Oversampling for Imbalanced Data via Optimal Transport

Yuguang Yan,[1,2*†] **Mingkui Tan,**[1*] **Yanwu Xu,**[3*]
Jiezhang Cao,[1] **Michael Ng,**[4] **Huaqing Min,**[1] **Qingyao Wu**[1‡]
[1]School of Software Engineering, South China University of Technology, China
[2]CVTE Research, Guangzhou Shiyuan Electronics Co., Ltd., China
[3]Artificial Intelligence Innovation Business, Baidu Inc., China
[4]Department of Mathematics, Hong Kong Baptist University, Hong Kong, China
{yan.yuguang,secaojiezhang}@mail.scut.edu.cn, {mingkuitan,qyw,hqmin}@scut.edu.cn,
xuyanwu@baidu.com, mng@math.hkbu.edu.hk

Abstract

The issue of data imbalance occurs in many real-world applications especially in medical diagnosis, where normal cases are usually much more than the abnormal cases. To alleviate this issue, one of the most important approaches is the oversampling method, which seeks to synthesize minority class samples to balance the numbers of different classes. However, existing methods barely consider global geometric information involved in the distribution of minority class samples, and thus may incur distribution mismatching between real and synthetic samples. In this paper, relying on optimal transport (Villani 2008), we propose an oversampling method by exploiting global geometric information of data to make synthetic samples follow a similar distribution to that of minority class samples. Moreover, we introduce a novel regularization based on synthetic samples and shift the distribution of minority class samples according to loss information. Experiments on toy and real-world data sets demonstrate the efficacy of our proposed method in terms of multiple metrics.

Introduction

The imbalanced data issue occurs in many real-world applications, where the samples of one class are much more than samples of other classes (He and Garcia 2008; Branco, Torgo, and Ribeiro 2016; Lin et al. 2018; González et al. 2019). Especially in the area of medical diagnosis, the abnormal samples with a disease are expensive and difficult to collect, while normal samples are much easier to obtain. As a result, we usually face an imbalanced learning problem where normal samples are much more than the abnormal ones (Bhattacharya, Rajan, and Shrivastava 2017).

Standard machine learning methods usually focus on reducing loss over the whole training data set. These methods usually pay more attention to the training loss on majority class samples while omitting the minority class samples, thus fail to achieve promising performance. This issue becomes even worse in medical diagnosis, since misclassifying an abnormal one is much severer than misclassifying a normal one, which will delay the medical treatment.

To alleviate the imbalanced issue, several kinds of algorithms have been proposed in the last decades. Among these methods, oversampling attracts much attention because of its simplicity and efficacy (Fernández et al. 2018). Oversampling aims to synthesize minority class samples to balance the numbers of different classes, so that standard machine learning methods can be performed on the augmented data set. Oversampling methods usually synthesize new samples based on a minority class sample and its nearest neighbors. However, they barely consider global geometric information in the distribution of minority class samples, and thus may incur distribution mismatching between real and synthetic samples obtained by oversampling methods.

In this paper, we aim to exploit global geometric information of data to oversample minority class samples via optimal transport (Villani 2008). Motivated by this, we propose a novel oversampling method called **O**ptimal **T**ransport for **O**ver**S**ampling (OTOS), which applies optimal transport to synthesize samples that follow a similar distribution to the one of minority class samples. Specifically, we move random points from a prior distribution to that of minority class samples, as shown in Figure 1, so that the transported samples can be taken as synthetic minority class data for training. In addition, we introduce a regularization based on the transported samples for optimal transport, and leverage loss information to concentrate more on those minority class samples close to the decision boundary.

We apply a projected gradient method to optimize the resultant constrained problem, and conduct extensive experiments on both toy and real-world data sets, including benchmark data from LIBSVM[1] and medical image data. Multiple metrics regarding imbalanced learning are adopted to demonstrate the effectiveness of our proposed method.

The principal contributions are summarized as follows:

- We exploit global geometric information of data via optimal transport to guarantee distribution matching between synthetic and real minority class samples.

*The co-first author.

†This work was done when Yuguang Yan was an intern at Medical Image and Signal Processing Group, CVTE Research.

‡The corresponding author.

[1]https://www.csie.ntu.edu.tw/∼cjlin/libsvmtools/datasets

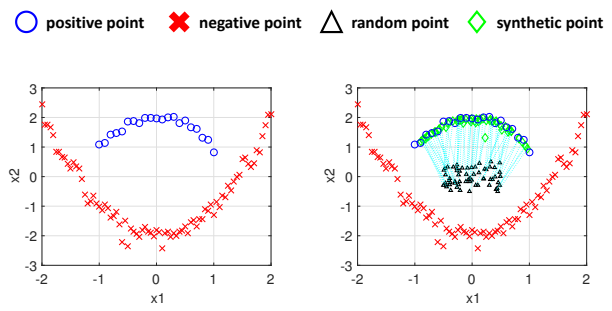

Figure 1: An illustration of our idea. Positive and negative points are minority and majority class samples, respectively. (a) Positive points and negative points. (b) Positive points, negative points, random points, and synthetic points obtained by optimal transport. Dot lines indicate the transport plan from random points to synthetic points, and random points are drawn from the uniform distribution $\mathcal{U}(-0.5, 0.5)$.

- We shift the empirical distribution of minority class samples based on training loss, rather than adopting a uniform distribution for them, which is commonly used in existing works (Courty et al. 2017b; Yan et al. 2018).
- We design a novel regularization with transported samples to avoid synthesizing noisy samples, which is achieved by enlarging the difference between the predicted values of a transported sample and a majority class sample.

Related Studies

Imbalanced Learning

Most existing methods of imbalanced learning belong to two categories: cost-sensitive approaches and oversampling approaches (Lemaître, Nogueira, and Aridas 2017). Cost-sensitive approaches try to assign different weights for classes, so that losses on minority class samples are emphasized to contribute more for training (Liu et al. 2017; Zhang et al. 2018). Nevertheless, these methods omit geometric information hidden in the structure of training data, which limits their performance on imbalanced problems.

Oversampling approaches seek to augment minority class data to balance the numbers of two classes (Das, Krishnan, and Cook 2015; Sen et al. 2016; Abdi and Hashemi 2016; Pérez-Ortiz et al. 2016; Bellinger, Drummond, and Japkowicz 2018). Among them, SMOTE is a classical method, which takes linear interpolations of a minority class sample and its nearest neighbors as new training samples (Chawla et al. 2002). bSMOTE improves SMOTE by finding so-called DANGER samples from the minority class (Han, Wang, and Mao 2005). ADASYN further extends SMOTE by considering different effects of training samples (He et al. 2008). In (Das, Krishnan, and Cook 2015), a Gibbs oversampling method is proposed. MWMOTE finds informative minority class samples and oversample data based on a clustering approach (Barua et al. 2014). (Peng 2015) proposes an adaptive sampling method to form multiple classifiers over different subsets. (Fernández et al. 2018) provides a summary regarding recent advances of SMOTE.

Compared to the above oversampling methods, the key differences in our work are two folds: firstly, we exploit global geometric information of data via optimal transport, which makes synthetic samples follow a similar distribution to that of minority class samples, while existing oversampling methods barely consider the global geometric information; secondly, our proposed method provides a global oversampling paradigm based on the Wasserstein barycenter (Cuturi and Doucet 2014; Peyré and Cuturi 2017), which does not rely on nearest neighbor searching commonly used in existing oversampling methods.

