Class Proportion Estimation with Application to Multiclass Anomaly Rejection

Tyler Sanderson ‡
University of Michigan in Ann Arbor

Clayton Scott
University of Michigan in Ann Arbor

Abstract

This work addresses two classification problems that fall under the heading of domain adaptation, wherein the distributions of training and testing examples differ. The first problem studied is that of class proportion estimation, which is the problem of estimating the class proportions in an unlabeled testing data set given labeled examples of each class. Compared to previous work on this problem, our approach has the novel feature that it does not require labeled training data from one of the classes. This property allows us to address the second domain adaptation problem, namely, multiclass anomaly rejection. Here, the goal is to design a classifier that has the option of assigning a "reject" label, indicating that the instance did not arise from a class present in the training data. We establish consistent learning strategies for both of these domain adaptation problems, which to our knowledge are the first of their kind. We also implement the class proportion estimation technique and demonstrate its performance on several benchmark data sets.

1 INTRODUCTION

This work studies two related classification problems that fall under the heading of domain adaptation, which is used to describe any learning problem where the distributions of training and testing instances differ. In particular, we study the problems of class proportion estimation (CPE) and multiclass anomaly rejection (MCAR). Both problems are studied in a multiclass setting, where the learner has access to a labeled training data set as well as an unlabeled testing data set. CPE is the problem of estimating the class proportions governing the unlabeled testing data, which may differ from those in the training data set. Unlike previous approaches to CPE, our approach has the novel feature that it does not require training data from one of the classes. This property allows us to address MCAR, where the goal is to design a classifier that may assign a "reject" label, indicating that the instance did not arise from a class present in the training data. We establish consistent learning strategies for both of these domain adaptation problems, which to our knowledge are the first of their kind. We also implement the CPE technique and demonstrate its performance on several benchmark data sets.

To begin, let us state the CPE problem. There are M classes, and a training sample for each class:

$$X_1^i, \ldots, X_{n_i}^i \overset{iid}{\sim} P_i, \tag{1}$$

where P_i is the ith class-conditional distribution, and X_j^i denotes the jth training sample from class i. In addition, there is an unlabeled testing sample

$$X_1^0, \ldots, X_{n_0}^0 \sim P_0 := \sum_{i=1}^{M} \pi_i P_i, \tag{2}$$

drawn from a mixture of the different classes. Here $\pi_i \geq 0$ and $\sum_i \pi_i = 1$. The critical feature of this problem is that the proportions π_i are unknown and different from the proportions represented in the training data, so that $n_i / \sum_\ell n_\ell$ is not a reasonable estimate. The goal is to estimate the π_i accurately, while making minimal assumptions on the P_i.

This form of domain adaptation arises frequently in applications where training and testing data are gathered according to different sampling plans. For example, training data gathered prospectively may have user-determined sample sizes, while testing data analyzed retrospectively have sample sizes that are beyond the user's control.

‡Current affiliation: Google Inc

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33. Copyright 2014 by the authors.

One motivation for class proportion estimation is design of a classifier for the test distribution. Suppose that there is a joint distribution on labels and instances with P_0 the marginal distribution on instances, P_i the class-conditional distributions, and π_i the prior distribution on labels. The risk of a classifier $f : \mathcal{X} \to \{1, \ldots, M\}$, $\mathcal{X} \subseteq \mathbb{R}^d$ denoting the feature space, may be expressed $R(f) := \sum_i \pi_i R_i(f)$ where $R_i(f) := P_i(\{x : f(x) \neq i\})$. The class-conditional errors R_i can be estimated since the training data provide examples from each class. However, the class proportions π_i need to be estimated in order to estimate the risk and thereby achieve good generalization.[1]

Our work is further motivated by MCAR, another domain adaptation problem. In particular, we consider the problem of having no training data from the last class ($n_M = 0$), which we consider to be the anomaly class. Many real problems fall into this category. For example, a classifier for object recognition will undoubtedly encounter object types in the real world not observed during training. The first $M - 1$ classes may be viewed as the known training classes, and predicting the Mth class amounts to a decision to "reject" an instance as not belonging to any of the known classes. This problem is more challenging than regular multiclass classification because estimation of $R_M(f)$ is no longer straightforward.

To summarize, this work makes the following contributions: It establishes the first methodology for CPE that is consistent in the case where a class is not observed. The first known consistent discrimination rule for MCAR is also introduced. Finally, we propose a practical implementation of our CPE methodology, and support this approach with experimental comparisons to existing methods.

On the technical side, our approach hinges on a reduction of CPE to another problem called *mixture proportion estimation*, reviewed below. To convert methods for CPE to a discrimination rule for MCAR, we also introduce a novel error estimation strategy for use with empirical risk minimization, and a corresponding uniform error analysis using multiclass VC theory.

2 RELATED WORK

Class proportion estimation goes back at least to Hall (1981), who introduced an approach for univariate data based on matching a weighted combination of class-conditional empirical distribution functions to the empirical distribution function of the unlabeled data. This idea was extended by Titterington (1983), who replaced empirical distribution functions by kernel density estimates, which allowed this "distribution matching" method to extend easily to multivariate data. The matching criterion is the L^2 distance between estimates of the marginal density P_0, and can be easily formulated as an unconstrained or constrained (if the class proportions are required to belong to a simplex) quadratic program. These authors established asymptotic normality of the estimated proportions under conditions that are typical of L^2 consistency for kernel density estimates. See Hall and Zhou (2003) for additional references on this strand of work.

