Discriminative Density-ratio Estimation

Yun-Qian Miao * Ahmed K. Farahat † Mohamed S. Kamel ‡

Abstract

The covariate shift is a challenging problem in supervised learning that results from the discrepancy between the training and test distributions. An effective approach which recently drew a considerable attention in the research community is to reweigh the training samples to minimize that discrepancy. In specific, many methods are based on developing Density-ratio (DR) estimation techniques that apply to both regression and classification problems. Although these methods work well for regression problems, their performance on classification problems is not satisfactory. This is due to a key observation that these methods focus on matching the sample marginal distributions without paying attention to preserving the separation between classes in the reweighted space. In this paper, we propose a novel method for Discriminative Density-ratio (DDR) estimation that addresses the aforementioned problem and aims at estimating the density-ratio of joint distributions in a class-wise manner. The proposed algorithm is an iterative procedure that alternates between estimating the class information for the test data and estimating new density ratio for each class. To incorporate the estimated class information of the test data, a soft matching technique is proposed. In addition, we employ an effective criterion which adopts mutual information as an indicator to stop the iterative procedure while resulting in a decision boundary that lies in a sparse region. Experiments on synthetic and benchmark datasets demonstrate the superiority of the proposed method in terms of both accuracy and robustness.

Keywords: Covariate Shift; Density-ratio Estimation; Cost-sensitive Classification.

1 Introduction

There are many real world applications where the test data demonstrates distribution shift from the training data. In such cases, traditional machine learning techniques usually encounter performance degradation because their models are fitted towards minimizing error or risk of error for the training samples. So it is important that the learning algorithms can demonstrate some degree of adaptivity to cope with distribution changes. This has resulted in intensive research under the names domain adaptation [1][2], transfer learning [3], and concept drift [4][5]. One particular case is the covariate shift problem [6][7], which assumes that the marginal distributions are changed between training and test data (i.e., $p_{ts}(x) \neq p_{tr}(x)$), while the class conditional distributions are not affected (i.e., $p_{ts}(y|x) = p_{tr}(y|x)$).

Usually the covariate shift happens in biased sample selection scenarios. For example, in building an action recognition system, the training samples are collected in a university lab, where young people make up a high percentage of the population. When the system is intended to be applied in reality, it is likely that we will face a more general population model.

To compensate for the distribution gap between the training and test data, the objective of model fitting is modified to minimize the expectation of the weighted error, where the weight of the training sample is justified using its density ratio [6], i.e., $\beta(x) = p_{ts}(x)/p_{tr}(x)$. Therefore, the solution to covariate shift is formulated as estimating the marginal density ratio and applying some cost-sensitive learning techniques [8][9].

Recently, a number of methods have been proposed to estimate the Density ratio (DR), given two sets of finite number of observation samples. There are two groups of methods for Density-ratio estimation in the literature. One is a two-step procedure: first estimate the training and test probability densities separately and then divide them. The second group of methods estimates the density ratio directly in one step. These one-shot methods usually achieve more accurate and robust results and are considered the state of the art [10][11][12][13].

In the literature, the aforementioned reweighting of training samples according to the density ratio is examined in a wide range of applications, including both the regression and classification tasks. We have seen these reweighting methods performing well in many regression tasks. However, existing research and our experiments found that these methods do not yield satisfactory re-
sults in classification scenarios. In many cases, they even recorded worse prediction accuracy than the simple unweighted approach. This motivated us to investigate the covariate shift classification problems in this work. Our key observation is that these conventional density-ratio methods focus on matching the training and test distributions without paying attention to preserving the separation among classes in the reweighted space. So these traditional density-ratio estimation methods may deteriorate the discrimination ability even if the marginal distributions might be matched well in general.

In this paper, we propose a novel method called Discriminative Density-ratio (DDR) estimation that addresses the aforementioned problem and aims to estimate the density ratio between the joint distributions in a class-wise manner. To do so, we divide the task into two parts: (1) estimating the density ratio between the training and test data for each class; and (2) estimating the class prior changes for the test data. As the class labels for the test data are unknown, the proposed method is based on an iterative procedure, which alternates between estimating the class information for the test data and estimating new class-wise density ratios.

In comparison to the conventional approach which matches sample marginal distributions, the proposed class-wise matching method has two benefits. First, it allows relaxing the assumption of the covariate shift that \( p_{ts}(y|x) = p_{tr}(y|x) \) and accordingly captures a mixture of distribution changes. Second, it focuses on the classification problems and considers preserving the separation among classes while matching the shifted distributions. Our experiments on synthetic and benchmark datasets confirm the effectiveness of the proposed DDR algorithm.

The paper is organized as follows. The rest of this section describes the notations used in the paper. Section 2 introduces the covariate shift problem. Section 3 reviews the state of the art on the density-ratio estimation and analyzes the limitations of previous work. Section 4 describes our proposed method. In Section 5, empirical evaluations are conducted. Section 6 concludes the paper.