Optimal Transport

Optimal transport (Villani 2008), which was firstly introduced by Monge in (Monge 1781), originally aims to study how to transport mass into a given place with the minimal cost. After that, Kantorovitch further developed optimal transport and applied it for economic applications (Kantorovitch 1958). The minimal cost in optimal transport is also known as the Wasserstein distance or Earth Mover Distance. To efficiently solve the optimization problem involved in optimal transport, some fast numerical algorithms are proposed in (Cuturi 2013; Benamou et al. 2015). Recently, optimal transport has been actively applied in machine learning problems (Peyré and Cuturi 2017). In (Courty et al. 2017a; 2017b), optimal transport is applied in unsupervised domain adaptation. Yan et al. leveraged optimal transport to address the problem of heterogeneous domain adaptation in (Yan et al. 2018). The Wasserstein distance is also used in the generative model to measure the distance between two distributions (Arjovsky, Chintala, and Bottou 2017).

Methodology

Problem Statement and Notations

Given training data $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top \in \mathbb{R}^{n\times d}$ and their labels $\mathbf{y} = [y_1, \ldots, y_n]^\top \in \{+1, -1\}^n$, where n is the number of training data, and d is the number of features. Among the training data, the positive data are represented as $\mathbf{X}^+ = [\mathbf{x}_1^+, \ldots, \mathbf{x}_{n^+}^+]^\top \in \mathbb{R}^{n^+\times d}$ with n^+ being the number of positive samples, and the negative data are represented as $\mathbf{X}^- = [\mathbf{x}_1^-, \ldots, \mathbf{x}_{n^-}^-]^\top \in \mathbb{R}^{n^-\times d}$ with n^- being the number of negative samples. Without loss of generality, we have $n^+ \ll n^-$, which means that the number of positive samples is much smaller than that of negative samples. We also call the positive label as the minority class, and the negative label as the majority class.

$\mathbf{1}_n$ denotes a vector in the space $\mathbb{R}^n$ with all the elements being 1. For a vector $\mathbf{a}$, $\mathrm{diag}(\mathbf{a})$ is a diagonal matrix with the diagonal elements being $\mathbf{a}$. For a matrix $\mathbf{A}$, let A_{ij} be the (i,j)-th element of $\mathbf{A}$. The trace of the square matrix $\mathbf{A}$ is defined as $\mathrm{tr}(\mathbf{A}) = \sum_i A_{ii}$. For two matrices $\mathbf{A}$ and $\mathbf{B}$, let $\mathbf{A}\otimes\mathbf{B}$ be the Kronecker product, $\mathbf{A}\odot\mathbf{B}$ be the element-wise product, and the inner product is defines as

$$\langle \mathbf{A}, \mathbf{B}\rangle = \sum_i \sum_j A_{ij}B_{ij} = \mathrm{tr}(\mathbf{A}^\top\mathbf{B}) = \mathrm{tr}(\mathbf{A}\mathbf{B}^\top).$$

Optimal Transport for Oversampling

Optimal transport aims to transport samples from one distribution to another distribution with the minimal transport cost (Peyré and Cuturi 2017; Courty et al. 2017b). As a result, we can obtain more samples in the target distribution. Motivated by this, we apply optimal transport to transport some random samples into the distribution of minority class samples to augment training data.

Specifically, let $\boldsymbol{\mu}^r$ be the empirical distributions of a set of random vectors $\{\mathbf{x}_i^r\}_{i=1}^{n^r}$ drawn from a prior distribution, and $\boldsymbol{\mu}^+$ be the empirical distribution of positive samples, which are minority class samples in this paper. We propose to transport samples from $\boldsymbol{\mu}^r$ into $\boldsymbol{\mu}^+$ to augment positive samples. Define δ be the Dirac function at one point, p_i^r and p_i^+ be probability masses with the simplex constraints $\sum_{i=1}^{n^r} p_i^r = 1$ and $\sum_{i=1}^{n^+} p_i^+ = 1$. The two empirical distributions are written as

$$\boldsymbol{\mu}^r = \sum_{i=1}^{n^r} p_i^r \delta_i^r, \quad \boldsymbol{\mu}^+ = \sum_{i=1}^{n^+} p_i^+ \delta_i^+. \tag{1}$$

A transport plan is represented as a joint distribution, which is in the following definition domain:

$$\mathcal{T} = \{\mathbf{T} \in (\mathbb{R}^+)^{n^r \times n^+} | \mathbf{T}\mathbf{1}_{n^+} = \boldsymbol{\mu}^r, \mathbf{T}^\top \mathbf{1}_{n^r} = \boldsymbol{\mu}^+\}, \tag{2}$$

and the entropy of $\mathbf{T}$ is defined as

$$H(\mathbf{T}) = -\sum_{ij} T_{ij}(\log T_{ij} - 1).$$

Optimal transport aims to find a transport matrix with the minimal cost, which is modelled as the following optimization problem:

$$\min_{\mathbf{T}\in\mathcal{T}} \ \langle \mathbf{T}, \mathbf{C}^{r+} \rangle - \epsilon H(\mathbf{T}), \tag{3}$$

where ϵ is a trade-off parameter, $\mathbf{C}^{r+}$ is the cost matrix with C_{ij}^{r+} being defined as

$$C_{ij}^{r+} = c(\mathbf{x}_i^r, \mathbf{x}_j^+) = \|\mathbf{x}_i^r - \mathbf{x}_j^+\|_2^2, \tag{4}$$

and the entropic regularization $H(\mathbf{T})$ is used to smoothen the solution and speed up the optimization (Cuturi 2013). In this way, the transported samples follow a similar distribution to that of positive samples and can be taken as augmented positive data. Specifically, let $n^r = n^- - n^+$, so the numbers of samples from two classes are balanced.

Oversampling by Data Transport

After obtaining the optimal transport matrix $\mathbf{T}$, we can transport the random vectors into the distribution of the positive samples based on the Wasserstein barycenter, which represents the random points in the distribution of positive data (Cuturi and Doucet 2014; Peyré and Cuturi 2017). Specifically, for the point $\mathbf{x}_i^r$, its representation in the positive data distribution is denoted by $\hat{\mathbf{x}}_i^r$, which is obtained by solving the following optimization problem:

$$\begin{aligned} \hat{\mathbf{x}}_i^r &= \arg\min_{\mathbf{x}\in\mathbb{R}^d} \sum_j T_{ij} c(\mathbf{x}, \mathbf{x}_j^+) \\ &= \arg\min_{\mathbf{x}\in\mathbb{R}^d} \sum_j T_{ij} \|\mathbf{x} - \mathbf{x}_j^+\|_2^2. \end{aligned} \tag{5}$$

Taking the partial derivative *w.r.t.* $\mathbf{x}$ to zero, we obtain

$$\sum_j T_{ij}\mathbf{x} - \sum_j T_{ij}\mathbf{x}_j^+ = \mathbf{0}, \tag{6}$$

and the solution is given by the following:

$$\hat{\mathbf{x}}_i^r = \frac{\sum_j T_{ij}\mathbf{x}_j^+}{\sum_j T_{ij}}. \tag{7}$$

Based on this, the transported data matrix can be written as

$$\hat{\mathbf{X}} = \mathrm{diag}(\mathbf{T}\mathbf{1}_{n^+})^{-1}\mathbf{T}\mathbf{X}^+ = \mathrm{diag}(\boldsymbol{\mu}^r)^{-1}\mathbf{T}\mathbf{X}^+. \tag{8}$$

For simplicity, define $\mathbf{D}_r = \mathrm{diag}(\boldsymbol{\mu}^r)^{-1}$, and Eq. (8) can be rewritten as

$$\hat{\mathbf{X}} = \mathbf{D}_r\mathbf{T}\mathbf{X}^+. \tag{9}$$

From Eq. (9), we observe that each synthetic positive sample is a convex combination of multiple positive samples. This indicates that our approach leverages global information from all the given minority class samples, which differs from nearest neighbor searching commonly used in existing oversampling methods. Therefore, global geometric information extracted from minority class samples are exploited in our approach.

