LETTER **Communicated by Terrence Sejnowski**

Classification from Triplet Comparison Data

Zhenghang Cui
cui@ms.k.u-tokyo.ac.jp
Nontawat Charoenphakdee
nontawat@ms.k.u-tokyo.ac.jp
Issei Sato
sato@k.u-tokyo.ac.jp
The University of Tokyo, Tokyo 113-0033, Japan, and RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan

Masashi Sugiyama
sugi@k.u-tokyo.ac.jp
RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan and The University of Tokyo, Tokyo 113-0033, Japan

Learning from triplet comparison data has been extensively studied in the context of metric learning, where we want to learn a distance metric between two instances, and ordinal embedding, where we want to learn an embedding in a Euclidean space of the given instances that preserve the comparison order as much as possible. Unlike fully labeled data, triplet comparison data can be collected in a more accurate and human-friendly way. Although learning from triplet comparison data has been considered in many applications, an important fundamental question of whether we can learn a classifier only from triplet comparison data without all the labels has remained unanswered. In this letter, we give a positive answer to this important question by proposing an unbiased estimator for the classification risk under the empirical risk minimization framework. Since the proposed method is based on the empirical risk minimization framework, it inherently has the advantage that any surrogate loss function and any model, including neural networks, can be easily applied. Furthermore, we theoretically establish an estimation error bound for the proposed empirical risk minimizer. Finally, we provide experimental results to show that our method empirically works well and outperforms various baseline methods.

1 Introduction

Recently, learning from comparison-feedback data has received increasing attention (Heim, 2016; Kleindessner, 2017). It is usually argued that humans perform better in the task of evaluating which instances are similar, rather

Neural Computation **32, 659–681** (2020)
https://doi.org/10.1162/neco_a_01262

than identifying each individual instance (Stewart, Brown, & Chater, 2005). It is also argued that humans can achieve much better and more reliable performance on assessing the similarity on a relative scale ("Instance A is more similar to instance B than to instance C") rather than on an absolute scale ("The similarity score between A and B is 0.9, while the one between A and C is 0.4") (Kleindessner, 2017). Collecting data in this manner has the advantage of avoiding the problem caused by individuals' different assessment scales. However, the collected absolute similarity scores may only provide information on a comparison level in some applications, such as sensor localization (Liu, Wu, & He, 2004). It was shown that keeping only the relative comparison information can help an algorithm be resilient against measurement errors and achieve high accuracy (Xiao, Li, & Luo, 2006).

In this letter, we focus on the problem of learning from triplet comparison data, a common form of comparison-feedback data. A triplet comparison (x_a, x_b, x_c) contains the information that instance x_a is more similar to x_b than to x_c. As one example, search engine query logs can readily provide feedback in the form of triplet comparisons (Schultz & Joachims, 2004). Given a list of website links $\{A, B, C\}$ for a query, if links A and B are clicked and the link C is not clicked, we can formulate a triplet comparison as (A, B, C). We can also collect unlabeled data sets first and collect triplet comparison afterward, such as the instrument data set (Mojsilovic & Ukkonen, 2019) and the car data set (Kleindessner, 2017). In these cases, data are collected in a totally unlabeled way.

Learning from triplet comparison data was initially studied in the context of metric learning (Schultz & Joachims, 2004), in which a consistent distance metric between two instances is assumed to be learned from data. The well-known triplet loss for face recognition was proposed in this line of research (Schroff, Kalenichenko, & Philbin, 2015; Yu, Liu, Gong, Ding, & Tao, 2018). When this loss function is used, an inductive mapping function can be efficiently learned from triplet comparison image data. At the same time, the problem of ordinal embedding has also been extensively studied (Agarwal et al., 2007; Van Der Maaten & Weinberger, 2012). It aims to learn an embedding of the given instances to the Euclidean space that preserves the order given by the data. Algorithms for large-scale ordinal embedding have been developed (Anderton & Aslam, 2019). In addition, many other problem settings have been considered for the situation of using only triplet comparison data, such as nearest-neighbor search (Haghiri, Ghoshdastidar, & von Luxburg, 2017), kernel function construction (Kleindessner & von Luxburg, 2017a) and outlier identification (Kleindessner & Von Luxburg, 2017b).

However, learning a binary classifier from triplet comparison data remained untouched until recently. A random forest construction algorithm (Haghiri, Garreau, & Luxburg, 2018) was proposed for both classification and regression. However, it first requires a labeled data set and needs

to actively access a triplet comparison oracle many times. For passively collected triplet comparison data, a boosting-based algorithm (Perrot & von Luxburg, 2018) was recently proposed without accessing a triplet comparison oracle. However, a set of labeled data is still indispensable to initiating the training process. To the best of our knowledge, this letter is the first to tackle the problem of learning a classifier only from passively obtained triplet comparison data without accessing either a labeled data set or an oracle.

We show that we can learn a binary classifier from only passively obtained triplet comparison data. We achieve this goal by developing a novel method for learning a binary classifier in this setting with theoretical justification. We use the direct risk minimization framework given for the classification problem. We then show that the classification risk can be empirically estimated in an unbiased way given only triplet comparison data. Theoretically, we establish an estimation error bound for the proposed empirical risk minimizer, showing that learning from triplet comparison data is consistent. Our method also returns an inductive model, which is different from clustering and ordinal embedding and can be applied to unseen test data points. The test data would consist of single instances instead of triplet comparisons since our primitive goal is to perform a binary classification task on unseen data points.

In summary, for the problem of classification using only triplet comparison data, our contributions in this letter are three-fold:

- We propose an empirical risk minimization method for binary classification using only passively obtained triplet comparison data, which gives us an inductive classifier.
- We theoretically establish an estimation error bound for our method, showing that the learning is consistent.
- We experimentally demonstrate the practical usefulness of our method.

2 Related Work

Our problem setting of learning a binary classifier from passively obtained triplet comparison data can be considered a type of a weakly supervised classification problem, where we do not have access to ground-truth labels (Zhou, 2017).

An approach based on constructing an unbiased risk estimator of the true classification risk from weakly supervised data has been explored in many problem settings; for example, positive-unlabeled classification (du Plessis, Niu, & Sugiyama, 2014; Niu, du Plessis, Sakai, Ma, & Sugiyama, 2016) and similarity-unlabeled classification (Bao, Niu, & Sugiyama, 2018) can be handled by the framework of learning from two sets of unlabeled data (Lu, Niu,

Menon, & Sugiyama, 2019). Nevertheless, our problem setting is not a special case addressed by Lu et al. (2019) since we have only one set of triplet comparison data. We later show that we can formulate three different distributions, which is significantly different from the framework that Lu et al. (2019) used and can be considered as a case of learning from three sets of unlabeled data.

Moreover, our problem setting is also different from similarity-dissimilarity-unlabeled classification (Shimada, Bao, Sato, & Sugiyama, 2019) in the sense that we have no access to unlabeled data and similarity and dissimilarity pairs, only triplet comparison information. Furthermore, it is important to note that our problem setting is also different from preference learning (Fürnkranz & Hüllermeier, 2010) since we do not want to learn a ranking function but construct a binary classifier. Although we can first learn a ranking function and then decide a proper threshold to construct a binary classifier (Narasimhan & Agarwal, 2013), it is not straightforward to choose a proper threshold. Therefore, instead of this two-stage method, we focus on a method that can directly learn a binary classifier from triplet comparison data.