### 1.1 Notation

Throughout this paper, scales, vectors, and matrices are shown in small, bold, and capital letters respectively. When discussing covariate shift classification, we use the following notations:

- \( \mathcal{X} \subseteq \mathbb{R}^d \): the d-dimension input space,
- \( x \in \mathcal{X} \) is an input sample,
- \( \mathcal{Y} \): the class label space, \( y \in \mathcal{Y} \) is an output variable.
- \( p_{tr} \): the probability density of the training data
- \( p_{ts} \): the probability density of the test data
- \( n_{tr} \): the number of training samples
- \( n_{ts} \): the number of test samples
- \( \pi_{tr} \): the set of training samples
- \( \pi_{ts} \): the set of test samples
- \( \beta \): the density-ratio between two distributions
- \( \gamma \): the ratio between two class priors

### 2 Learning Under Covariate Shift

With the empirical risk minimization framework [13], the general purpose of a supervised learning problem is to minimize the expected risk of

\[
R(\theta, p, l) = \int \int l(x, y, \theta) p(x, y) \, dx \, dy,
\]

where \( \theta \) is a learned model, \( l(x, y, \theta) \) is a loss function for the problem with a joint distribution \( p(x, y) \).

When we are facing the case where the training distribution \( p_{tr}(x, y) \) differs from the distribution of test data \( p_{ts}(x, y) \), in order to obtain the optimal model in the test domain \( \theta_{ts}^* \), we can derive the following reweighting scheme:

\[
\theta_{ts}^* = \arg \min_{\theta \in \Theta} \sum_{(x,y) \in \pi_{ts}} \frac{p_{ts}(y|x)}{p_{tr}(y|x)} \frac{p_{ts}(x)}{p_{tr}(x)} l(x, y, \theta)
\]

Further, covariate shift assumes that the class conditional distributions are the same across the training and test data (i.e. \( p_{ts}(y|x) = p_{tr}(y|x) \)), but that the marginal distributions are different. Hence we can derive that:

\[
\theta_{ts}^* \approx \arg \min_{\theta \in \Theta} \sum_{(x,y) \in \pi_{ts}} \frac{p_{ts}(y|x)}{p_{tr}(y|x)} \frac{p_{ts}(x)}{p_{tr}(x)} l(x, y, \theta)
\]

Now, the learning objective in the new test domain would be evaluated by the importance-weighted training
samples to reflect the changes of distribution, where the importance of a sample is equal to the density-ratio $\beta$.

Having the weighted training instances, there are plenty of cost-sensitive learning algorithms that can be applied. Instead of minimizing the loss of misclassification, the cost-sensitive learning aims at minimizing the instance-dependent cost of wrong prediction [15, 8, 9]. For example, the Support Vector Machines [16] and Regularized Least Squares [12] can naturally embed weighted samples in the training process.

3 Density-ratio Estimation

Because of the increasing demand from practical application domains to develop machine learning systems that adapt to unseen cases, the Density-ratio (DR) estimation has attracted considerable attention in the research community, and there are numerous methods being proposed to solve the problem in the literature. The simple approach is to solve the density-ratio estimation problem in two steps: estimating the training and test distributions separately and taking a division. However, this naïve method encounters several problems [10]:

1. The information from the given limited number of samples may be sufficient to infer the density-ratio, but insufficient to infer two probability density functions. The estimation of probability density is usually a more general and challenging problem.

2. A small estimation error in the denominator can lead to a large variance in the density-ratio.

3. The naïve approach would be highly unreliable for high-dimension problems because of the well-known “curse-of-dimensionality” problem.

Therefore, researchers have been putting efforts on proposing new methods to estimate the density-ratio directly without going through the estimation of two probability densities. Along this direction, Huang et al. proposed a Kernel Mean Matching (KMM) algorithm [10], which directly gives the estimates of sample importance by minimizing the mean discrepancy in a reproduced kernel space. By explicitly modeling the function of density-ratio, another group of methods have been developed with the formulation of various objective functions, which include the Kullback-Leibler Importance Estimation Procedure (KLIEP) [12], Least-Squares Importance Fitting (LSIF) [17], unconstrained Least-Squares Importance Fitting (uLSIF) [17]. Among these density-ratio estimation methods, various advantages are demonstrated on their different applicable fields. In general, the method of uLSIF was shown to have excellent numerical stability and efficient run-time solution [18].

Besides the covariate shift problem mentioned above, the density-ratio estimation has shown noticeable potential in many data mining and machine learning fields. Some highlighted applicable fields are outliers detection [19], change-point detection for time series data, feature selection and feature extraction based on mutual information estimation [18].