Distribution Shifting based on Training Loss

For $\boldsymbol{\mu}^r$, we simply adopt a uniform distribution, *i.e.*, $\boldsymbol{\mu}^r = [\frac{1}{n^r}, \ldots, \frac{1}{n^r}]^\top$. For $\boldsymbol{\mu}^+$, without more information about the underlying distribution of positive data, one usually adopts a uniform distribution (Courty et al. 2017b; Yan et al. 2018), *i.e.*, $\boldsymbol{\mu}^+ = [\frac{1}{n^+}, \ldots, \frac{1}{n^+}]^\top$, in Problem (3). However, this approach omits the loss information of samples. For SVM, a sample closer to the decision boundary usually has a larger loss and more information than one far away from the decision boundary.

Motivated by this intuition, we firstly pretrain an SVM classifier, and then shift the positive data distribution based on training loss to concentrate more on the ones with larger loss values. Formally, we pretrain an SVM classifier with the hinge loss, which is given as

$$\begin{aligned} \min_{\mathbf{w},\boldsymbol{\xi}} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & y_i\mathbf{w}^\top\mathbf{x}_i \geq 1 - \xi_i,\ \xi_i \geq 0, \quad i = 1, \ldots, n, \end{aligned} \tag{10}$$

where C is the trade-off parameter for loss. Let $\xi_i^+ = \max(1 - y_i^+\mathbf{w}^\top\mathbf{x}_i^+, 0)$ be the hinge loss of the positive sample $\mathbf{x}_i^+$, we apply the softmax function to shift the distribution of $\{\mathbf{x}_i^+\}_{i=1}^{n_+}$ as

$$\boldsymbol{\mu}^+ = \Big[\frac{e^{\xi_1^+}}{\sum_{i=1}^{n_+} e^{\xi_i^+}}, \ \ldots, \ \frac{e^{\xi_{n_+}^+}}{\sum_{i=1}^{n_+} e^{\xi_i^+}}\Big]^\top. \tag{11}$$

Consequently, the positive samples with larger losses contribute more in optimal transport, making the transported samples more informative for training an effective classifier.

Learning with Transported Samples

Concentrating on those samples close to the decision boundary can take better advantage of training samples. However, a potential issue is that a few minority class samples that are close to majority class samples may highly affect the results of optimal transport, making some transported samples too close to majority class samples and confusing the training of the classifier. This issue will become even severer with noisy minority samples. In order to alleviate this, for each pair of a transported positive sample and a negative sample, we propose to enlarge the difference between the predicted values of them obtained by the pretrained model. To achieve this, we design the following regularization:

$$\Omega(\mathbf{T}) = \frac{1}{2}\sum_{i=1}^{n^r}\sum_{j=1}^{n^-}\Big|\Big|\big(\mathbf{w}^\top\hat{\mathbf{x}}_i - \mathbf{w}^\top\mathbf{x}_j^-\big) - 1\Big|\Big|^2, \tag{12}$$

and seek to solve the optimization problem as

$$\min_{\mathbf{T}\in\mathcal{T}} \quad \mathcal{L}(\mathbf{T}) \triangleq \Omega(\mathbf{T}) + \lambda\langle\mathbf{T}, \mathbf{C}^{r+}\rangle - \epsilon H(\mathbf{T}), \tag{13}$$

where λ and ϵ are trade-off parameters. By rearranging the above objective function, we obtain

$$\begin{aligned}\mathcal{L}(\mathbf{T}) &= \frac{1}{2}\sum_{i=1}^{n^r}\sum_{j=1}^{n^-}\Big|\Big|\big(\mathbf{w}^\top\hat{\mathbf{x}}_i - \mathbf{w}^\top\mathbf{x}_j^-\big) - 1\Big|\Big|^2 \\ &\quad + \lambda\mathrm{tr}(\mathbf{T}^\top\mathbf{C}^{r+}) - \epsilon H(\mathbf{T}) \\ &= \frac{1}{2}n^-\mathrm{tr}\Big(\hat{\mathbf{X}}\mathbf{w}\mathbf{w}^\top\hat{\mathbf{X}}^\top\Big) - n^-\mathrm{tr}\Big(\hat{\mathbf{X}}(\mathbf{1}_{n^r}^\top\otimes\mathbf{w})\Big) \\ &\quad - \mathrm{tr}\Big(\hat{\mathbf{X}}\big(\mathbf{1}_{n^r}^\top\otimes(\mathbf{w}\mathbf{w}^\top(\mathbf{X}^-)^\top\mathbf{1}_{n^-})\big)\Big) \\ &\quad + \lambda\mathrm{tr}(\mathbf{T}^\top\mathbf{C}^{r+}) - \epsilon H(\mathbf{T}) + \text{constant}.\end{aligned} \tag{14}$$

By substituting Eq. (9) into Eq. (14), we simplify $\mathcal{L}(\mathbf{T})$ as

$$\begin{aligned}\mathcal{L}(\mathbf{T}) &= \frac{1}{2}n^-\mathrm{tr}\Big(\mathbf{D}_r\mathbf{T}\mathbf{X}^+\mathbf{w}\mathbf{w}^\top(\mathbf{X}^+)^\top\mathbf{T}^\top\mathbf{D}_r^\top\Big) \\ &\quad - \mathrm{tr}\Big(\mathbf{D}_r\mathbf{T}\mathbf{X}^+\big(\mathbf{1}_{n^r}^\top\otimes((\mathbf{w}\mathbf{w}^\top(\mathbf{X}^-)^\top\mathbf{1}_{n^-}) + n^-\mathbf{w})\big)\Big) \\ &\quad + \lambda\mathrm{tr}(\mathbf{T}^\top\mathbf{C}^{r+}) - \epsilon H(\mathbf{T}) + \text{constant} \\ &= \frac{1}{2}n^-\mathrm{tr}\Big(\mathbf{T}\mathbf{X}^+\mathbf{w}\mathbf{w}^\top(\mathbf{X}^+)^\top\mathbf{T}^\top\mathbf{D}_r^\top\mathbf{D}_r\Big) \\ &\quad - \mathrm{tr}\Big(\mathbf{T}\mathbf{X}^+\big(\mathbf{1}_{n^r}^\top\otimes((\mathbf{w}\mathbf{w}^\top(\mathbf{X}^-)^\top\mathbf{1}_{n^-}) + n^-\mathbf{w})\big)\mathbf{D}_r\Big) \\ &\quad + \lambda\mathrm{tr}(\mathbf{T}^\top\mathbf{C}^{r+}) - \epsilon H(\mathbf{T}) + \text{constant}.\end{aligned} \tag{15}$$

