm{vex}}(\boldsymbol{\alpha}) + \bar{J}_{\mathrm{cave}}(\boldsymbol{\alpha}, \boldsymbol{b}^t, \boldsymbol{c}^t),$$

which can be performed by solving the following convex quadratic problem:

$$\begin{aligned}\min_{\boldsymbol{\alpha}} \quad & -\sum_{\ell=1}^{b} \alpha_\ell \left(\frac{1}{n'} \sum_{i=1}^{n'} b_i^t \varphi_\ell(\boldsymbol{x}'_i) + \frac{1}{n} \sum_{j=1}^{n} c_j^t \varphi_\ell(\boldsymbol{x}_j) \right) \\ & + \frac{1}{n'} \sum_{i=1}^{n'} \xi'_i + \frac{1}{n} \sum_{j=1}^{n} \xi_j + \frac{\lambda}{2} \sum_{\ell=1}^{b} \alpha_\ell^2 \\ \text{s.t.} \quad & \xi'_i \ge 0,\ \xi'_i \ge \sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}'_i) + 1,\ \ \forall i = 1, \ldots, n' \\ & \xi_j \ge 0,\ \xi_j \ge \sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}_j) - 1\ \ \forall j = 1, \ldots, n.\end{aligned}$$

In practice, Gaussian kernels centered at the sample points in $\mathcal{X}_p$ and $\mathcal{X}_{p'}$ are chosen as the basis functions. All hyper-parameters are set by cross-validation. We call this proposed method *direct sign density difference (DSDD)* estimation.

C. Generalization Error Bounds

Suppose that we have a test distribution, that shares the same class conditional distribution, but has a class prior $p_{\mathrm{te}}(y = 1) = \theta p(y = 1) + (1 - \theta) p'(y = 1)$, with $0 \le \theta \le 1$. We can then express θ as

$$\theta = \frac{p_{\mathrm{te}}(y = 1) - p'(y = 1)}{p(y = 1) - p'(y = 1)}.$$

We consider the decision function of the form

$$g(\boldsymbol{x}) = \sum_{i=1}^{n+n'} \alpha_i k(\boldsymbol{x}, \boldsymbol{c}_i), \tag{7}$$

where k is a kernel function, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{n+n'})$, and $\boldsymbol{c}_i = \boldsymbol{x}_i$ for $1 \le i \le n$ and $\boldsymbol{c}_i = \boldsymbol{x}'_{i-n}$ for $n + 1 \le i \le n + n'$. We consider the following *surrogate loss* [9]:

$$\ell_\eta(z) = \min\left(1, \max(0, 1 - z\eta)\right).$$

For any $\eta > 0$, $\ell_\eta(z)$ lower bounds the *indicator loss* and approaches the *indicator loss* when η approaches zero. Then we have the following theorem (its proof is omitted due to lack of space; we decomposed the test distribution into the weighted sum of two training distributions, and then applied standard error bounds using the Rademacher complexity [9]):

Theorem 1: Assume that

$$\exists B_k > 0, \forall \boldsymbol{x} \in \mathbb{R}^d, k(\boldsymbol{x}, \boldsymbol{x}) \le B_k^2.$$

Let $\boldsymbol{\alpha}^*$ be an optimal solution to DSDD, $g(\boldsymbol{x})$ be the decision function defined in Eq. (7) with parameter $\boldsymbol{\alpha}^*$, and

$$B_{\mathcal{F}} = \sqrt{\boldsymbol{\alpha}^{*\top} K \boldsymbol{\alpha}^*}, \quad B'_{\mathcal{F}} = \|\boldsymbol{\alpha}^*\|_1,$$

where K is the kernel matrix. Assume that the ground truth class labels $y_1, \ldots, y_n, y'_1, \ldots, y'_{n'}$ are available for evaluation. Then, with probability at least $1 - \delta$, we have