3 Learning a Classifier from Triplet Comparison Data

In this section, we first review the fully supervised classification setting. Then we introduce the problem setting and assumption for the data generation process of triplet comparison data. Finally, we describe the proposed method for training a binary classifier from only passively obtained triplet comparison data.

3.1 Preliminary. We first briefly introduce the traditional binary classification problem. We denote $\mathcal{X} \subset \mathbb{R}^d$ as a d-dimensional sample space and $\mathcal{Y} = \{+1, -1\}$ as a binary label space. In the fully supervised setting, we usually assume the labeled data $(x, y) \in \mathcal{X} \times \mathcal{Y}$ are drawn from the joint probability distribution with density $p(x, y)$ (Vapnik, 1995). The goal is to obtain a classifier $f : \mathcal{X} \to \mathbb{R}$ that minimizes the classification risk

$$R(f) = \mathop{\mathbb{E}}_{(x,y)\sim p(x,y)} [\ell(f(x), y)], \tag{3.1}$$

where the expectation is over the joint density $p(x, y)$ and $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$ is a loss function that measures how well the classifier estimates the true class label.

In the fully supervised classification setting, we are given both positive and negative training data collectively drawn from the joint density $p(x, y)$. However, in our case, we still want to train a binary classifier that minimizes the classification risk, although we do not have fully labeled data.

3.2 Generation Process of Triplet Comparison Data. We formulate the underlying generation process of triplet comparison data in order to perform empirical risk minimization (ERM). Three samples in a triplet (x_a, x_b, x_c) are first generated independently and then shown to a user. The user can mark the triplet to be proper or not. Denoting the similarity between two samples x_a and x_b as σ_{ab}, the larger σ_{ab} is, the more similar the two samples are. Then a proper triplet means $\sigma_{ab} > \sigma_{ac}$. Specifically, it means that three labels (y_a, y_b, y_c) in a triplet appear to be one of the following cases:

$$\mathcal{Y}_1 \triangleq \{(+1, +1, -1), (-1, -1, +1), (+1, +1, +1), \\ (-1, -1, -1), (+1, -1, -1), (-1, +1, +1)\}.$$

Otherwise, it means the first sample is more similar to the third sample than to the second sample; thus, the user chooses to mark the triplet as not proper. Similarly, it means (y_a, y_b, y_c) appears to be one of the following cases:

$$\mathcal{Y}_2 \triangleq \{(+1, -1, +1), (-1, +1, -1)\}.$$

First, three data samples are generated independently from the underlying joint density $p(x, y)$; then $\mathcal{D} = \{(x_a, x_b, x_c)\}$ are collected without knowing the underlying true labels (y_a, y_b, y_c). However, we can collect information from user feedback about which case a triplet belongs to. Notice that in this problem setting, we assume the user always gives rational feedback. This means the user never recognizes samples with different labels to be more similar to each other. After receiving feedback from users, we can actually obtain two distinct data sets. The data the user chooses to keep the order are denoted as

$$\mathcal{D}_1 \triangleq \{(x_a, x_b, x_c) | (y_a, y_b, y_c) \in \mathcal{Y}_1\}.$$

Similarly, the data the user chooses to flip the order are denoted as

$$\mathcal{D}_2 \triangleq \{(x_a, x_b, x_c) | (y_a, y_b, y_c) \in \mathcal{Y}_2\}.$$

Note that the ratio of $n_1 \triangleq |\mathcal{D}_1|$ to $n_2 \triangleq |\mathcal{D}_2|$ is fixed because we assume the three samples in a triplet are generated independently from $p(x, y)$; thus, the ratio $\frac{n_1}{n_2}$ is only dependent on the underlying class prior probabilities, which are fixed, unknown values.

The two data sets can be considered to be generated from two underlying distributions, as indicated by the following lemma.

Lemma 1. *Corresponding to the data generation process described above, let*

$$p_1(x_a, x_b, x_c) = \frac{p(x_a, x_b, x_c, (y_a, y_b, y_c) \in \mathcal{Y}_1)}{\pi_T},$$
$$p_2(x_a, x_b, x_c) = \pi_+ p_+(x_a)p_-(x_b)p_+(x_c) + \pi_- p_-(x_a)p_+(x_b)p_-(x_c), \quad (3.2)$$

where $\pi_T = 1 - \pi_+\pi_-$, $\pi_+ \triangleq p(y = +1)$ *and* $\pi_- \triangleq p(y = -1)$ *are the class prior probabilities that satisfy* $\pi_+ + \pi_- = 1$ *and* $p_+(x) \triangleq p(x|y = +1)$ *and* $p_-(x) \triangleq p(x|y = -1)$ *are class-conditional probabilities. Then it follows that*

$$\mathcal{D}_1 = \{(x_{1,a}, x_{1,b}, x_{1,c})\}_{i=1}^{n_1} \overset{i.i.d.}{\sim} p_1(x_a, x_b, x_c),$$
$$\mathcal{D}_2 = \{(x_{2,a}, x_{2,b}, x_{2,c})\}_{i=1}^{n_2} \overset{i.i.d.}{\sim} p_2(x_a, x_b, x_c).$$

Detailed derivation is given in appendix A.

We denote the pointwise data collected from $\mathcal{D}_1$ and $\mathcal{D}_2$ by ignoring the triplet comparison relation as $\mathcal{D}_{1,a} \triangleq \{x_{1,a}\}_{i=1}^{n_1}$, $\mathcal{D}_{1,b} \triangleq \{x_{1,b}\}_{i=1}^{n_1}$, $\mathcal{D}_{1,c} \triangleq \{x_{1,c}\}_{i=1}^{n_1}$, $\mathcal{D}_{2,a} \triangleq \{x_{2,a}\}_{i=1}^{n_2}$, $\mathcal{D}_{2,b} \triangleq \{x_{2,b}\}_{i=1}^{n_2}$ and $\mathcal{D}_{2,c} \triangleq \{x_{2,c}\}_{i=1}^{n_2}$, the marginal densities of which can be expressed by the following theorem.

Theorem 1. *Samples in* $\mathcal{D}_{1,a}$, $\mathcal{D}_{1,c}$, $\mathcal{D}_{2,a}$, *and* $\mathcal{D}_{2,c}$ *are independently drawn from*

$$\tilde{p}_1(x) = \pi_+ p_+(x) + \pi_- p_-(x); \quad (3.3)$$

samples in $\mathcal{D}_{1,b}$ *are independently drawn from*

$$\tilde{p}_2(x) = \frac{(\pi_+^3 + 2\pi_+^2\pi_-)p_+(x) + (2\pi_+\pi_-^2 + \pi_-^3)p_-(x)}{\pi_T}; \quad (3.4)$$

and samples in $\mathcal{D}_{2,b}$ *are independently drawn from*

$$\tilde{p}_3(x) = \pi_- p_+(x) + \pi_+ p_-(x). \quad (3.5)$$

The proof is given in appendix B.