We define the matrix variables $\mathbf{\Theta}$, $\mathbf{\Phi}$ and $\mathbf{\Psi}$ as

$$\begin{aligned}\mathbf{\Theta} &= \lambda(\mathbf{C}^{r+})^\top - \mathbf{X}^+\big(\mathbf{1}_{n^r}^\top\otimes((\mathbf{w}\mathbf{w}^\top(\mathbf{X}^-)^\top\mathbf{1}_{n^-}) + n^-\mathbf{w})\big)\mathbf{D}_r, \\ \mathbf{\Phi} &= \mathbf{X}^+\mathbf{w}\mathbf{w}^\top(\mathbf{X}^+)^\top, \\ \mathbf{\Psi} &= \mathbf{D}_r^\top\mathbf{D}_r.\end{aligned} \tag{16}$$

As a result, $\mathcal{L}(\mathbf{T})$ is reformulated as

$$\mathcal{L}(\mathbf{T}) = \frac{1}{2}n^-\mathrm{tr}\Big(\mathbf{T}\mathbf{\Phi}\mathbf{T}^\top\mathbf{\Psi}\Big) + \mathrm{tr}(\mathbf{T}\mathbf{\Theta}) - \epsilon H(\mathbf{T}) + \text{constant}. \tag{17}$$

Optimization Details

For simplicity, we define

$$f(\mathbf{T}) = \frac{1}{2}n^-\mathrm{tr}\Big(\mathbf{T}\mathbf{\Phi}\mathbf{T}^\top\mathbf{\Psi}\Big) + \mathrm{tr}(\mathbf{T}\mathbf{\Theta}), \tag{18}$$

and $\mathcal{L}(\mathbf{T})$ can be rewritten as

$$\mathcal{L}(\mathbf{T}) = f(\mathbf{T}) - \epsilon H(\mathbf{T}) + \text{constant}. \tag{19}$$

Minimizing $\mathcal{L}(\mathbf{T})$ *w.r.t.* $\mathbf{T}$ is non-trivial because of the equality constraints. To solve it, we apply a projected gradient descent algorithm based on the exponentiated gradient and the Kullback-Leibler divergence (Benamou et al. 2015; Peyré, Cuturi, and Solomon 2016). Specifically, at the τ-th iteration, we firstly update $\mathbf{T}_\tau$ by the exponentiated gradient method as follows:

$$\tilde{\mathbf{T}}_\tau := \mathbf{T}_\tau \odot \exp\Big(-\alpha\nabla\mathcal{L}(\mathbf{T}_\tau)\Big), \tag{20}$$

where $\alpha > 0$ is a step size. After that, we project $\tilde{\mathbf{T}}_\tau$ into the definition domain $\mathcal{T}$ with the Kullback-Leibler metric as

$$\mathbf{T}_{\tau+1} := \Pi_{\mathcal{T}}^{\mathrm{KL}}(\tilde{\mathbf{T}}_\tau) = \arg\min_{\mathbf{T}'\in\mathcal{T}}\mathrm{KL}(\mathbf{T}'|\tilde{\mathbf{T}}_\tau). \tag{21}$$

According to (Benamou et al. 2015), the projection operation in Eq. (21) can be rewritten as the following regularized optimal transport problem:

$$\begin{aligned}\mathbf{T}_{\tau+1} &:= \Pi_{\mathcal{T}}^{\mathrm{KL}}(\tilde{\mathbf{T}}_\tau) \\ &= \arg\min_{\mathbf{T}'\in\mathcal{T}}\langle-\epsilon\log(\tilde{\mathbf{T}}_\tau), \mathbf{T}'\rangle - \epsilon H(\mathbf{T}'),\end{aligned} \tag{22}$$

which can be efficiently solved by the Sinkhorn's fixed point algorithm (Sinkhorn 1967; Cuturi 2013). In Problem (22), the transport cost matrix $-\epsilon\log(\tilde{\mathbf{T}}_\tau)$ can be simplified as

$$\begin{aligned}-\epsilon\log(\tilde{\mathbf{T}}_\tau) &= -\epsilon\log\Big(\mathbf{T}_\tau\odot\exp\big(-\alpha\nabla\mathcal{L}(\mathbf{T}_\tau)\big)\Big) \\ &= \nabla f(\mathbf{T}_\tau) \\ &= \mathbf{\Theta}^\top + n^-\mathbf{\Psi}\mathbf{T}_\tau\mathbf{\Phi},\end{aligned} \tag{23}$$

where we set $\epsilon\alpha = 1$.

Algorithm 1 summarizes the procedure of OTOS.

Algorithm 1 Optimal Transport for OverSampling (OTOS)

1: Initialize $\mathbf{T} = \boldsymbol{\mu}^r(\boldsymbol{\mu}^+)^\top$, $\tau = 1$.
2: Train $\mathbf{w}$ over $\mathbf{X}^+$ and $\mathbf{X}^-$ by solving Problem (10).
3: Compute $\boldsymbol{\mu}^+$ based on Eq. (11).
4: Construct $\mathbf{\Theta}$, $\mathbf{\Phi}$ and $\mathbf{\Psi}$ according to Eq. (16).
5: **repeat**
6: Calculate Eq. (23) based on $\mathbf{T}_\tau$.
7: Obtain $\mathbf{T}_{\tau+1}$ by solving Problem (22).
8: $\tau := \tau + 1$.
9: **until** Convergence.
10: Synthesize samples $\hat{\mathbf{X}}$ based on Eq. (9).
11: Train $\mathbf{w}$ over $\mathbf{X}^+$, $\mathbf{X}^-$ and $\hat{\mathbf{X}}$.

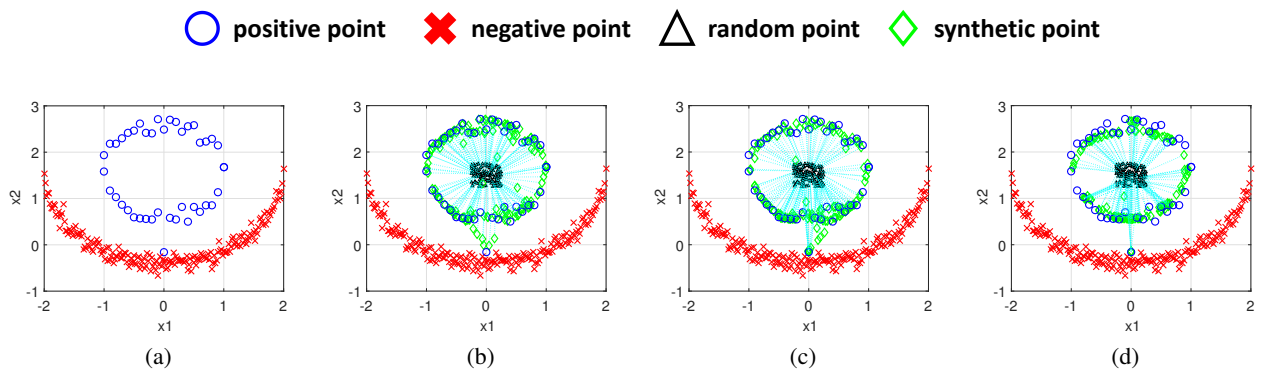

Figure 2: Results on a 2D toy data set. Positive and negative points are minority and majority class samples, respectively. (a) Positive points and negative points. (b) Positive points, negative points, and synthetic points obtained by OTOS-α. (c) Positive points, negative points, and synthetic points obtained by OTOS-β. (d) Positive points, negative points, and synthetic points obtained by OTOS. Dot lines indicate the transport plan from random points to synthetic points.