3.1 Limitations of Previous Work. Reviewing the existing work on the importance reweighting strategy for covariate shift adaptation, the first issue is the strong assumption on the class conditional distributions (the posterior). The posteriors are assumed to be fixed between the training and test data, while the marginal distributions exhibit changes (i.e., $p_{ts}(y|x) = p_{tr}(y|x)$ and $p_{ts}(x) \neq p_{tr}(x)$). According to the Bayes’ rule, there is the following equation that describes the relationship between the prior, posterior, marginal, likelihood, and joint distributions as

\[ p(x, y) = p(y|x)p(x) = p(x|y)p(y). \]

The root cause of covariate shift is the sampling bias, i.e., $p_{ts}(x) \neq p_{tr}(x)$. In this giving condition, we can not assure the posterior will remain unchanged, even if we know that the concepts behind the data are stable. This means that when the distributions of covariate are shifted, there is a high possibility that the class conditional distributions and/or priors will also change.

Moreover, focusing on the classification problem, the objective is to discriminatively separate the instances into different classes. However, in the conventional weighting approach dealing with covariate shift adaptation, the distribution matching is performed on the whole input space. In other words, the existing algorithms focus on matching the training and test distributions without considering to preserve the separation among classes in the reweighted space.

These two problems heavily hold back the effectiveness of the weighting methods in correcting the covariate shift problem, especially for classification tasks. Several research reported this fact [11, 20, 18], but none of them presented a clear solution.

4 Proposed Approach

Having the intuition of preserving the separations between classes while pursuing the match of distributions, we propose an approach named Discriminative Density-ratio (DDR) estimation, which aims to estimate the density ratio between the joint distributions in a class-wise manner. Our proposed approach uses an iterative procedure. The class labels of the test samples are estimated using the updated density-ratio estimates and
in turn the density ratios are estimated for each class.
Following Eq. (2.2), instead of assuming unchanged
class conditional distributions and simplifying Eq. (2.2)
into the density ratio on \( x \), we decompose the joint
distributions from the perspective of class likelihood and
define a more general weighting scheme \( w \) to reflect the
density ratio of joint distributions as

\[
    w = \frac{p_{ts}(x, y)}{p_{tr}(x, y)} = \frac{p_{ts}(x|y) \cdot p_{ts}(y)}{p_{tr}(x|y) \cdot p_{tr}(y)}.
\]  (4.5)

For a classification problem, assuming class labels
are from a finite discrete set \( y \in [c_1, c_2, \ldots, c_m] \), then the density ratio of joint distributions can be evaluated
in a class-wise manner as

\[
    w = [w_{c_1}, w_{c_2}, \ldots, w_{c_m}]^T,
\]

where

\[
    w_{c_i} = \frac{p_{ts}(x|y = c_i)}{p_{tr}(x|y = c_i)} \cdot \frac{p_{ts}(y = c_i)}{p_{tr}(y = c_i)} \quad i = 1...m.
\]  (4.7)

Let \( \beta(x|y = c_i) = \frac{p_{ts}(x|y = c_i)}{p_{tr}(x|y = c_i)} \) be the density ratio
for class \( c_i \), and \( \gamma(y = c_i) = \frac{p_{ts}(y = c_i)}{p_{tr}(y = c_i)} \) be the ratio that
reflects the changes of priors. Then, Eq. (4.7) can be written as

\[
    w_{c_i} = \beta(x|y = c_i) \cdot \gamma(y = c_i).
\]  (4.8)

As a result, we can induce the weights for all training
samples in this class-wise manner, which reflects the changes of the joint distributions between the training
and test data. Now, estimating the density ratio of joint
distributions is divided into two sub-tasks: the estimation of the class-wise density ratio \( \beta \) and the estimation of the prior-ratio \( \gamma \).

However, the reality is that the test data do not
have label information. In order to proceed with the
class-wise matching, we propose an iterative procedure
that alternates between estimating the class information
for the test data and estimating new class-wise density
ratios. The success of this iterative procedure greatly
depends on two proposed components: (1) a soft distribution matching algorithm which incorporates the posteriors of the test data and, (2) an effective mutual information based stopping criterion. The details of the iterative procedures as well as the two components are explained in the rest of this section.

4.1 Iterative Estimation Procedure. The iterative procedure proceeds as follows (Algorithm 1 shows the complete steps).

Algorithm 1 Discriminative Density-ratio Estimation

**Input:** \( X_{tr}, Y_{tr}, X_{ts} \)

**Output:** \( w(x), x \in X_{tr} \)

**Steps:**

1. initialization: \( w^{(0)} = 1; \ t = 0 \)
2. while stop-criterion-not-met do
   3. learn a model \( \theta^{(t)} = \text{Learn}(x_{tr}, y_{tr}, w^{(t)}) \)
   4. predict \( \hat{y}^{(t+1)}(y|x \in X_{ts}) = \theta^{(t)}(x|x \in X_{ts}) \)
   5. estimate paired class: \( \beta^{(t+1)}(x|y = c_i) \leftarrow \text{SoftDR} \left( X_{tr}, Y_{tr}, X_{ts}; \hat{y}^{(t+1)}(y|x \in X_{ts}), c_i \right) \)
   6. estimate \( \gamma^{(t+1)}(y = c_i) \leftarrow \frac{\hat{y}^{(t+1)}(y = c_i)}{p_{tr}(y = c_i)} \)
   7. update \( w^{(t+1)} = \beta^{(t+1)} \cdot \gamma^{(t+1)} \)
   8. \( t = t + 1 \)
9. end while