Theorem 1 indicates that from triplet comparison data, we can essentially obtain samples that can be drawn independently from three different distributions. We denote the three aggregated data sets as

$$\tilde{\mathcal{D}}_1 = \mathcal{D}_{1,a} \cup \mathcal{D}_{1,c} \cup \mathcal{D}_{2,a} \cup \mathcal{D}_{2,c},$$
$$\tilde{\mathcal{D}}_2 = \mathcal{D}_{1,b}, \quad \tilde{\mathcal{D}}_3 = \mathcal{D}_{2,b}.$$

3.3 Unbiased Risk Estimator for Triplet Comparison Data. We now attempt to express the classification risk,

$$R(f) \triangleq \mathop{\mathbb{E}}_{(x,y)\sim p(x,y)}[\ell(f(x), y)], \tag{3.6}$$

on the basis of the three pointwise densities presented in section 3.2.

The classification risk can be separately expressed as the expectations over $p_+(x)$ and $p_-(x)$. Although we do not have access to data drawn from these two distributions, we can obtain data from three related densities $\tilde{p}_1(x)$, $\tilde{p}_2(x)$, and $\tilde{p}_3(x)$, as indicated in theorem 1. Letting

$$A \triangleq \frac{\pi_+^3 + 2\pi_+^2\pi_-}{\pi_{\mathrm{T}}}, \quad B \triangleq \frac{2\pi_+\pi_-^2 + \pi_-^3}{\pi_{\mathrm{T}}}, \tag{3.7}$$

we can express the relationship between these densities as

$$\begin{bmatrix} \tilde{p}_1(x) \\ \tilde{p}_2(x) \\ \tilde{p}_3(x) \end{bmatrix} = \begin{bmatrix} \pi_+ & \pi_- \\ A & B \\ \pi_- & \pi_+ \end{bmatrix} \begin{bmatrix} p_+(x) \\ p_-(x) \end{bmatrix}. \tag{3.8}$$

Our goal is to solve equation 3.8 so that we can express $p_+(x)$ and $p_-(x)$ in terms of the three densities from which we have independent and identically distributed (i.i.d.) data samples. To this end, we can rewrite the classification risk, which we want to minimize, in terms of $\tilde{p}_1(x)$, $\tilde{p}_2(x)$, and $\tilde{p}_3(x)$. An answer to equation 3.8 is given by the following lemma.

Lemma 2. *We can express $p_+(x)$ and $p_-(x)$ in terms of $\tilde{p}_1(x)$, $\tilde{p}_2(x)$, and $\tilde{p}_3(x)$ as*

$$\begin{aligned} p_+(x) &= \frac{1}{(ac - b^2)}((c\pi_+ - b\pi_-)\tilde{p}_1(x) + (cA - bB)\tilde{p}_2(x) \\ &\quad + (c\pi_- - b\pi_+)\tilde{p}_3(x)), \\ p_-(x) &= \frac{1}{(ac - b^2)}((a\pi_- - b\pi_+)\tilde{p}_1(x) + (aB - bA)\tilde{p}_2(x) \\ &\quad + (a\pi_+ - b\pi_-)\tilde{p}_3(x)), \end{aligned} \tag{3.9}$$

provided $ac - b^2 \neq 0$ where

$$a \triangleq \pi_+^2 + A^2 + \pi_-^2, \quad b \triangleq 2\pi_+\pi_- + AB, \quad c \triangleq \pi_-^2 + B^2 + \pi_+^2.$$

Detailed derivation is given in appendix C.

As a result of lemma 2, we can express the classification risk using only triplet comparison data. Letting $\ell_+(x) \triangleq \ell(f(x), +1)$ and $\ell_-(x) \triangleq \ell(f(x), -1)$, we have the following theorem.

Theorem 2. *The classification risk can be equivalently expressed as*

$$R(f) = \frac{1}{(ac-b^2)}\Big\{ \mathop{\mathbb{E}}_{x\sim\tilde{p}_1(x)}[\pi_{test}(c\pi_+ - b\pi_-)\,\ell_+(x) + (1-\pi_{test})(a\pi_- - b\pi_+)\,\ell_-(x)] + \mathop{\mathbb{E}}_{x\sim\tilde{p}_2(x)}[\pi_{test}(cA - bB)\,\ell_+(x) + (1-\pi_{test})(aB - bA)\,\ell_-(x)] + \mathop{\mathbb{E}}_{x\sim\tilde{p}_3(x)}[\pi_{test}(c\pi_- - b\pi_+)\,\ell_+(x) + (1-\pi_{test})(a\pi_+ - b\pi_-)\,\ell_-(x)]\Big\}, \tag{3.10}$$

where $\pi_{test} \triangleq p_{test}(y=+1)$ *denotes the class prior of the test data set.*

The proof is given in appendix D.

In this letter, we consider the common case in which $\pi_{\text{test}} = \pi_+$, which means the test data set shares the same class prior as the training data set. However, even when $\pi_{\text{test}} \neq \pi_+$, which means the class prior shift (Sugiyama, 2012) occurs, our method can still be used when π_{test} is known.

The process of obtaining the empirical risk minimizer of equation 3.10, $\hat{f} = \arg\min R(f)$, is similar to other ERM-based learning approaches. As long as the risk representation that we want to minimize is continuous and differentiable with respect to the model parameters, such as the linear-in-parameter model or neural networks, we can use powerful stochastic optimization algorithms (Kingma & Ba, 2014).

4 Estimation Error Bound

In this section, we establish an estimation error bound for the proposed unbiased risk estimator. Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ represent a function class specified by a model. First, let $\mathfrak{R}(\mathcal{F})$ be the (expected) Rademacher complexity of $\mathcal{F}$, which is defined as

$$\mathfrak{R}(\mathcal{F}) \triangleq \mathop{\mathbb{E}}_{Z_1,\cdots,Z_n\sim\mu} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{f\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(Z_i) \right], \tag{4.1}$$

where n is a positive integer, $Z_1, \cdots, Z_n$ are i.i.d. random variables drawn from a probability distribution with density μ, and $\boldsymbol{\sigma} = (\sigma_1, \cdots, \sigma_n)$ are Rademacher variables, which are random variables that take the value of $+1$ or -1 with even probabilities.

We assume for any probability density μ, the specified model $\mathcal{F}$ satisfies $\mathfrak{R}(\mathcal{F}) \leq \frac{C_{\mathcal{F}}}{\sqrt{n}}$ for some constant $C_{\mathcal{F}} > 0$. Also, let $f^* \triangleq \arg\min_{f\in\mathcal{F}} R(f)$ be the true risk minimizer and $\hat{f} \triangleq \arg\min_{f\in\mathcal{F}} \hat{R}_{T,\ell}(f)$ the empirical risk minimizer.