$$\begin{aligned}&\mathbb{E}_{p_{\mathrm{te}}}[\ell(yg(\boldsymbol{x}))] - \frac{\theta}{n} \sum_{i=1}^{n} \ell_\eta(y_i g(\boldsymbol{x}_i)) - \frac{1 - \theta}{n'} \sum_{i=1}^{n'} \ell_\eta(y'_i g(\boldsymbol{x}'_i)) \\ &\le \left(\frac{\theta}{\sqrt{n}} + \frac{1 - \theta}{\sqrt{n'}} \right) \frac{2 B_k B_{\mathcal{F}}}{\eta} \\ &\quad + \left(\frac{\theta}{\sqrt{n}} + \frac{1 - \theta}{\sqrt{n'}} \right) \min\left(3, 1 + \frac{4 B_k^2 B'_{\mathcal{F}}}{\eta}\right) \sqrt{\frac{\ln(2/\delta)}{2}},\end{aligned}$$

where the expectation $\mathbb{E}_{p_{\mathrm{te}}}[\ell(yg(\boldsymbol{x}))]$ follows the test distribution $p_{\mathrm{te}}(\boldsymbol{x}, y)$.

From the above, we see that the order of the bound is $O\left(1/\sqrt{n} + 1/\sqrt{n'}\right)$. Compared to supervised classification from i.i.d. data such as support vector machines [9], which has an order of $O(1/\sqrt{n + n'})$, our bounds converge slower. However, we *do not require class labels for training* in our problem setting.

IV. Experiments

We first illustrate the operation of our method on a toy example. Then we use real-world benchmark data to show the superiority of our algorithm.

A. Toy Problem

We illustrate the problem and our method with a simple example. Suppose that the class-conditional densities for the two classes are given as

$$p(\boldsymbol{x}|y=1) = \mathcal{N}_{\boldsymbol{x}}\left(-\mathbf{1}_2, \boldsymbol{I}_{2\times 2}\right),$$
$$p(\boldsymbol{x}|y=-1) = \mathcal{N}_{\boldsymbol{x}}\left(\mathbf{1}_2, \boldsymbol{I}_{2\times 2}\right),$$

where $\mathcal{N}_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal density with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ w.r.t. $\boldsymbol{x}$. $\mathbf{1}_2$ is a 2×1 vector of ones and $\boldsymbol{I}_{2\times 2}$ is a 2×2 identity matrix. We generate 2 sets of 30 samples with class-priors $p(y=1)=0.3$ and $p'(y=1)=0.7$, respectively. The result is illustrated in Figure 1. As can be seen from this example, we are able to obtain a labeling of the classes that roughly corresponds to the true (unknown) labels of the data.

B. Benchmark Datasets

We compare our method against several competing methods on benchmark datasets. For each experiment, we constructed the datasets $\mathcal{X}_p$ and $\mathcal{X}_{p'}$ by drawing n and n' samples from the positive and negative classes of the binary classification datasets according to a prior of $p(y=1)$ and $p'(y=1)$. The labeling was then performed using these two datasets. A label was assigned to each sample according to the sign of the density difference. Since the exact class label can not be determined if the class-priors are unknown, the labeling error rate was calculated:

$$\mathrm{LER} := \min\left(\mathrm{MCR}, 1-\mathrm{MCR}\right),$$

where MCR represents the misclassification rate with the assigned labels. $1-$MCR is the misclassification rate assuming that all labels are flipped.

We compared the following methods:

- **Direct Sign Density Difference (DSDD) Estimation** (proposed): Directly estimate $\mathrm{sign}\left(p(\boldsymbol{x}) - p'(\boldsymbol{x})\right)$ using the method described in Section III. Hyperparameters are selected via cross validation.
- **Least-Squares Density Difference (LSDD) Estimation**: Estimate $\mathrm{sign}\left[p(\boldsymbol{x}) - p'(\boldsymbol{x})\right]$ by estimating $p(\boldsymbol{x}) - p'(\boldsymbol{x})$ using the least squares fitting method [2]. Hyperparameters are selected via cross validation.
- **Kernel Density Estimation (KDE):** Estimate $\mathrm{sign}\left[p(\boldsymbol{x}) - p'(\boldsymbol{x})\right]$ by estimating the densities $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$ with KDE. Hyperparameters are selected using least-squares cross validation.
- **K-Means (KM)**: Cluster the data into two clusters using the K-means algorithm [10].
- **Spectral Clustering (SC)**: Cluster the data into two clusters using the spectral clustering algorithm [11]. The affinity matrix was constructed with 7 nearest neighbors.
- **Squared-loss Mutual Information based Clustering (SMIC)** : Cluster the data according to the SMIC method [12]. SMIC was chosen since it provides model selection, avoiding the need for subjective parameter tuning.