- The procedure begins with learning a classification model based on the training samples \( \{X_{tr}, Y_{tr}\} \) whose weights are set to 1 \( (w^{(0)} = 1) \) in the first iteration (Step 3).
- The classification model is then used to estimate the posteriors of the test data. It is noticeable that there is a distribution change between the training and test data, but we assume that the model can still give reasonable predictions for the posteriors of test samples \( [21] \) (Step 4).
- The class-wise density ratios are then estimated. Since the test data has extra information which is the posterior probabilities, we propose to utilize this extra information by extending the current density-ratio estimation techniques to incorporate the weighted test data. The details of this new method are explained in the soft matching section, named Soft Density-ratio (SoftDR) estimation (Step 5).
- The new prior ratios are estimated using an approach similar to that of Chang and Ng \( [22] \), in which we use the classifier’s prediction results to update the priors (Step 6).
- Finally, the weight of training sample \( w^{(t)} \) is updated (Step 7).

The aforementioned steps are repeated until a stopping
criterion is met.

4.2 Soft Matching. Several density-ratio estimation
methods have been proposed to estimate the importance
(weights) of samples, in order to match two shifted distributions. The core concept of these methods relies on the kernel function to evaluate the similarity between
samples, i.e. \( k(x_i, x_j) \). The situation we face is that the test samples to be matched have soft decisions on the belongings to a class. An extension to the above kernel function can be used to utilize this information.

Assume there are confidence scores (or probabilities) \( w_i \) and \( w_j \) associated with samples \( x_i \) and \( x_j \). Let \( \phi(\cdot) \) be the mapping function associated with the kernel. Then, the kernel function between two weighted samples can be calculated as

\[
\begin{align*}
k((x_i, w_i), (x_j, w_j)) &= [w_i \cdot \phi(x_i)]^T [w_j \cdot \phi(x_j)] \\
&= w_i \cdot w_j \cdot \phi(x_i)^T \cdot \phi(x_j) \\
(4.9)\end{align*}
\]

Using the kernel of Eq. (4.9) with the original matching method allows the algorithm to perform soft Density-ratio (SoftDR) estimation. It is notable that the test sample confidence scores are induced by the posteriors and the training sample confidence scores are all set to 1, because we have labels for the training data.

Without loss of generality, we illustrate the soft matching extension to Unconstrained Least-squares Importance Fitting (uLSIF) [17] as an example. In uLSIF, the density ratio is modeled as a linear combination of a series of basis functions as

\[
\hat{\beta}(x) = \sum_{l=1}^{b} \alpha_l \varphi_l(x) = \sum_{l=1}^{b} \alpha_l k(x, x_l),
\]

where \( x_l \) is a set of predefined reference points.

Then, uLSIF learns the parameter \( \{\alpha_l\}_{l=1}^{b} \) by minimizing the squared loss of density-ratio function fitting. This leads to the following unconstrained optimization problem:

\[
(4.10) \quad \min_{\{\alpha_l\}_{l=1}^{b}} \left[ \frac{1}{2} \sum_{l,l'=1}^{b} \alpha_l \alpha_{l'} \hat{S}_{l,l'} - \sum_{l=1}^{b} \alpha_l \hat{s}_l + \frac{\lambda}{2} \sum_{l=1}^{b} \alpha_l^2 \right],
\]

where

\[
(4.11) \quad \hat{S}_{l,l'} := \frac{1}{n_{ts}} \sum_{x_l, x_{l'} \in \pi_{tr}} k(x_l, x_{l'}) k(x_l, x_{l'}),
\]

\[
(4.12) \quad \hat{s}_l := \frac{1}{n_{ts}} \sum_{x_l \in \pi_{ts}} k(x_l, x_l).
\]

Because the training samples are given the label information, dividing them into groups according to their class labels means that the training samples have weights of 1. But the test samples are classified by a model in each iteration, whose output confidence values reflect their probability of belonging to a class. To reflect the uncertainty of test samples belonging to a class, we propose to add the soft matching ability to the uLSIF algorithm. Using the concept of weighted kernel functions (Eq. 4.10), it can be observed that the objective function and \( S_{l,l'} \) are the same, except that \( \hat{s}_l \) (Eq. 4.12) needs to be modified as

\[
(4.13) \quad \hat{s}_{l,c} := \frac{1}{n_{ts}} \sum_{x_j \in \pi_{ts}} p(c|x_j) \cdot k(x_j, x_l),
\]

where \( p(c|x_j) \) is the posterior of sample \( x_j \) having a class label \( c \).