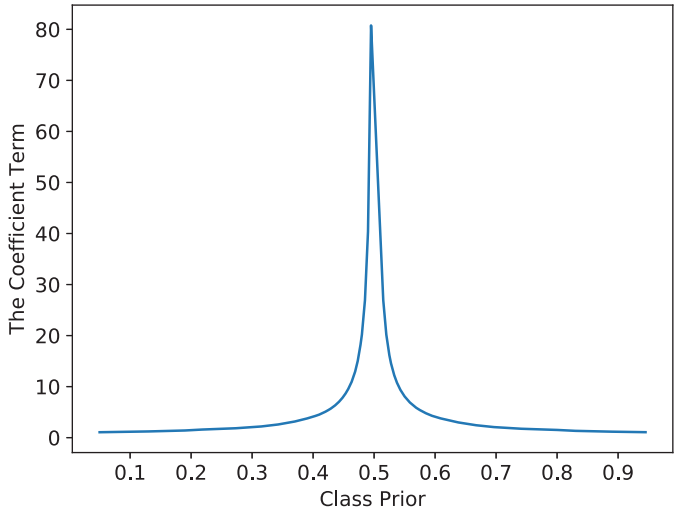

Figure 1: Behavior of the coefficient term.

Theorem 3. *Assume the loss function ℓ is ρ-Lipschitz with respect to the first argument ($0 < \rho < \infty$), and all functions in the model class $\mathcal{F}$ are bounded—that is, there exists a constant C_b such that $||f||_\infty \leq C_b$ for any $f \in \mathcal{F}$. Let $C_\ell \triangleq \sup_{t\in\{\pm 1\}} \ell(C_b, t)$. Then for any $\delta > 0$, with probability at least $1-\delta$,*

$$R(\hat{f}) - R(f^*) \leq \left(\frac{2\rho C_{\mathcal{F}}}{\sqrt{n}} + \sqrt{\frac{C_\ell^2 \log \frac{2}{\delta}}{2n}} \right) \cdot \frac{C_R}{|ac - b^2|}, \tag{4.2}$$

where

$$\begin{aligned} C_R = {} & |\pi_{test}(c\pi_+ - b\pi_-)| + |(1-\pi_{test})(a\pi_- - b\pi_+)| + |\pi_{test}(cA - bB)| \\ & + |(1-\pi_{test})(aB - bA)| + |\pi_{test}(c\pi_- - b\pi_+)| \\ & + |(1-\pi_{test})(a\pi_+ - b\pi_-)|. \end{aligned} \tag{4.3}$$

The proof is given in appendix E.

Since n appears in the denominator, it is obvious that when the class prior is fixed, the bound will get tighter as the triplet comparison data increase. However, it is not clear how the bound will behave when we fix the amount of triplet comparison data and change the class prior. Thus in Figure 1, we show the behavior of the coefficient term $\frac{C_R}{|ac-b^2|}$ with respect to the same class prior of both training and test data sets. From the illustration, we can

capture the rough trend that the bound gets tighter when the class prior becomes further from 0.5. We will investigate this behavior in experiments.

5 On Class Prior

In the previous sections, the class prior π_+ is assumed known. For this simple case, we can directly use the proposed algorithm to separate test data as well as identify correct classes. However, it may not be true for many real-world applications. Two situations can be considered. For the worst case, no information about the class prior is given. Although we can still estimate a result for the class prior from data and obtain a classifier that is able to separate data for different classes, we cannot identify the correct class without the information about which class has a higher class prior. A better situation is that we have the information about which class has a higher class prior. By setting this class as the positive one, we can successfully train a classifier to identify the correct class. Thus, we assume that the positive class has a higher class prior, which means $\pi_+ > \frac{1}{2}$.

5.1 Class Prior Estimation from Triplet Comparison Data. Noticing $\pi_{\mathrm{T}} = 1 - \pi_+ + \pi_+^2$, we can obtain $\pi_+^2 - \pi_+ + (1 - \pi_{\mathrm{T}}) = 0$. By assuming $\pi_+ > \pi_-$, we have

$$\pi_+ = \frac{1 + \sqrt{1 - 4(1 - \pi_{\mathrm{T}})}}{2}. \tag{5.1}$$

Since we can unbiasedly estimate π_{T} by $\frac{n_1}{n_1+n_2}$, the class prior π_+ can thus be estimated once the triplet comparison data set is given.

6 Experiments

In this section, we conducted experiments using real-world data sets to evaluate and investigate the performance of the proposed method for triplet classification.

6.1 Baseline Methods

6.1.1 KMEANS. As a simple baseline, we used k-means clustering (Macqueen, 1967) with $k = 2$ on all the data instances of triplets while ignoring all the relation information.

6.1.2 ITML. Information-theoretic metric learning (Davis, Kulis, Jain, Sra, & Dhillon, 2007) is a metric learning method that requires pairwise the relationship between data instances. From a triplet (x_a, x_b, x_c), we constructed pairwise constraints as (x_a, x_b) being similar and (x_a, x_c) being dissimilar. Using the metric returned by the algorithm, we conducted k-means

clustering on test data. We used the identity matrix for prior knowledge and fix the slack variable as $\gamma = 1$.

6.1.3 TL. Triplet loss (Schroff et al., 2015) is a loss function proposed in the context of deep metric learning, which can learn a metric directly from triplet comparison data. Using the metric returned by the algorithm, we conducted k-means clustering on test data.

6.1.4 SERAPH. Semisupervised metric learning paradigm with hypersparsity (Niu, Dai, Yamada, & Sugiyama, 2014) is a metric learning method based on entropy regularization. We formulated a pairwise relationship in the same manner as with ITML. Using the metric returned by ITML, we conducted k-means clustering on test data.

6.1.5 SU. SU learning (Bao et al., 2018) is a method for learning a binary classifier from similarity and unlabeled data. We used the same method for estimating the class prior and considered the less similar sample in a triplet as unlabeled data.

6.2 Data Sets.

6.2.1 UCI Data Sets. We used six data sets from the UCI Machine Learning Repository (Asuncion & Newman, 2007). They are binary classification data sets, and we use the given labels for further triplet comparison data generation.

6.2.2 Image Data Sets. We used three image data sets.

The MNIST (LeCun, Bottou, Bengio, & Haffner, 1998) data set consists of 70,000 examples associated with a label from 10 digits. Each data instance is a 28×28 gray-scale image; thus, the input dimension is 784. To form a binary classification problem, we treat even numbers as the positive class and odd numbers as the negative class. The data were standardized to have zero mean and unit variance.

The Fashion MNIST (Xiao, Rasul, & Vollgraf, 2017) data set consists of 70,000 examples associated with a label from 10 fashion item classes. Each data instance is a 28×28 gray-scale image; thus, the input dimension is 784. To form a binary classification problem, we treat five classes—T-shirt/top, Pullover, Dress, Coat, and Shirt—as positive class since they all represent upper—body clothing. The data were standardized to have zero mean and unit variance.

The CIFAR-10 (Krizhevsky & Hinton, 2009) data set consists of 60,000 examples associated with a label from 10 classes. Each image is given in a $32 \times 32 \times 3$ format; thus, the input dimension is 3,072. To form a binary classification problem, we treated four classes—airplane, automobile, ship, and truck—as positive classes since they all represent artificial objects.