For experiments, the *UCI benchmark datasets*[1] were used. We compare the performance of the methods by varying the class balance on these datasets. Two class balances were selected: one with a large difference between the classes ($p(y=1)=0.2$ and $p'(y=1)=0.8$) and one with a small difference between the classes ($p(y=1)=0.35$ and $p'(y=1)=0.65$). The average and standard deviation of the labeling error rate for the two experiments, with $|\mathcal{X}_p| = |\mathcal{X}_{p'}| = 40$ are given in Tables I and II.

From the results we see that methods which follow the approach proposed in Section II of estimating the sign of the density difference (i.e., DSDD, LSDD, and KDE) generally work better than methods using the cluster structure of the data (i.e., KM, SC, and SMIC). The thyroid dataset lends itself to interpretation of why these methods work better. The labels in the thyroid dataset correspond to healthy and diseased. The diseased label is caused by either a hyper-functioning or hypo-functioning thyroid. These two underlying causes cause within-class multimodality which may cause clustering-based methods to fail.

Among the methods which estimate the sign of the density difference, we see that DSDD generally performs better than LSDD and LSDD in turn performs better than KDE. This is as expected since KDE solves a more general problem than LSDD, and LSDD solves a more general problem than DSDD. This pattern is even more pronounced on the more difficult case where the class balances are close to each other (Table II).

V. Conclusion

The problem of unsupervised labeling of two unbalanced datasets was considered. We first showed that this problem can be solved if two unlabeled datasets having different class balances are available. More specifically, we showed that the solution can be obtained by estimating of the sign of the difference between probability densities. We then introduced a method to directly estimate the sign of the density difference and avoid density estimation. The method was shown on various datasets to outperform competing methods that either estimate the density difference or use the cluster structure of the data.

Because the sign of density difference corresponds to the Bayes optimal classifier under equal class balance, it may be estimated by any classifier that separates $\mathcal{X}_p$ and $\mathcal{X}_{p'}$. Following this idea, we tested the *support vector machine* (SVM) for estimating the sign of density difference. However, this did not work well due to the high overlap of $\mathcal{X}_p$ and $\mathcal{X}_{p'}$—both the datasets are mixtures of two classes, only with different mixing ratios.

From this classification point of view, we can actually see that our objective function (6) corresponds to the *robust*

[1]The datasets were obtained from http://archive.ics.uci.edu/ml/.

TABLE I: Labeling error rate for experiments with a class prior of $p(y=1)=0.2$ and $p'(y=1)=0.8$. The size of each dataset was $|\mathcal{X}_p|=40$ and $|\mathcal{X}_{p'}|=40$. The best method in terms of the mean error and comparable methods according to the two-sided paired t-test at the significance level 5% are specified by bold face. The standard deviation of the labeling error rate is given in brackets.

Dataset	DSDD	LSDD	KDE	KM	SC	SMIC
australian	**.142**(.045)	.174(.110)	.211(.126)	.266(.147)	.381(.033)	.303 (.103)
banana	.179(.097)	**.170**(.070)	.237(.147)	.431(.068)	.427(.141)	.424 (.141)
diabetes	.246(.122)	**.223**(.079)	**.226**(.051)	.372(.080)	.380(.094)	.370 (.131)
german	.268(.059)	.281(.127)	**.211**(.051)	.437(.114)	.448(.128)	.439 (.052)
heart	**.176**(.051)	**.174**(.047)	.211(.074)	.261(.131)	.310(.032)	.327 (.107)
image	**.198**(.078)	.206(.047)	**.201**(.049)	.385(.093)	.351(.119)	.384 (.135)
ionosphere	**.157**(.059)	.184(.106)	.194(.123)	.329(.145)	.319(.113)	.311 (.174)
saheart	.310(.093)	**.205**(.048)	.238(.113)	.422(.121)	.395(.113)	.384 (.072)
thyroid	**.102**(.052)	.121(.116)	.207(.074)	.328(.113)	.326(.109)	.305 (.074)
twonorm	.044(.085)	.051(.072)	.200(.028)	**.036**(.054)	.043(.069)	.048 (.071)

TABLE II: Labeling error rate for experiments with a class prior of $p(y=1)=0.35$ and $p'(y=1)=0.65$. The size of each dataset was $|\mathcal{X}_p|=40$ and $|\mathcal{X}_{p'}|=40$. The test setup is the same as that in Table I.