Among the existing density-ratio estimators, uLSIF is robust and computationally efficient, and hence it is used in our experiments. For other density-ratio estimation methods, such as the Kernel Mean Matching (KMM) [16] and the Kullback-Leibler Importance Estimation Procedure (KLIEP) [12], the soft matching can also be implemented in a similar way by modifying the kernel functions as show in Eq. (4.9).

### 4.3 Stopping Criterion

One naïve criterion for stopping the algorithm is based on the convergence of the weights, i.e. \( \|w^{(t+1)} - w^{(t)}\| \leq \epsilon \). However, we observed that this criterion only works when there is a clear separation between classes. For real datasets, this criterion will usually lead to a poor local solution. Instead, we propose to adopt the Mutual Information (MI) [23], as an indicator for a desirable location of the decision boundary.

Given a test sample \( x_t \), we define its posteriors using the current model as a \( m \) dimension vector (corresponding to \( m \) classes) as \( \hat{p}_t = [\hat{p}_{t1}, \hat{p}_{t2}, \ldots, \hat{p}_{tm}]^T \).

Then, the information entropy of this probability vector is defined as

\[
(4.14) \quad H(\hat{p}_t) = -\sum_{i=1}^{m} \hat{p}_{ti} \ln(\hat{p}_{ti}).
\]

MI between the test samples \( X_{ts} \) and their estimated labels \( \hat{Y}_{ts} \) using the model’s output \( \hat{p}_{ts} \) is defined as

\[
(4.15) \quad \text{MI} \left( X_{ts}, \langle \hat{Y}_{ts}, \hat{p}_{ts} \rangle \right) = H(\hat{p}_0) - \frac{1}{n_{ts}} \sum_{t=1}^{n_{ts}} H(\hat{p}_t),
\]

where \( \hat{p}_0 \) is the posterior vector for sample \( x_t \), \( \hat{p}_0 \) is the class prior, and \( H(\cdot) \) is the information entropy.

Mutual information has been studied in the context of discriminative clustering [24], semi-supervised learning [25] and domain adaptation [26]. Maximizing this criterion implicitly means that the output of the current model has the least amount of confusing labels and the classification boundaries lie at sparse regions. Facing the covariate shift scenarios and the unknown but
drifted test distributions, we expect that this criterion could serve as a good indicator that can be utilized as an effective stopping condition.

5 Experiments

In this section, we conduct three sets of experiments. The first one is on a synthetic 2-class 4-cluster data. The second experiment evaluates the sampling bias scenarios on different benchmark datasets. Following, a more challenging cross-dataset task is studied.

5.1 Synthetic Data. Our first experiment is designed with samples generated from 2-dimensional Gaussian mixture models, in which both the class priors and likelihoods exhibit changes. The 2-class 4-cluster distributions of the training and test data are given as Table 1.

Figure 1 illustrates the difference of importance estimation results between the unweighted approach, the conventional DR, and the proposed DDR methods, and their corresponding classification boundaries. We can observe that the conventional DR method assigns higher importance weights to the misclassified blue points because they lie in a dense region of test points (middle figure), while our proposed DDR method assigns small importance weights to these points and accordingly learns a much better decision boundary (right figure). The left figure clearly shows that classification using the unweighted approach is biased to training samples and leads to a suboptimal solution to the test data.

Table 2 reports the numerical classification results for various numbers of training samples using the Naive Bayes Classifier. The number of test samples is fixed to 2000. Both DR and DDR are based on the density-ratio estimating method of uLSIF. The results show that weighting the covariate shifted data with the proposed DDR method performs consistently better than the conventional DR method in term of the classification accuracy.

As a reference line, we also report the experimental result that is based on 5-fold cross validation on the test data, which is the performance for test data without exposure to distribution changes (the column ‘Oracle-test’ in Table 2).

5.2 Biased Sampling. Further, we evaluate our proposed DDR method on a set of benchmark datasets. The datasets ‘GermanCredit’, ‘DelveSplice’, ‘Ionosphere’, ‘Australian’, ‘BreastCancer’, ‘Diabe’ are from the UCI Machine Learning Repository. The datasets

Table 1: The distributions of the training and test data of the 2-class 4-cluster problem.

|       | Prior | Likelihood                          |
|-------|-------|-------------------------------------|
| $P_{tr}$ |       |                                     |
| class-1 | 0.5   | $0.9 \cdot \mathcal{N}(\frac{1}{5}, I) + 0.1 \cdot \mathcal{N}(\frac{4}{5}, I)$ |
| class-2 | 0.5   | $0.1 \cdot \mathcal{N}(\frac{1}{1}, I) + 0.9 \cdot \mathcal{N}(\frac{4}{1}, I)$ |
| $P_{ts}$ |       |                                     |
| class-1 | 0.6   | $0.5 \cdot \mathcal{N}(\frac{1}{5}, I) + 0.5 \cdot \mathcal{N}(\frac{4}{5}, I)$ |
| class-2 | 0.4   | $0.5 \cdot \mathcal{N}(\frac{1}{1}, I) + 0.5 \cdot \mathcal{N}(\frac{4}{1}, I)$ |