Experiments

To verify our proposed method, we firstly conduct empirical studies on a toy data set, and then apply our method on several real-world data sets, which include benchmark data sets from LIBSVM and medical image data.

Compared Methods

- **SVM**. We perform SVM (Fan et al. 2008) on an imbalanced data set to train a classifier. SVM is a straightforward method without considering the imbalance issue.
- **ROS**. Random oversampling (ROS) randomly selects samples from the minority class and adds them to training data.
- **SMOTE**. In the method of synthetic minority oversampling technique (SMOTE) (Chawla et al. 2002), for a minority class sample, the linear interpolations of it and its nearest neighbors are taken as new samples for training.
- **bSMOTE**. Borderline-SMOTE (bSMOTE) (Han, Wang, and Mao 2005) firstly finds some so-called DANGER samples for the minority class, and then takes the linear interpolations of them and their nearest neighbors as new training samples.
- **ADASYN**. Adaptive synthetic (ADASYN) (He et al. 2008) extends SMOTE by adaptively adjusting the numbers of artificial samples for each minority class sample. If a minority class sample has more nearest neighbors belonging to the majority class, ADASYN will synthesize more artificial samples based on the sample and its nearest neighbors.
- **MWMOTE**. Majority weighted minority oversampling technique (MWMOTE) (Barua et al. 2014) identifies informative minority class samples, and synthesizes samples according to the weighted informative minority class samples based on a clustering method.
- **OTOS-α**. OTOS-α is a simplified version of OTOS. OTOS-α adopts a uniform distribution for $\boldsymbol{\mu}^+$. It firstly obtains $\mathbf{T}$ by solving Problem (3), and then synthesizes samples based on Eq. (9).
- **OTOS-β**. OTOS-β is also a simplified version of OTOS. OTOS-β replaces the uniform distribution of $\boldsymbol{\mu}^+$ in OTOS-α by the distribution in Eq. (11).

Experiments on Toy Data

To demonstrate the effects of the proposed method, we firstly conduct experiments on a 2D toy data set, in which 42 positive points are minority class samples, and 201 negative points are majority class samples. Figure 2(a) shows the positive and negative points.

From Figure 2(b), the synthetic points obtained by OTOS-α follow a similar distribution to that of the minority class samples. Figure 2(c) presents synthetic points obtained by OTOS-β. Compared to OTOS-α, OTOS-β synthesizes more samples near to those samples that are close to the borderline between two classes, since it pays more attention to those samples based on the distribution in Eq. (11). Nevertheless, the nearest positive point to the negative ones highly affects the result of OTOS-β, making it sensitive to noisy samples. Figure 2(d) shows the results of OTOS. Compared to OTOS-α, OTOS synthesizes more samples close to the borderline between two classes. In addition, by introducing the regularization in Eq. (12) into optimal transport, OTOS is more robust to noisy points than OTOS-β.

Experiments on Real-World Data

Data Sets Tables 1 and 2 present the statistical information of the adopted benchmark and medical image data sets, respectively. In the tables, "size" represents $\#samples \times \#features$, and "ratio" is $\frac{\#majority\ class\ samples}{\#minority\ class\ samples}$. In the following, we describe the details of the data sets.

- **Benchmark data**. We adopt six benchmark data sets from LIBSVM (australian, breast-cancer, diabetes, german, svmguide2, and svmguide4) in the experiments. For the data sets that are split into training and testing subsets, we only adopt training subsets for simplicity. For the multi-class data sets, we take one class as positive and the others as negative to construct imbalanced binary classification tasks.

Table 1: Statistical information of the benchmark data sets.

name	size	ratio
australian	690×14	1.25
breast-cancer	683×10	1.86
diabetes	768×8	1.87
german	$1,000 \times 24$	2.33
svmguide2	391×20	1.30
svmguide4	300×10	4.66

Table 2: Statistical information of the medical image data sets.

name	size	ratio
ORIGA	$650 \times 2,048$	2.87
iSee-AMD	$8,480 \times 2,048$	10.78
iSee-DR	$8,016 \times 2,048$	30.31
iSee-glaucoma	$8,208 \times 2,048$	17.32

- **Medical data**. We also use four fundus image data sets, among which iSee-AMD, iSee-DR, iSee-glaucoma are used to detect Age-Related Macular Degeneration (AMD), Diabetic Retinopathy (DR) and glaucoma, respectively, and ORIGA is used to detect glaucoma. We extract features from the pool5 layer of the ResNet-152 (He et al. 2016) pretrained on ImageNet (Deng et al. 2009) to obtain a 2,048-dimensional vector for one image.

Experimental Settings For all the compared methods, we synthesize minority class samples until that the numbers of minority and majority class samples are the same, and use linear SVM with the default parameter $C = 1$ as the classifier. For our method, we draw random samples from a prior uniform distribution $\mathcal{U}(0, 1)$. The parameters λ and ϵ are selected in the range $10^{\{-1,0,1,2,3,4,5\}}$, and the best results are adopted. We repeat all the experiments 10 times and report the mean and standard derivation values, and results of each time are obtained by the mean of 10-fold cross-validation.

Evaluation Metrics We adopt multiple evaluation metrics to test the performance of the proposed method. Specifically, let y_i be a true label and $\hat{y}_i$ be a predicted label, we count the numbers of true positive (TP), false positive (FP), false negative (FN) and true negative (TN) samples, which are formally defined as follows:

$$\begin{aligned} TP &= \big|\{\mathbf{x}_i | y_i = +1 \wedge \hat{y}_i = +1, \quad i = 1, \ldots, n\}\big|, \\ FP &= \big|\{\mathbf{x}_i | y_i = -1 \wedge \hat{y}_i = +1, \quad i = 1, \ldots, n\}\big|, \\ FN &= \big|\{\mathbf{x}_i | y_i = +1 \wedge \hat{y}_i = -1, \quad i = 1, \ldots, n\}\big|, \\ TN &= \big|\{\mathbf{x}_i | y_i = -1 \wedge \hat{y}_i = -1, \quad i = 1, \ldots, n\}\big|. \end{aligned} \tag{24}$$

Based on the above notations, we define the following metrics:

$$Sensitivity = \frac{TP}{TP + FN}, \tag{25}$$

$$F1 = \frac{2 \cdot TP}{2 \cdot TP + FN + FP}, \tag{26}$$

$$G\text{-}mean = \sqrt{\frac{TP}{TP + FN} \cdot \frac{TN}{TN + FP}}. \tag{27}$$

Table 3: Significance test results (win/tie/loss) with paired t-test at 0.05 level for OTOS against other methods.

method	Sensitivity	F1	G-mean
SVM	9/0/1	10/0/0	10/0/0
ROS	7/2/1	8/2/0	8/2/0
SMOTE	7/3/0	9/1/0	8/2/0
bSMOTE	9/0/1	9/1/0	10/0/0
ADASYN	5/3/2	8/2/0	6/4/0
MWMOTE	9/1/0	7/3/0	8/2/0
OTOS-α	8/2/0	4/5/1	6/4/0
OTOS-β	4/6/0	4/6/0	4/6/0

Results and Discussions Table 3 summarizes the significance test results for OTOS against other methods on all the metrics and adopted data sets, and Tables 4 and 5 present the results on the benchmark and medical image data sets, respectively. The best results are indicated with boldface type and the second best results are underlined. We also apply paired t-test at 0.05 level for performing the significance tests between OTOS and other methods. "•" means that OTOS significantly outperforms a baseline method, and "◦" means that a baseline method significantly outperforms OTOS. We draw several interesting observations as follows.