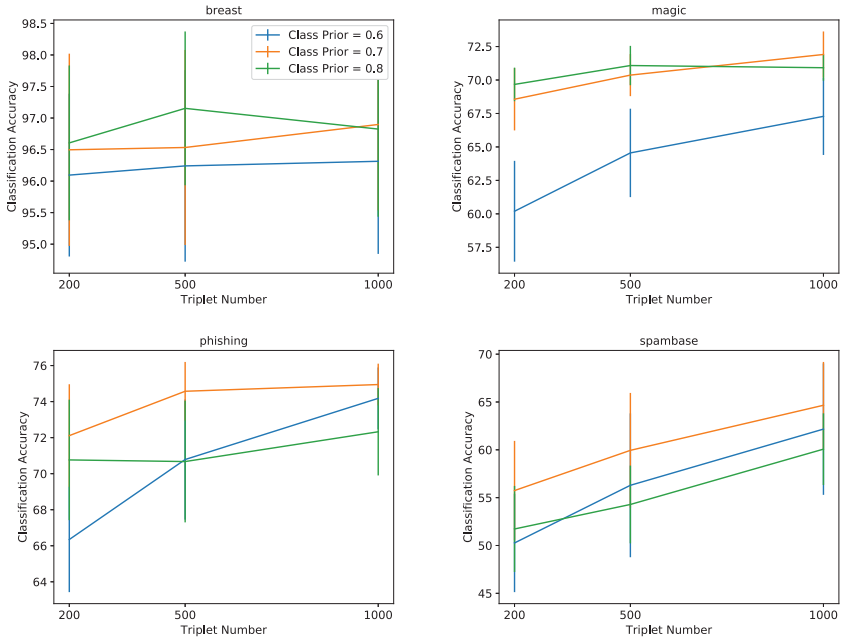

Figure 2: Average classification error and standard error over 20 trials.

Although these data sets have labels, using the triplet comparison data composed of labeled data fulfills the purpose of experiments, which is to assess whether the proposed method can work properly. As mentioned in section 1, the proposed method can be applied to situations where we do not have access to the labels.

6.3 Proposed Method. For the proposed method, we used a fully connected neural network with only one hidden layer of width 100 and rectified linear units (ReLUs) (Nair & Hinton, 2010) for all the data sets except for CIFAR-10. The width of the hidden layer was set to be 100 throughout all experiments. Adam (Kingma & Ba, 2014) was used for optimization. The neural network architecture used for CIFAR-10 is specified in appendix F. Two surrogate losses were used as indicated in Tables 1, 2, and 3.

6.4 Results. The proposed method estimates the unknown class prior first. For baseline methods, performances are measured by the clustering accuracy $1 - \min(r, 1 - r)$ where r is the error rate. The results of different triplet numbers are listed in Tables 1, 2, and 3. The best and equivalent methods are shown in bold on the one-sided t-test with a significance level of 5%. Also, as shown in Figure 2, the performance of the proposed method with

Table 1: Experimental Results with Class Prior as 0.7 and 1000 Training Triplets.

	Proposed Methods		Baselines				
Data Set	Squared	Double Hinge	KMEANS	ITML	TL	SERAPH	SU
Adult	65.54 (0.41)	64.19 (0.61)	71.94 (0.10)	71.04 (1.00)	61.48 (1.36)	71.04 (1.00)	**75.88 (0.50)**
Breast	**97.41 (0.28)**	**96.90 (0.31)**	96.20 (0.34)	95.84 (0.29)	93.87 (0.78)	96.72 (0.23)	65.26 (0.76)
Diabetes	**70.71 (0.84)**	64.87 (0.74)	66.69 (0.70)	65.91 (0.69)	64.38 (1.60)	67.44 (0.78)	34.42 (0.73)
Magic	61.75 (1.00)	**71.91 (0.39)**	65.08 (0.17)	64.79 (0.17)	65.42 (0.22)	64.96 (0.19)	34.77 (0.19)
Phishing	**76.58 (0.30)**	74.95 (0.27)	63.43 (0.50)	63.75 (0.23)	57.85 (0.92)	63.42 (0.53)	34.17 (0.22)
Spambase	**62.08 (1.87)**	**64.66 (1.04)**	**63.59 (0.24)**	**63.24 (0.31)**	59.59 (1.57)	**63.28 (0.34)**	60.27 (0.30)
MNIST	79.86 (0.35)	**80.78 (0.34)**	65.24 (0.25)	0.00 (0.00)	58.26 (1.24)	0.00 (0.00)	50.80 (0.03)
Fashion	89.73 (0.33)	**91.62 (0.33)**	74.90 (1.00)	0.00 (0.00)	76.83 (1.31)	0.00 (0.00)	49.85 (0.08)
CIFAR10	**76.39 (1.57)**	66.28 (2.51)	64.17 (0.01)	0.00 (0.00)	60.17 (1.26)	0.00 (0.00)	59.50 (0.50)

Table 2: Experimental Results with Class Prior as 0.7 and 500 Training Triplets.

	Proposed Methods		Baselines				
Data Set	Squared	Double Hinge	KMEANS	ITML	TL	SERAPH	SU
Adult	62.72 (0.57)	59.74 (1.44)	71.44 (0.60)	71.79 (0.20)	58.53 (1.17)	70.54 (1.09)	**76.30 (0.04)**
Breast	**96.90 (0.44)**	**96.53 (0.35)**	**96.28 (0.29)**	**96.79 (0.24)**	89.67 (1.97)	**96.68 (0.27)**	64.12 (0.91)
Diabetes	**69.64 (0.68)**	67.08 (0.91)	66.27 (0.65)	64.87 (0.66)	63.15 (1.56)	67.44 (0.68)	33.90 (0.67)
Magic	63.86 (1.44)	**70.37 (0.36)**	64.86 (0.15)	65.03 (0.13)	66.36 (0.30)	64.94 (0.14)	34.83 (0.15)
Phishing	**75.52 (0.31)**	74.57 (0.37)	63.08 (0.47)	63.31 (0.41)	56.37 (1.18)	62.73 (0.76)	33.89 (0.20)
Spambase	61.18 (1.11)	59.95 (1.38)	**63.55 (0.32)**	**64.17 (0.31)**	59.35 (1.48)	**63.53 (0.35)**	58.96 (0.44)
MNIST	**74.23 (0.32)**	**75.19 (0.50)**	64.74 (0.55)	0.00 (0.00)	56.07 (0.87)	0.00 (0.00)	50.87 (0.26)
Fashion	83.83 (0.55)	**87.86 (0.66)**	75.40 (0.34)	0.00 (0.00)	76.66 (1.39)	0.00 (0.00)	49.88 (0.08)
CIFAR10	**66.28 (1.77)**	**62.63 (2.53)**	**64.16 (0.01)**	0.00 (0.00)	61.26 (1.13)	0.00 (0.00)	59.05 (0.65)

Table 3: Experimental Results with Class Prior as 0.7 and 200 Training Triplets.