---

http://archive.ics.uci.edu/ml/datasets.html
Table 2: Classification accuracies over 30 runs for the 2-class 4-cluster data with variant number of training samples. For each test case, the best-performing method other than the reference method ‘Oracle-cvtest’ is highlighted in bold (according to a t-test with 5% significant level).

| $n_{tr}$ | Unweighted | uLSIF | DDR-uLSIF | Oracle-cvtest |
|----------|------------|-------|-----------|---------------|
| 100      | 0.9533±0.0143 | 0.9587±0.0182 | 0.9717±0.0060 | 0.9770±0.0035 |
| 200      | 0.9549±0.0094 | 0.9655±0.0095 | 0.9745±0.0046 | 0.9778±0.0036 |
| 300      | 0.9545±0.0101 | 0.9645±0.0085 | 0.9735±0.0032 | 0.9771±0.0026 |
| 400      | 0.9553±0.0072 | 0.9641±0.0087 | 0.9739±0.0042 | 0.9762±0.0035 |
| 500      | 0.9585±0.0065 | 0.9676±0.0064 | 0.9736±0.0048 | 0.9770±0.0030 |
| 1000     | 0.9596±0.0056 | 0.9681±0.0052 | 0.9737±0.0046 | 0.9776±0.0031 |

‘USPS’ and ‘MNIST’ are from the LibSVM data collection.\(^4\)

The covariate shift classification tasks are formulated by splitting the training and test data with a deliberate biased sampling selection procedure (following the setup of \(^20\)). In all experiments, before any further processing, all the data are normalized to the range $[-1,1]^d$. Then, the half of data are uniformly sampled to form the testing section. And, the rest of data are sub-sampled to form the biased training set with the probability of $P(s=1|x) = \frac{e^{x^\top\omega}}{1+e^{x^\top\omega}}$, where $s=1$ means the sample is included in the training set, and $v = \frac{4\omega^\top(x-x_0)}{\sigma^2 x_0^2 x^\top x}$, $\omega \in \mathbb{R}^d$ is a projection vector randomly chosen from $[-1,1]^d$. For each run of experiment, ten vectors of $\omega$ are randomly generated and we select the vector $\omega$ which maximizes the difference between the unweighted method and the weighted method with ideal sampling weights.

We employ the uLSIF method in our experiments because of its superiority in speed and numerical stability. The classifier we used is Importance-Weighted Least-Squares Probabilistic Classifier (IWLSPC) \(^27\). The number of kernel basis functions is set to 100 by random sampling from the test data. The other hyper parameters (the kernel width $\sigma$ and regularization parameter $\lambda$) are chosen by 5-fold Importance Weighted Cross-Validation (IWCV) \(^12\).

We evaluate the performance of our DDR method by comparing with the conventional density-ratio estimation method using the exact same settings. The classification results using the model learned from the unweighted training data are included as the baseline. Because of the deliberate biased sampling selection procedure, we know the probability of each sample being included into the training section is $P(s=1|x)$. Therefore, the perfect sample importance is known as the reciprocal of being selected, i.e., $imp = \frac{1}{P(s=1|x)}$. We report results of using this oracle importance weights as the reference (the column ‘Oracle-imp’ in Table 3).

All experiments are repeated 30 times with different training-test data splits. The significance of the improvement in classification accuracy is tested using a t-test at a significance level of 5%. The results are summarized in Table 3. It can be observed that the proposed DDR approach outperforms the unweighted method and the conventional density-ratio estimator in almost all cases. There are 4 out of 10 cases where the accuracies are improved by more than 10%.

5.3 Cross-dataset Tasks. Training a model with samples from one dataset and adapting the model to another dataset which is collected at different conditions, is usually seen as a very challenging problem. We evaluate our DDR approach in the cross-dataset classification task using the two handwritten digits recognition datasets: USPS and MNIST. The USPS dataset contains 9,298 handwritten digit images with the size 16 × 16. The MNIST dataset has a total of 70,000 handwritten digit images (the first 20,000 samples are used in our experiment). The size of each image is 28 × 28.

Because the two datasets have different image sizes and intensity levels, a preprocessing step is applied first as: (1) resize the image size of MNIST from 28 × 28 into the same size of USPS, 16 × 16; (2) normalize the feature (intensity of pixel) into the range of $[-1,1]$. Then, we conduct two scenarios of experiments: one using USPS for training and MNIST for testing, the other using MNIST for training and USPS for testing. The classification method being used is SVM with linear kernels \(^16\). The parameter $c$ in SVM is a trade-off between model generalization and training error, and its value is chosen using 5-fold importance weighted cross-validation.