- OTOS achieves the best Sensitivity results on 7 data sets, the best F1 results on 7 data sets, and the best G-mean results on all the adopted data sets. This demonstrates the effectiveness of OTOS.
- For all the three metrics, most of the best or the second best results are obtained by OTOS-α, OTOS-β or OTOS, which indicates that optimal transport is able to synthesize high-quality minority class samples to enhance the classification performance.
- Compared to OTOS-α and OTOS-β, OTOS usually gets better or highly comparable results, which verifies the effects of the shifted distribution in Eq. (11) and the regularization in Eq. (12).
- OTOS-β performs better than OTOS-α on most experiments, which validates that concentrating on minority class samples close to a borderline between two classes is beneficial for training an effective classifier.
- From Table 3, OTOS significantly outperforms other methods on most comparisons, which demonstrates the consistent superiority of OTOS over other methods.

Conclusion

In this paper, we propose a novel oversampling method for imbalanced data via optimal transport. We exploit global geometric information in the distribution of minority class samples, and take transported samples as synthetic minority class data. Moreover, we design a regularization based on the transported samples, and shift the distribution of minority class samples according to loss information. Experimental results on benchmark and medical image data sets demonstrate the effectiveness of our proposed method in terms of multiple metrics.

Table 4: Results on the benchmark data sets.

method	australian			breast-cancer		
	Sensitivity	F1	G-mean	Sensitivity	F1	G-mean
SVM	$0.9230 \pm 0.0040\circ$	$0.8479 \pm 0.0029\bullet$	$0.8575 \pm 0.0028\bullet$	$0.9630 \pm 0.0055\bullet$	$0.9556 \pm 0.0030\bullet$	$0.9675 \pm 0.0028\bullet$
ROS	$0.9247 \pm 0.0055\circ$	$0.8486 \pm 0.0032\bullet$	$0.8581 \pm 0.0027\bullet$	0.9722 ± 0.0062	0.9572 ± 0.0044	0.9703 ± 0.0038
SMOTE	0.9217 ± 0.0050	$0.8480 \pm 0.0032\bullet$	$0.8578 \pm 0.0027\bullet$	0.9730 ± 0.0057	0.9580 ± 0.0040	0.9709 ± 0.0034
bSMOTE	$\mathbf{0.9287 \pm 0.0032}\circ$	0.8509 ± 0.0020	$0.8604 \pm 0.0020\bullet$	$0.9626 \pm 0.0030\bullet$	$0.9545 \pm 0.0031\bullet$	$0.9668 \pm 0.0024\bullet$
ADASYN	$0.9247 \pm 0.0017\circ$	$0.8494 \pm 0.0011\bullet$	$0.8584 \pm 0.0011\bullet$	$\mathbf{0.9843 \pm 0.0059}\circ$	$0.9548 \pm 0.0049\bullet$	0.9713 ± 0.0039
MWMOTE	$0.9037 \pm 0.0122\bullet$	0.8485 ± 0.0080	0.8601 ± 0.0075	$0.9617 \pm 0.0045\bullet$	$0.9547 \pm 0.0029\bullet$	$0.9666 \pm 0.0026\bullet$
OTOS-α	0.9173 ± 0.0034	0.8520 ± 0.0030	0.8624 ± 0.0029	$0.9578 \pm 0.0077\bullet$	$0.9559 \pm 0.0046\bullet$	$0.9665 \pm 0.0041\bullet$
OTOS-β	0.9167 ± 0.0022	0.8512 ± 0.0017	0.8620 ± 0.0014	$0.9570 \pm 0.0038\bullet$	$0.9549 \pm 0.0028\bullet$	$0.9657 \pm 0.0024\bullet$
OTOS	0.9180 ± 0.0036	$\mathbf{0.8525 \pm 0.0020}$	$\mathbf{0.8627 \pm 0.0020}$	0.9713 ± 0.0042	$\mathbf{0.9604 \pm 0.0025}$	$\mathbf{0.9719 \pm 0.0022}$
method	diabetes			german		
	Sensitivity	F1	G-mean	Sensitivity	F1	G-mean
SVM	$0.5535 \pm 0.0091\bullet$	$0.6186 \pm 0.0069\bullet$	$0.6953 \pm 0.0060\bullet$	$0.4893 \pm 0.0072\bullet$	$0.5559 \pm 0.0074\bullet$	$0.6560 \pm 0.0056\bullet$
ROS	$0.7073 \pm 0.0071\bullet$	$0.6656 \pm 0.0059\bullet$	$0.7426 \pm 0.0049\bullet$	$0.7253 \pm 0.0106\bullet$	$0.6046 \pm 0.0042\bullet$	$0.7167 \pm 0.0034\bullet$
SMOTE	$0.7081 \pm 0.0076\bullet$	$0.6660 \pm 0.0059\bullet$	$0.7424 \pm 0.0048\bullet$	$0.7177 \pm 0.0079\bullet$	$0.6025 \pm 0.0056\bullet$	$0.7146 \pm 0.0046\bullet$
bSMOTE	$0.6777 \pm 0.0070\bullet$	$0.6559 \pm 0.0058\bullet$	$0.7337 \pm 0.0047\bullet$	$0.6900 \pm 0.0077\bullet$	$0.6023 \pm 0.0067\bullet$	$0.7137 \pm 0.0049\bullet$
ADASYN	$0.7365 \pm 0.0125\bullet$	0.6697 ± 0.0080	0.7459 ± 0.0069	$0.7407 \pm 0.0135\bullet$	$0.6057 \pm 0.0105\bullet$	$0.7173 \pm 0.0093\bullet$
MWMOTE	$0.6727 \pm 0.0102\bullet$	$0.6445 \pm 0.0067\bullet$	$0.7239 \pm 0.0055\bullet$	$0.5633 \pm 0.0215\bullet$	$0.5776 \pm 0.0142\bullet$	$0.6839 \pm 0.0122\bullet$
OTOS-α	$0.7115 \pm 0.0099\bullet$	0.6679 ± 0.0063	0.7443 ± 0.0051	$0.6977 \pm 0.0118\bullet$	$0.6041 \pm 0.0071\bullet$	$0.7161 \pm 0.0061\bullet$
OTOS-β	0.7546 ± 0.0088	0.6709 ± 0.0072	0.7466 ± 0.0061	$0.7507 \pm 0.0068\bullet$	$0.6094 \pm 0.0048\bullet$	$0.7204 \pm 0.0042\bullet$
OTOS	$\mathbf{0.7635 \pm 0.0098}$	$\mathbf{0.6733 \pm 0.0064}$	$\mathbf{0.7495 \pm 0.0053}$	$\mathbf{0.7590 \pm 0.0077}$	$\mathbf{0.6156 \pm 0.0055}$	$\mathbf{0.7255 \pm 0.0042}$
method	svmguide2			svmguide4		
	Sensitivity	F1	G-mean	Sensitivity	F1	G-mean
SVM	$0.0000 \pm 0.0000\bullet$	$0.0000 \pm 0.0000\bullet$	$0.0000 \pm 0.0000\bullet$	$0.0000 \pm 0.0000\bullet$	$0.0000 \pm 0.0000\bullet$	$0.0000 \pm 0.0000\bullet$
ROS	$0.7229 \pm 0.0170\bullet$	$0.7943 \pm 0.0165\bullet$	$0.8181 \pm 0.0133\bullet$	$0.7160 \pm 0.0227\bullet$	$0.4155 \pm 0.0178\bullet$	$0.6612 \pm 0.0156\bullet$
SMOTE	$0.7206 \pm 0.0191\bullet$	$0.7954 \pm 0.0124\bullet$	$0.8179 \pm 0.0105\bullet$	$0.7460 \pm 0.0165\bullet$	$0.4207 \pm 0.0104\bullet$	$0.6719 \pm 0.0091\bullet$
bSMOTE	$0.0494 \pm 0.0050\bullet$	$0.0902 \pm 0.0089\bullet$	$0.1689 \pm 0.0201\bullet$	$0.2700 \pm 0.0271\bullet$	$0.2913 \pm 0.0340\bullet$	$0.4185 \pm 0.0520\bullet$
ADASYN	$0.7671 \pm 0.0104\bullet$	$0.8180 \pm 0.0065\bullet$	$0.8371 \pm 0.0052\bullet$	$0.7140 \pm 0.0378\bullet$	$0.4760 \pm 0.0228\bullet$	$0.7088 \pm 0.0194\bullet$
MWMOTE	$0.6935 \pm 0.0356\bullet$	$0.7765 \pm 0.0190\bullet$	$0.8019 \pm 0.0154\bullet$	$0.6180 \pm 0.0114\bullet$	$0.3536 \pm 0.0056\bullet$	$0.6017 \pm 0.0059\bullet$
OTOS-α	$0.8188 \pm 0.0077\bullet$	$0.8382 \pm 0.0058\bullet$	$0.8553 \pm 0.0054\bullet$	$0.7460 \pm 0.0190\bullet$	$0.5802 \pm 0.0137\bullet$	$0.7758 \pm 0.0123\bullet$
OTOS-β	0.8312 ± 0.0079	0.8408 ± 0.0061	0.8575 ± 0.0058	$0.7720 \pm 0.0235\bullet$	$0.5751 \pm 0.0159\bullet$	$0.7807 \pm 0.0135\bullet$
OTOS	$\mathbf{0.8347 \pm 0.0098}$	$\mathbf{0.8442 \pm 0.0065}$	$\mathbf{0.8610 \pm 0.0060}$	$\mathbf{0.8020 \pm 0.0371}$	$\mathbf{0.6090 \pm 0.0248}$	$\mathbf{0.8037 \pm 0.0206}$