	Proposed Methods		Baselines				
Data Set	Squared	Double Hinge	KMEANS	ITML	TL	SERAPH	SU
Adult	58.12 (0.90)	55.10 (1.00)	70.54 (1.50)	70.04 (1.17)	58.28 (0.94)	68.54 (1.67)	**75.27 (0.51)**
Breast	**96.68 (0.32)**	**96.50 (0.35)**	**95.91 (0.34)**	**96.24 (0.24)**	94.27 (0.68)	**96.64 (0.28)**	66.20 (0.80)
Diabetes	**69.25 (0.98)**	65.36 (0.89)	64.97 (0.87)	**67.27 (0.72)**	63.47 (1.22)	**67.11 (0.82)**	35.23 (0.94)
Magic	60.54 (1.88)	**68.56 (0.53)**	64.88 (0.13)	65.15 (0.14)	66.31 (0.42)	64.97 (0.15)	34.60 (0.34)
Phishing	**72.22 (0.62)**	**72.11 (0.65)**	63.70 (0.26)	63.71 (0.21)	57.02 (1.41)	63.17 (0.77)	34.03 (0.32)
Spambase	57.69 (1.68)	55.74 (1.19)	**63.78 (0.34)**	**63.04 (0.35)**	60.78 (1.63)	**63.74 (0.25)**	58.92 (0.43)
MNIST	67.14 (0.67)	**70.96 (0.53)**	64.49 (1.00)	0.00 (0.00)	57.88 (1.43)	0.00 (0.00)	50.10 (0.62)
Fashion	76.67 (0.40)	**83.74 (0.55)**	74.90 (1.00)	0.00 (0.00)	73.24 (1.80)	0.00 (0.00)	47.97 (0.76)
CIFAR10	**63.14 (1.68)**	58.83 (2.16)	**64.16 (0.01)**	0.00 (0.00)	61.23 (1.18)	0.00 (0.00)	58.65 (0.66)

respect to the class prior and the size of training data set followed the prediction by the theory in most of the cases.

7 Conclusion

In this letter, we proposed a novel method for learning a classifier from only passively obtained triplet comparison data. We established an estimation error bound for the proposed method and confirmed that the estimation error decreases as the amount of triplet comparison data increases. We also empirically confirmed that the performance of the proposed method surpassed multiple baseline methods on various data sets. For future work, it would be interesting to investigate alternative methods that can handle a multiclass case.

Appendix A: Proof of Lemma 1

From the data generation process, we can consider the generation distribution for data of $\mathcal{D}_1$ as

$$\begin{aligned} p_1(x_a, x_b, x_c) &= p(x_a, x_b, x_c | (y_a, y_b, y_c) \in \mathcal{Y}_1) \\ &= \frac{p(x_a, x_b, x_c, (y_a, y_b, y_c) \in \mathcal{Y}_1)}{p((y_a, y_b, y_c) \in \mathcal{Y}_1)} \\ &= \frac{p(x_a, x_b, x_c, (y_a, y_b, y_c) \in \mathcal{Y}_1)}{\pi_+^3 + 2\pi_+^2\pi_- + 2\pi_+\pi_-^2 + \pi_-^3}. \end{aligned} \tag{A.1}$$

Note that the denominator in equation A.1 can be rewritten as

$$\begin{aligned} \pi_{\mathrm{T}} &\triangleq \pi_+^3 + 2\pi_+^2\pi_- + 2\pi_+\pi_-^2 + \pi_-^3 \\ &= (\pi_+^3 + \pi_-^3) + 2(\pi_+^2\pi_- + \pi_+\pi_-^2) \\ &= \pi_+^2 + \pi_+\pi_- + \pi_-^2 \\ &= 1 - \pi_+\pi_-. \end{aligned} \tag{A.2}$$

Then we have

$$p_1(x_a, x_b, x_c) = \frac{p(x_a, x_b, x_c, (y_a, y_b, y_c) \in \mathcal{Y}_1)}{\pi_{\mathrm{T}}}. \tag{A.3}$$

Moreover, the distribution $p(x_a, x_b, x_c, (y_a, y_b, y_c) \in \mathcal{Y}_1)$ at the numerator of equation A.3 can be explicitly expressed as

$$\begin{aligned} &p(x_a, x_b, x_c, (y_a, y_b, y_c) \in \mathcal{Y}_1) \\ &= \pi_+^3 p_+(x_a) p_+(x_b) p_+(x_c) + \pi_+^2\pi_- p_+(x_a) p_+(x_b) p_-(x_c) \end{aligned}$$

$$+ \pi_+\pi_-^2 p_+(x_a)p_-(x_b)p_-(x_c) + \pi_+^2\pi_- p_-(x_a)p_+(x_b)p_+(x_c)$$
$$+ \pi_+\pi_-^2 p_-(x_a)p_-(x_b)p_+(x_c) + \pi_-^3 p_-(x_a)p_-(x_b)p_-(x_c), \quad \text{(A.4)}$$

from the assumption that three instances in each triplet comparison are generated independently.

Similarly, the underlying density for data of $\mathcal{D}_2$ can be expressed as

$$\begin{aligned} p_2(x_a, x_b, x_c) &= p(x_a, x_b, x_c | (y_a, y_b, y_c) \in \mathcal{Y}_2) \\ &= \frac{p(x_a, x_b, x_c, (y_a, y_b, y_c) \in \mathcal{Y}_2)}{p((y_a, y_b, y_c) \in \mathcal{Y}_2)} \\ &= \frac{\pi_+^2\pi_- p_+(x_a)p_-(x_b)p_+(x_c) + \pi_+\pi_-^2 p_-(x_a)p_+(x_b)p_-(x_c)}{\pi_+^2\pi_- + \pi_+\pi_-^2} \\ &= \pi_+ p_+(x_a)p_-(x_b)p_+(x_c) + \pi_- p_-(x_a)p_+(x_b)p_-(x_c). \quad \text{(A.5)} \end{aligned}$$

□

Appendix B: Proof of Theorem 1

For simplicity, we give the proof of $\mathcal{D}_{2,a}$; the other five cases follow the similar proof. Notice

$$\mathcal{D}_2 \underset{i.i.d.}{\sim} p_2(x_a, x_b, x_c) = \pi_+ p_+(x_a)p_-(x_b)p_+(x_c) + \pi_- p_-(x_a)p_+(x_b)p_-(x_c). \quad \text{(B.1)}$$

In order to decompose the triplet comparison data distribution into pointwise distribution, we marginalize $p_2(x_a, x_b, x_c)$ with respect to x_b and x_c:

$$\begin{aligned} &\int p_2(x_a, x_b, x_c)dx_b dx_c \\ &= \pi_+ p_+(x_a) \int p_-(x_b)dx_b \int p_+(x_c)dx_c + \pi_- p_-(x_a) \int p_+(x_b)dx_b \int p_-(x_c)dx_c \\ &= \pi_+ p_+(x_a) \int \frac{p(x_b, y=-1)}{p(y=-1)}dx_b \int \frac{p(x_c, y=+1)}{p(y=+1)}dx_c \\ &\quad + \pi_- p_-(x_a) \int \frac{p(x_b, y=+1)}{p(y=+1)}dx_b \int \frac{p(x_c, y=-1)}{p(y=-1)}dx_c \\ &= \pi_+ p_+(x_a) + \pi_- p_-(x_a) \\ &= \tilde{p}_1(x_a). \quad \text{(B.2)} \end{aligned}$$

□

Appendix C: Proof of Lemma 2

Notice that the equation has an infinite number of solutions. Letting

$$T \triangleq \begin{bmatrix} \pi_+ & \pi_- \\ A & B \\ \pi_- & \pi_+ \end{bmatrix}, \tag{C.1}$$

we resort to finding the Moore-Penrose pseudo-inverse (Moore, 1920; Penrose, 1955), which provides the minimum Euclidean norm solution to the above system of linear equations.