Table 4 presents the average and standard deviations of the classification accuracies of 30 runs. Each

\(^4\)http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/
Table 3: Biased sampling on benchmark datasets: Classification accuracies over 30 runs. For each dataset, the best-performing method is highlighted in bold (according to a $t$-test with 5% significant level, the setup of ‘Oracle-imp’ is a reference and not involved to comparison).

| Dataset        | Unweighted | uLSIF     | DDR-uLSIF | Oracle-imp |
|----------------|------------|-----------|-----------|------------|
| GermanCredit   | 0.6970±0.0383 | 0.6888±0.0554 | 0.7013±0.0130 | 0.6939±0.0495 |
| DelveSplice     | 0.5527±0.0601 | 0.5797±0.0797 | 0.6679±0.1184 | 0.6336±0.0874 |
| Ionosphere      | 0.6818±0.0565 | 0.6759±0.0588 | 0.6979±0.0666 | 0.7483±0.0866 |
| Australian      | 0.8121±0.0298 | 0.8187±0.0450 | 0.8319±0.0275 | 0.8342±0.0272 |
| BreastCancer    | 0.8189±0.1498 | 0.7963±0.1379 | 0.9219±0.1107 | 0.8942±0.1361 |
| Diabete         | 0.7372±0.0245 | 0.7346±0.0273 | 0.7286±0.0279 | 0.7149±0.0378 |
| USPS5v6         | 0.9581±0.0084 | 0.9508±0.0297 | 0.9747±0.0062 | 0.9689±0.0163 |
| USPS3v8         | 0.6262±0.0623 | 0.7443±0.1549 | 0.9283±0.0813 | 0.7861±0.1561 |
| MNIST5v6        | 0.7979±0.1978 | 0.8888±0.1353 | 0.9477±0.0100 | 0.9124±0.1384 |
| MNIST3v8        | 0.5591±0.1079 | 0.5725±0.1228 | 0.7936±0.1640 | 0.6809±0.1819 |

run is based on randomly selecting 90% of training samples and test samples from the datasets. The report results show that the DDR method can significantly boost recognition accuracies. Compared to the conventional DR approach, for the scenario “USPS to MNIST” 7 out of 10 test cases achieve an improvement in accuracy of 2% to 7%. For the scenario “MNIST to USPS”, 8 out of 10 test cases gain an improvement in accuracy of 2% to 15%.

6 Conclusion

This paper proposes a novel algorithm for covariate shift classification problems which estimates the density-ratio in a discriminative manner. Instead of matching the marginal distributions without paying attention to the separations among classes, our proposed method estimates density ratio between joint distributions in a class-wise manner. Therefore, our method allows relaxing the strong assumption of covariate shift and preserves the separation between classes while minimizing the distribution discrepancy between the training and test data. In order to proceed with the class-wise matching, the proposed algorithm deploys an iterative procedure that alternates between estimating the class information for the test data and estimating new class-wise density ratios. Two modules contribute to the success of the proposed DDR method. One is the soft matching algorithm which extends current density-ratio estimation algorithms to incorporate sample posterior. Another important component is the employment of the mutual information as an indicator for stopping the iterative procedure. Experiments on synthetic and benchmark data confirm the superiority of the proposed algorithm.

Although our method focused on the covariate shift adaptation problem, we vision that the concept of discriminative distribution matching is also useful to other scenarios of transfer learning.

References

[1] Hal Daumé III and Daniel Marcu. Domain adaptation for statistical classifiers. Journal of Artificial Intelligence Research, 26(1):101–126, 2006.
[2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. Machine Learning, 79(1):151–175, 2010.
[3] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. Knowledge and Data Engineering, IEEE Transactions on, 22(10):1345–1359, 2010.
[4] Alexey Tsymbal. The problem of concept drift: definitions and related work. Technical Report TCD-CS-2004-15, Department of Computer Science, Trinity College Dublin, Ireland, 2004.
[5] R. P. J. C. Bose, Wil van der Aalst, Indrė Žliobaitė, and Mykola Pechenizkiy. Handling concept drift in process mining. In Advanced Information Systems Engineering, pages 391–405. Springer, 2011.
[6] Hidetoshi Shimodaia. Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of Statistical Planning and Inference, 90(2):227–244, 2000.
[7] Jose G Moreno-Torres, Troy Raeder, Rocio Alaiz-Rodriguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. Pattern Recognition, 45(1):521–530, 2012.
[8] Bianca Zadrozny, John Langford, and Naoki Abe. Cost-sensitive learning by cost-proportionate example weighting. In ICDM, 2003.
[9] Yanmin Sun, Mohamed S Kamel, Andrew KC Wong, and Yang Wang. Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition, 40(12):3358–3378, 2007.
Table 4: Cross-dataset tasks: classification accuracies over 30 runs on the USPS and MNIST datasets. For each test case, the best-performing method is highlighted in bold according to a t-test with 5% significant level. The italic means that it records the best average accuracy but not statistically significant.