Dataset	DSDD	LSDD	KDE	KM	SC	SMIC
australian	**.244**(.116)	.259(.088)	.355(.104)	.265(.080)	.376(.065)	.308 (.107)
banana	**.338**(.094)	**.339**(.100)	.365(.067)	.433(.049)	.427(.069)	.424 (.070)
diabetes	**.340**(.075)	.361(.124)	.345(.034)	.373(.063)	.380(.048)	.371 (.114)
german	.375(.042)	.380(.093)	**.354**(.057)	.437(.024)	.445(.057)	.438 (.041)
heart	.270(.133)	**.247**(.084)	.354(.052)	.264(.059)	.315(.081)	.327 (.089)
image	**.331**(.078)	.350(.067)	.350(.039)	.384(.031)	.354(.049)	.382 (.050)
ionosphere	**.291**(.099)	.356(.066)	.345(.048)	.330(.070)	.322(.058)	.314 (.107)
saheart	.378(.093)	**.353**(.057)	.363(.066)	.419(.082)	.395(.022)	.385 (.040)
thyroid	**.227**(.098)	.251(.087)	.302(.022)	.326(.061)	.329(.047)	.307 (.076)
twonorm	.164(.188)	.153(.121)	.352(.096)	**.036**(.053)	.042(.122)	.049 (.120)

SVM [13] that minimizes the ramp loss (a clipped hinge loss). Thanks to the robustness brought by the ramp loss, the overlapped datasets $\mathcal{X}_p$ and $\mathcal{X}_{p'}$ can be separated more reliably, and thus we obtained good estimation of the sign of density difference.

Furthermore, this view conversely shows that the robust SVM is actually a suitable classification method because it directly estimates the Bayes optimal classifier, the sign of density difference. Labeling and classification are different problems, but one can actually give insight into the other. In the future work, we will further investigate the relation between labeling and classification.

ACKNOWLEDGMENT

MCdP and GN were supported by MEXT scholarships, and MS was supported by KAKENHI 25700022.

REFERENCES

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

[2] M. Sugiyama, T. Kanamori, T. Suzuki, M. C. du Plessis, S. Liu, and I. Takeuchi, "Density-difference estimation," *Neural Computation*, 2013, to appear.

[3] B. Silverman, *Density estimation for statistics and data analysis*. London, UK: Chapman and Hall, 1986.

[4] W. Härdle, M. Müller, S. Sperlich, and A. Werwatz, *Nonparametric and semiparametric models*. Springer, 2004.

[5] P. Hall and M. P. Wand, "On nonparametric discrimination using density differences," *Biometrika*, vol. 75, no. 3, pp. 541–547, 1988.

[6] N. H. Anderson, P. Hall, and D. Titterington, "Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates," *Journal of Multivariate Analysis*, vol. 50, no. 1, pp. 41–54, 1994.

[7] A. L. Yuille and A. Rangarajan, "The concave-convex procedure (CCCP)," in *NIPS 14*, 2002.

[8] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, Mar. 2004.

[9] P. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, pp. 463–482, 2002.

[10] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. Berkeley, CA, USA: University of California Press, 1967, pp. 281–297.

[11] J. Shi and J. Malik, "Normalized cuts and image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 8, pp. 888–905, 2000.

[12] M. Sugiyama, N. Gang, M. Yamada, M. Kimura, and H. Hachiya, "Information-maximization clustering based on squared-loss mutual information," *Neural Computation*, 2013, to appear.

[13] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. New York, NY, USA: Cambridge University Press, 2004.