Let T^* denote the conjugate transpose. We have

$$T^*T = \begin{bmatrix} \pi_+^2 + A^2 + \pi_-^2 & 2\pi_+\pi_- + AB \\ 2\pi_+\pi_- + AB & \pi_-^2 + B^2 + \pi_+^2 \end{bmatrix} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}. \tag{C.2}$$

In the next step, we need to take the inverse of the above 2×2 matrix. To achieve a proper inverse matrix, we need to introduce another assumption that $\pi_+ \neq \frac{1}{2}$, which guarantees $ac - b^2 \neq 0$. Then

$$(T^*T)^{-1} = \frac{1}{(ac - b^2)} \begin{bmatrix} c & -b \\ -b & a \end{bmatrix}. \tag{C.3}$$

Finally, the Moore-Penrose pseudo-inverse is given by

$$(T^*T)^{-1}T^* = \frac{1}{(ac - b^2)} \begin{bmatrix} c\pi_+ - b\pi_- & cA - bB & c\pi_- - b\pi_+ \\ -b\pi_+ + a\pi_- & -bA + aB & -b\pi_- + a\pi_+ \end{bmatrix}. \tag{C.4}$$

Thus, we can express $p_+(x)$ and $p_-(x)$ in terms of $\tilde{p}_1(x)$, $\tilde{p}_2(x)$, and $\tilde{p}_3(x)$ as

$$\begin{aligned} p_+(x) &= \frac{1}{(ac - b^2)} \left((c\pi_+ - b\pi_-)\tilde{p}_1(x) + (cA - bB)\tilde{p}_2(x) + (c\pi_- - b\pi_+)\tilde{p}_3(x)\right), \\ p_-(x) &= \frac{1}{(ac - b^2)} \left((a\pi_- - b\pi_+)\tilde{p}_1(x) + (aB - bA)\tilde{p}_2(x) + (a\pi_+ - b\pi_-)\tilde{p}_3(x)\right). \end{aligned} \tag{C.5}$$

□

Appendix D: Proof of Theorem 2

Using equation 3.9, we can rewrite the classification risk as

$$\begin{aligned}
R_\ell(f) &= \mathop{\mathbb{E}}_{p(x,y)}[\ell(f(x),y)] \\
&= \pi_{\text{test}} \mathop{\mathbb{E}}_{p_+(x)}[\ell_+(x)] + (1-\pi_{\text{test}}) \mathop{\mathbb{E}}_{p_-(x)}[\ell_-(x)] \\
&= \frac{\pi_{\text{test}}}{(ac-b^2)}\{(c\pi_+ - b\pi_-) \mathop{\mathbb{E}}_{\tilde{p}_1(x)}[\ell_+(x)] + (cA - bB) \mathop{\mathbb{E}}_{\tilde{p}_2(x)}[\ell_+(x)] \\
&\quad + (c\pi_- - b\pi_+) \mathop{\mathbb{E}}_{\tilde{p}_3(x)}[\ell_+(x)]\} + \frac{1-\pi_{\text{test}}}{(ac-b^2)}\{(a\pi_- - b\pi_+) \mathop{\mathbb{E}}_{\tilde{p}_1(x)}[\ell_-(x)] \\
&\quad + (aB - bA) \mathop{\mathbb{E}}_{\tilde{p}_2(x)}[\ell_-(x)] + (a\pi_+ - b\pi_-) \mathop{\mathbb{E}}_{\tilde{p}_3(x)}[\ell_-(x)]\}, \qquad \text{(D.1)}
\end{aligned}$$

which can be then simplified as equation 3.10. □

Appendix E: Proof of Theorem 3

Letting

$$\begin{aligned}
C_1 &\triangleq \frac{\pi_{\text{test}}}{(c\pi_+ - b\pi_-)(ac-b^2)}, \quad C_2 \triangleq \frac{1-\pi_{\text{test}}}{(a\pi_- - b\pi_+)(ac-b^2)}, \\
C_3 &\triangleq \frac{\pi_{\text{test}}}{(cA-bB)(ac-b^2)}, \quad C_4 \triangleq \frac{(1-\pi_{\text{test}})}{(aB-bA)(ac-b^2)}, \\
C_5 &\triangleq \frac{\pi_{\text{test}}}{(c\pi_- - b\pi_+)(ac-b^2)}, \quad C_6 \triangleq \frac{(1-\pi_{\text{test}})}{(a\pi_+ - b\pi_-)(ac-b^2)},
\end{aligned}$$

and

$$\begin{aligned}
R_a(f) &= \mathop{\mathbb{E}}_{x\sim\tilde{p}_1(x)}[C_1\ell(f(x),+1) + C_2\ell(f(x),-1)], \\
R_b(f) &= \mathop{\mathbb{E}}_{x\sim\tilde{p}_2(x)}[C_3\ell(f(x),+1) + C_4\ell(f(x),-1)], \\
R_c(f) &= \mathop{\mathbb{E}}_{x\sim\tilde{p}_3(x)}[C_5\ell(f(x),+1) + C_6\ell(f(x),-1)], \qquad \text{(E.1)}
\end{aligned}$$

we can simplify the unbiased risk estimator in the form

$$R(f) = R_a(f) + R_b(f) + R_c(f). \qquad \text{(E.2)}$$

Then

$$R(\hat{f}) - R(f^*) \leq 2\sup_{f\in\mathcal{F}} |R_a(f) - \hat{R}_a(f)| + 2\sup_{f\in\mathcal{F}} |R_b(f) - \hat{R}_b(f)| + 2\sup_{f\in\mathcal{F}} |R_c(f) - \hat{R}_c(f)|.$$

For the first term,

$$\begin{aligned}
\sup_{f\in\mathcal{F}} |R_a(f) - \hat{R}_a(f)| &= \sup_{f\in\mathcal{F}} \left| \mathop{\mathbb{E}}_{p_a(x)} [C_1\ell(f(x), +1) + C_2\ell(f(x), -1)] - \frac{1}{n}\sum_{i=1}^{n} \hat{L} \right| \\
&\leq |C_1| \sup_{f\in\mathcal{F}} \left| \mathop{\mathbb{E}}_{p_a(x)} [\ell(f(x), +1)] - \frac{1}{n}\sum_{i=1}^{n} \ell(\widehat{f(x), +1}) \right| \\
&\quad + |C_2| \sup_{f\in\mathcal{F}} \left| \mathop{\mathbb{E}}_{p_a(x)} [\ell(f(x), -1)] - \frac{1}{n}\sum_{i=1}^{n} \ell(\widehat{f(x), -1}) \right| \\
&\leq |C_1| 2\mathcal{R} + |C_1| \sqrt{\frac{C_\ell^2 \log \frac{2}{\delta}}{2n}} + |C_2| 2\mathcal{R} + |C_2| \sqrt{\frac{C_\ell^2 \log \frac{2}{\delta}}{2n}} \\
&= (|C_1| + |C_2|) \left(\frac{2\rho C_{\mathcal{F}}}{\sqrt{n}} + \sqrt{\frac{C_\ell^2 \log \frac{2}{\delta}}{2n}} \right). \qquad \text{(E.3)}
\end{aligned}$$