Table 5: Results on the medical image data sets.

method	ORIGA			iSee-AMD		
	Sensitivity	F1	G-mean	Sensitivity	F1	G-mean
SVM	$0.4000 \pm 0.0161\bullet$	$0.4074 \pm 0.0162\bullet$	$0.5647 \pm 0.0129\bullet$	$0.3267 \pm 0.0285\bullet$	$0.3928 \pm 0.0139\bullet$	$0.5581 \pm 0.0227\bullet$
ROS	0.4188 ± 0.0257	0.4257 ± 0.0205	0.5797 ± 0.0179	$0.4878 \pm 0.0158\bullet$	$0.3814 \pm 0.0057\bullet$	$0.6606 \pm 0.0093\bullet$
SMOTE	$0.4188 \pm 0.0312\bullet$	$0.4253 \pm 0.0251\bullet$	$0.5798 \pm 0.0234\bullet$	0.4997 ± 0.0122	$0.3880 \pm 0.0106\bullet$	0.6690 ± 0.0084
bSMOTE	$0.4169 \pm 0.0374\bullet$	$0.4235 \pm 0.0331\bullet$	$0.5779 \pm 0.0282\bullet$	$0.4925 \pm 0.0140\bullet$	$0.3905 \pm 0.0073\bullet$	$0.6649 \pm 0.0076\bullet$
ADASYN	0.4237 ± 0.0297	0.4275 ± 0.0275	0.5817 ± 0.0233	0.4957 ± 0.0177	$0.3892 \pm 0.0080\bullet$	$0.6665 \pm 0.0108\bullet$
MWMOTE	0.4281 ± 0.0172	0.4295 ± 0.0180	0.5824 ± 0.0150	$0.3469 \pm 0.0220\bullet$	0.3990 ± 0.0112	$0.5743 \pm 0.0169\bullet$
OTOS-α	0.4338 ± 0.0170	0.4372 ± 0.0138	0.5901 ± 0.0119	$0.4354 \pm 0.0160\bullet$	$\mathbf{0.4215 \pm 0.0079}\circ$	$0.6387 \pm 0.0104\bullet$
OTOS-β	0.4362 ± 0.0287	0.4404 ± 0.0256	0.5931 ± 0.0212	$0.4875 \pm 0.0150\bullet$	0.4030 ± 0.0094	$0.6653 \pm 0.0091\bullet$
OTOS	$\mathbf{0.4475 \pm 0.0265}$	$\mathbf{0.4462 \pm 0.0273}$	$\mathbf{0.5983 \pm 0.0208}$	$\mathbf{0.5057 \pm 0.0149}$	0.4064 ± 0.0066	$\mathbf{0.6764 \pm 0.0083}$
method	iSee-DR			iSee-glaucoma		
	Sensitivity	F1	G-mean	Sensitivity	F1	G-mean
SVM	$0.1688 \pm 0.0134\bullet$	$0.2258 \pm 0.0177\bullet$	$0.3912 \pm 0.0249\bullet$	$0.1148 \pm 0.0137\bullet$	$0.1574 \pm 0.0174\bullet$	$0.3239 \pm 0.0230\bullet$
ROS	$0.2428 \pm 0.0140\bullet$	$0.2101 \pm 0.0105\bullet$	$0.4775 \pm 0.0149\bullet$	$0.3080 \pm 0.0226\bullet$	$0.2178 \pm 0.0095\bullet$	$0.5254 \pm 0.0201\bullet$
SMOTE	$0.2432 \pm 0.0108\bullet$	$0.2135 \pm 0.0074\bullet$	$0.4754 \pm 0.0121\bullet$	$0.3066 \pm 0.0139\bullet$	$0.2200 \pm 0.0101\bullet$	$0.5262 \pm 0.0126\bullet$
bSMOTE	$0.2492 \pm 0.0171\bullet$	$0.2172 \pm 0.0117\bullet$	$0.4837 \pm 0.0173\bullet$	$0.3039 \pm 0.0115\bullet$	$0.2159 \pm 0.0073\bullet$	$0.5233 \pm 0.0103\bullet$
ADASYN	0.2500 ± 0.0148	$0.2209 \pm 0.0164\bullet$	0.4837 ± 0.0172	$0.2986 \pm 0.0201\bullet$	$0.2167 \pm 0.0118\bullet$	$0.5197 \pm 0.0185\bullet$
MWMOTE	$0.1784 \pm 0.0139\bullet$	$0.2291 \pm 0.0148\bullet$	$0.4099 \pm 0.0158\bullet$	$0.1902 \pm 0.0214\bullet$	$0.2069 \pm 0.0139\bullet$	$0.4221 \pm 0.0230\bullet$
OTOS-α	$0.2512 \pm 0.0148\bullet$	$\mathbf{0.2474 \pm 0.0098}$	0.4891 ± 0.0145	$0.2905 \pm 0.0181\bullet$	$\mathbf{0.2427 \pm 0.0114}$	$0.5176 \pm 0.0169\bullet$
OTOS-β	$\mathbf{0.2696 \pm 0.0138}$	0.2409 ± 0.0121	0.5034 ± 0.0124	0.3239 ± 0.0188	$0.2334 \pm 0.0099\bullet$	0.5416 ± 0.0149
OTOS	0.2652 ± 0.0156	0.2438 ± 0.0146	$\mathbf{0.5014 \pm 0.0164}$	$\mathbf{0.3332 \pm 0.0161}$	0.2406 ± 0.0096	$\mathbf{0.5492 \pm 0.0140}$