| Test Case | USPS to MNIST | MNIST to USPS |
|-----------|---------------|---------------|
|           | Unweighted    | uLSIF         | DDR-uLSIF     | Unweighted    | uLSIF         | DDR-uLSIF     |
| 0 vs 1    | 0.8695±0.0619 | 0.9337±0.0682 | 0.9455±0.0098 | 0.9697±0.0129 |              |               |
| 1 vs 2    | 0.5947±0.0280 | 0.6008±0.0227 | 0.9114±0.0207 | 0.9245±0.0301 |              |               |
| 2 vs 3    | 0.7200±0.0756 | 0.7688±0.0580 | 0.6432±0.0393 | 0.6412±0.0334 |              |               |
| 3 vs 4    | 0.7991±0.0421 | 0.8226±0.0323 | 0.7878±0.0390 | 0.8129±0.0258 |              |               |
| 4 vs 5    | 0.6373±0.0624 | 0.7279±0.0656 | 0.8322±0.0317 | 0.8794±0.0173 |              |               |
| 5 vs 6    | 0.5644±0.0593 | 0.5861±0.0705 | 0.5901±0.0280 | 0.6787±0.0507 |              |               |
| 6 vs 7    | 0.6025±0.0820 | 0.6041±0.0807 | 0.6039±0.0813 | 0.5105±0.0276 | 0.5150±0.0320 |               |
| 7 vs 8    | 0.6290±0.0631 | 0.6164±0.0642 | 0.6352±0.0697 | 0.6240±0.0407 | 0.7775±0.0440 |               |
| 8 vs 9    | 0.6977±0.0783 | 0.7376±0.0825 | 0.8066±0.0826 | 0.8255±0.0622 | 0.8340±0.0509 |               |
| 9 vs 0    | 0.9092±0.0314 | 0.9133±0.0302 | 0.9142±0.0287 | 0.6482±0.1067 | 0.6775±0.1131 | 0.8099±0.1087 |

[10] Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Schölkopf. Correcting sample selection bias by unlabeled data. In Advances in Neural Information Processing Systems 19, pages 601–608, 2007.

[11] Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. Dataset shift in machine learning, pages 131–160, 2000.

[12] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. The Journal of Machine Learning Research, 8:985–1005, 2007.

[13] Yaoliang Yu and Csaba Szepesvári. Analysis of kernel mean matching under covariate shift. In ICML, pages 607–614, 2012.

[14] Vladimir N Vapnik. Statistical learning theory. Wiley, 1998.

[15] Charles Elkan. The foundations of cost-sensitive learning. In International Joint Conference on Artificial Intelligence, volume 17, pages 973–978, 2001.

[16] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011.

[17] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. Efficient direct density ratio estimation for non-stationarity adaptation and outlier detection. Advances in neural information processing systems, 21:809–816, 2008.

[18] Masashi Sugiyama, Takafumi Kanamori, Taiji Suzuki, Shohei Hido, Jun Sese, Ichiro Takeuchi, and Liwei Wang. A density-ratio framework for statistical data processing. IPSJ Transactions on Computer Vision and Applications, 1(0):183–208, 2009.

[19] Shohei Hido, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama, and Takafumi Kanamori. Statistical outlier detection using direct density ratio estimation. Knowledge and information systems, 26(2):309–336, 2011.

[20] Corinna Cortes, Mehryar Mohri, Michael Riley, and Ashin Rostamizadeh. Sample selection bias correction theory. In Algorithmic Learning Theory, pages 38–53. Springer, 2008.

[21] Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In International Conference on Artificial Intelligence and Statistics, pages 129–136, 2010.

[22] Yee Seng Chan and Hwee Tou Ng. Word sense disambiguation with distribution estimation. In IJCAI, pages 1010–1015, 2005.

[23] William M Wells, Paul Viola, Hideki Atsumi, Shin Nakajima, and Ron Kikinis. Multi-modal volume registration by maximization of mutual information. Medical image analysis, 1(1):35–51, 1996.

[24] Inderjit S Dhillon, Subramanyam Mallela, and Dharmendra S Modha. Information-theoretic co-clustering. In Proceedings of the ninth ACM SIGKDD, pages 89–98. ACM, 2003.

[25] Ran El-Yaniv and Oren Souviron. Iterative double clustering for unsupervised and semi-supervised learning. In Machine Learning: ECML 2001, pages 121–132. Springer, 2001.

[26] Yuan Shi and Fei Sha. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In ICML, 2012.

[27] Hirotaka Hachiya, Masashi Sugiyama, and Naonori Ueda. Importance-weighted least-squares probabilistic classifier for covariate shift adaptation with application to human activity recognition. Neurocomputing, 80:93–101, 2012.