Combining three terms, theorem 3 is proved. □

Appendix E: CNN Structure for CIFAR10

The following structure is used:

- Convolution (3 in/32 out-channels, kernel size 3) with ReLU
- Convolution (32 in/32 out-channels, kernel size 3) with ReLU
- Max-pooling (kernel size 2, stride 2)
- Repeat twice:
  - Convolution (32 in/32 out-channels, kernel size 3) with ReLU
  - Convolution (32 in/32 out-channels, kernel size 3) with ReLU
  - Max-pooling (kernel size 2, stride 2)
- Fully connected (512 units) with ReLU
- Fully connected (1 unit) □

Acknowledgments

Z.C. was supported by the IST-RA program, the University of Tokyo. N.C. was supported by a MEXT scholarship. I.S. was supported by JST CREST,

grant JPMJCR17A1, Japan. M.S. was supported by the International Research Center for Neurointelligence (WPI-IRCN) at the University of Tokyo Institutes for Advanced Study. We thank Ikko Yamane and Han Bao for fruitful discussions on this work.

References

Agarwal, S., Wills, J., Cayton, L., Lanckriet, G., Kriegman, D., & Belongie, S. (2007). Generalized non-metric multidimensional scaling. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics* (pp. 11–18).

Anderton, J., & Aslam, J. (2019). Scaling up ordinal embedding: A landmark approach. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 282–290).

Asuncion, A., & Newman, D. (2007). UCI machine learning repository. Irvine, CA: University of California, Irvine. http://archive.ics.uci.edu/ml

Bao, H., Niu, G., & Sugiyama, M. (2018). Classification from pairwise similarity and unlabeled data. In *Proceedings of the 35th International Conference on Machine Learning* (pp. 452–461).

Davis, J. V., Kulis, B., Jain, P., Sra, S., & Dhillon, I. S. (2007). Information-theoretic metric learning. In *Proceedings of the 24th International Conference on Machine Learning* (pp. 209–216).

du Plessis, M. C., Niu, G., & Sugiyama, M. (2014). Analysis of learning from positive and unlabeled data. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, *27* (pp. 703–711). Red Hook, NY: Curran.

Fürnkranz, J., & Hüllermeier, E. (2010). *Preference learning*. Berlin: Springer.

Haghiri, S., Garreau, D., & Luxburg, U. (2018). Comparison-based random forests. In *Proceedings of the 35th International Conference on Machine Learning* (pp. 1871–1880).

Haghiri, S., Ghoshdastidar, D., & von Luxburg, U. (2017). Comparison based nearest neighbor search. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (pp. 851–859).

Heim, E. (2016). *Efficiently and effectively learning models of similarity from human feedback*. PhD diss., University of Pittsburgh.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*. https://openreview.net/group?id=ICLR.cc

Kleindessner, M. (2017). *Machine learning in a setting of ordinal distance information*. PhD diss., Eberhard Karls Universität Tübingen.

Kleindessner, M., & von Luxburg, U. (2017a). Kernel functions based on triplet comparisons. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems, 30* (pp. 6807–6817). Red Hook, NY: Curran.

Kleindessner, M., & von Luxburg, U. (2017b). Lens depth function and *k*-relative neighborhood graph: Versatile tools for ordinal data analysis. *Journal of Machine Learning Research*, *18*(1), 1889–1940.

Krizhevsky, A., & Hinton, G. (2009). *Learning multiple layers of features from tiny images* (Technical Report Vol. 1, no. 4). Toronto: University of Toronto.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Liu, C., Wu, K., & He, T. (2004). Sensor localization with ring overlapping based on comparison of received signal strength indicator. In *Proceedings of the 2004 IEEE International Conference on Mobile Ad-hoc and Sensor Systems* (pp. 516–518). Piscataway, NJ: IEEE.

Lu, N., Niu, G., Menon, A. K., & Sugiyama, M. (2019). On the minimal supervision for training any binary classifier from only unlabeled data. In *Proceedings of the International Conference on Learning Representations 2019*. https://openreview.net/group?id=ICLR.cc

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281–297). Berkeley: University of California, Berkeley.

Mojsilovic, S., & Ukkonen, A. (2019). Relative distance comparisons with confidence judgements. In *Proceedings of the 2019 SIAM International Conference on Data Mining* (pp. 459–467). Philadelphia: SIAM.

Moore, E. H. (1920). On the reciprocal of the general algebraic matrix. *Bull. Am. Math. Soc.*, *26*, 394–395.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning* (pp. 807–814).

Narasimhan, H., & Agarwal, S. (2013). On the relationship between binary classification, bipartite ranking, and binary class probability estimation. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems, 26* (pp. 2913–2921). Red Hook, NY: Curran.

Niu, G., Dai, B., Yamada, M., & Sugiyama, M. (2014). Information-theoretic semi-supervised metric learning via entropy regularization. *Neural Computation*, *26*(8), 1717–1762.

Niu, G., du Plessis, M. C., Sakai, T., Ma, Y., & Sugiyama, M. (2016). Theoretical comparisons of positive-unlabeled learning against positive-negative learning. In *NeurIPS* (pp. 1199–1207).

Penrose, B. Y. R. (1955). A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, *51* (3), 406–413.

Perrot, M., & von Luxburg, U. (2018). Boosting for comparison-based learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (pp. 1844–1850). Palo Alto, CA: AAAI Press.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 815–823). Piscataway, NJ: IEEE.

Schultz, M., & Joachims, T. (2004). Learning a distance metric from relative comparisons. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems, 17* (pp. 41–48). Red Hook, NY: Curran.

Shimada, T., Bao, H., Sato, I., & Sugiyama, M. (2019). *Classification from pairwise similarities/dissimilarities and unlabeled data via empirical risk minimization*. arXiv:1904.11717.

Stewart, N., Brown, G. D., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review*, *112*(4), 881.

Sugiyama, M. (2012). Learning under non-stationarity: Covariate shift adaptation by importance weighting. In J. E. Gontle, W. Härdle, & Y. Mori (Eds.), *Handbook of computational statistics* (pp. 927–952). Berlin: Singer.

Van Der Maaten, L., & Weinberger, K. (2012). Stochastic triplet embedding. In *Proceedings of the 2012 IEEE International Workshop on Machine Learning for Signal Processing* (pp. 1–6). Piscataway, NJ: IEEE.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Berlin: Springer-Verlag.

Xiao, H., Rasul, K., & Vollgraf, R. (2017). *Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms*. arXiv:1708.07747.

Xiao, L., Li, R., & Luo, J. (2006). Sensor localization based on nonmetric multidimensional scaling. *STRESS* 2:1.

Yu, B., Liu, T., Gong, M., Ding, C., & Tao, D. (2018). Correcting the triplet selection bias for triplet loss. In *Proceedings of the European Conference on Computer Vision* (pp. 71–87). Berlin: Springer.

Zhou, Z. H. (2017). A brief introduction to weakly supervised learning. *National Science Review*, *5*(1), 44–53.

Received July 17, 2019; accepted November 9, 2019.