Acknowledgments

This work was supported by National Natural Science Foundation of China (61876208, 61502177 and 61602185), Recruitment Program for Young Professionals, Guangdong Provincial Scientific and Technological funds (2017B090901008, 2017A010101011, 2017B090910005), Fundamental Research Funds for the Central Universities D2172480, Pearl River S&T Nova Program of Guangzhou 201806010081, CCF-Tencent Open Research Fund RAGR20170105, Program for Guangdong Introducing Innovative and Enterpreneurial Teams 2017ZT07X183, HKRGC GRF (1202715, 12306616, 12200317, 12300218), HKBU RC-ICRS/16-17/03, and Guangzhou Shiyuan Electronics Co., Ltd. We also thank EyeSee Medical Science & Technology Chengdu Co., Ltd., iMed Team, and Singapore Eye Research Institute for providing research data.

References

Abdi, L., and Hashemi, S. 2016. To combat multi-class imbalanced problems by means of over-sampling techniques. *IEEE Transactions on Knowledge and Data Engineering* 28(1):238–251.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *ICML*, 214–223.

Barua, S.; Islam, M. M.; Yao, X.; and Murase, K. 2014. MWMOTE–majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Transactions on Knowledge and Data Engineering* 26(2):405–425.

Bellinger, C.; Drummond, C.; and Japkowicz, N. 2018. Manifold-based synthetic oversampling with manifold conformance estimation. *Machine Learning* 107(3):605–637.

Benamou, J.-D.; Carlier, G.; Cuturi, M.; Nenna, L.; and Peyré, G. 2015. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing* 37(2):A1111–A1138.

Bhattacharya, S.; Rajan, V.; and Shrivastava, H. 2017. Icu mortality prediction: A classification algorithm for imbalanced datasets. In *AAAI*, 1288–1294.

Branco, P.; Torgo, L.; and Ribeiro, R. P. 2016. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys* 49(2):31.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16:321–357.

Courty, N.; Flamary, R.; Habrard, A.; and Rakotomamonjy, A. 2017a. Joint distribution optimal transportation for domain adaptation. In *NIPS*, 3733–3742.

Courty, N.; Flamary, R.; Tuia, D.; and Rakotomamonjy, A. 2017b. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(9):1853–1865.

Cuturi, M., and Doucet, A. 2014. Fast computation of wasserstein barycenters. In *ICML*, 685–693.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *NIPS*, 2292–2300.

Das, B.; Krishnan, N. C.; and Cook, D. J. 2015. Racog and wracog: Two probabilistic oversampling techniques. *IEEE Transactions on Knowledge and Data Dngineering* 27(1):222.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R.; and Lin, C.-J. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* 9(Aug):1871–1874.

Fernández, A.; Garcia, S.; Herrera, F.; and Chawla, N. V. 2018. SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research* 61:863–905.

González, S.; García, S.; Li, S.-T.; and Herrera, F. 2019. Chain based sampling for monotonic imbalanced classification. *Information Sciences* 474:187–204.

Han, H.; Wang, W.-Y.; and Mao, B.-H. 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *ICIC*, 878–887.

He, H., and Garcia, E. A. 2008. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*.

He, H.; Bai, Y.; Garcia, E. A.; and Li, S. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *IJCNN*, 1322–1328.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Kantorovitch, L. 1958. On the translocation of masses. *Management Science* 5(1):1–4.

Lemaître, G.; Nogueira, F.; and Aridas, C. K. 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research* 18(1):559–563.

Lin, C.-T.; Hsieh, T.-Y.; Liu, Y.-T.; Lin, Y.-Y.; Fang, C.-N.; Wang, Y.-K.; Yen, G.; Pal, N. R.; and Chuang, C.-H. 2018. Minority oversampling in kernel adaptive subspaces for class imbalanced datasets. *IEEE Transactions on Knowledge and Data Engineering*.

Liu, M.; Xu, C.; Luo, Y.; Xu, C.; Wen, Y.; and Tao, D. 2017. Cost-sensitive feature selection via f-measure optimization reduction. In *AAAI*, 2252–2258.

Monge, G. 1781. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.

Peng, Y. 2015. Adaptive sampling with optimal cost for class-imbalance learning. In *AAAI*, volume 15, 2921–2927.

Pérez-Ortiz, M.; Gutiérrez, P. A.; Tino, P.; and Hervás-Martínez, C. 2016. Oversampling the minority class in the feature space. *IEEE Transactions on Neural Networks and Learning Systems* 27(9):1947–1961.

Peyré, G., and Cuturi, M. 2017. Computational optimal transport.

Peyré, G.; Cuturi, M.; and Solomon, J. 2016. Gromov-wasserstein averaging of kernel and distance matrices. In *ICML*, 2664–2672.

Sen, A.; Islam, M. M.; Murase, K.; and Yao, X. 2016. Binarization with boosting and oversampling for multiclass classification. *IEEE Transactions on Cybernetics* 46(5):1078–1091.

Sinkhorn, R. 1967. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly* 74(4):402–405.

Villani, C. 2008. *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

Yan, Y.; Li, W.; Wu, H.; Min, H.; Tan, M.; and Wu, Q. 2018. Semi-supervised optimal transport for heterogeneous domain adaptation. In *IJCAI*, 737–753.

Zhang, Y.; Zhao, P.; Cao, J.; Ma, W.; Huang, J.; Wu, Q.; and Tan, M. 2018. Online adaptive asymmetric active learning for budgeted imbalanced data. In *SIGKDD*, 2768–2777.