Two other works in the machine learning literature have also addressed CPE. Latinne et al. (2001) introduced an EM algorithm in a logistic regression framework that adjusts class proportions to maximize the test data likelihood given the trained model. Du Plessis and Sugiyama (2012) developed an algorithm based on distribution matching but with a Kullback-Leibler criterion. None of the above cited works consider the case where one of the classes is unobserved, nor do they establish a consistent discrimination rule. Only Hall and Titterington provide theoretical analysis for CPE; Hall's analysis considers univariate data, while Titterington's assumes the existence of densities.

Multiclass anomaly rejection should not be confused with a problem known as "classification with reject option" (Chow, 1970). Despite the name, that problem is *not* concerned with rejection of anomalous instances. Rather, the classifier is allowed to *abstain* from labeling instances that are ambiguous, that is, near the boundary between two observed classes. The objective in that problem is to minimize the error rate conditioned on a label being assigned.

The framework of "zero-shot learning" can correctly classify previously unobserved classes, provided that additional semantic information about those classes is also available (Palatucci et al., 2009). The framework of Görnitz et al. (2013) develops semi-supervised one-class classifiers that leverage unlabeled data and are capable of rejecting anomalies, but no consistency result is known. In the binary case ($M = 2$), MCAR amounts to learning with positive and unlabeled examples (LPUE). Consistency for LPUE can be established with respect to the Neyman-Pearson criterion (Blanchard et al., 2010), but this analysis has not been extended to other performance measures or the multiclass setting. In the next section we recount a key contribution of Blanchard et al. (2010) that enables

[1] Note that there are two possible settings for evaluation. In a transductive setting, the goal is to assign labels to the given test examples, while in a semi-supervised setting, the goal is to use these unlabeled examples to design a general-purpose classifier for classifying future draws from P_0. We focus on the semi-supervised setting, which can be specialized to the transductive setting.

our own.

3 MIXTURE PROPORTION ESTIMATION

We will show that class proportion estimation reduces to mixture proportion estimation, which is now reviewed. Let $(\mathcal{X}, \mathfrak{S})$ be a measurable space, and let F, G, and H be distributions on $\mathcal{X}$ such that

$$F = (1-\nu)G + \nu H \tag{3}$$

where $0 \leq \nu \leq 1$. Mixture proportion estimation is the following problem: given iid training samples of sizes m and n from F and H respectively, and no information about G, estimate ν. This problem was first addressed in a distribution-free framework by Blanchard et al. (2010) and later applied to the problem of classification with label noise (Scott et al., 2013). In this section, we relate the necessary results from Blanchard et al. (2010) while following the notation of Scott et al. (2013).

Without additional assumptions, ν is not an identifiable parameter. Indeed, if $F = (1-\nu)G + \nu H$ holds, then any alternate decomposition of the form $F = (1-\nu+\delta)G' + (\nu-\delta)H$, with $G' = (1-\nu+\delta)^{-1}((1-\nu)G + \delta H)$, and $\delta \in [0, \nu)$, is also valid. With no knowledge of G, we cannot decide which representation is the correct one. Therefore, the idea is to impose a condition on G such that ν becomes identifiable. Toward this end, the following definition is introduced.

Definition 1. *Let G, H be probability distributions. G is said to be* irreducible *with respect to H if there exists no decomposition of the form $G = \gamma H + (1-\gamma)F'$, where F' is some probability distribution and $0 < \gamma \leq 1$.*

Some commentary on this definition is offered below. The following was established in Blanchard et al. (2010).

Proposition 1. *Let F, H be probability distributions. If $F \neq H$, there is a unique $\nu^* \in [0,1)$ and G such that the decomposition $F = (1-\nu^*)G + \nu^* H$ holds, and such that G is irreducible with respect to H. If we additionally define $\nu^* = 1$ when $F = H$, then in all cases,*

$$\nu^* := \max\{\alpha \in [0,1] : \exists \text{ a distribution } G' \text{ s.t. } F = (1-\alpha)G' + \alpha H\}.$$

By this result, the following is well-defined.

Definition 2. *For any two probability distributions F, H, define*

$$\nu^*(F,H) := \max\{\alpha \in [0,1] : \exists \text{ a distribution } G' \text{ s.t. } F = (1-\alpha)G' + \alpha H\}.$$

Thus, G is irreducible with respect to H if and only if $\nu^*(G,H) = 0$. Further, it is not hard to show that for any two distributions F and H, $\nu^*(F,H) = \inf_{A \in \mathfrak{S}} F(A)/H(A)$ (Scott et al., 2013). Similarly, when F and H have densities f and h, $\nu^*(F,H)$ is the essential infimum of $f(x)/h(x)$. These identities make it possible to check irreducibility in different scenarios. For example, $\nu^*(G,H) = 0$ whenever the support of G does not contain the support of H. Even if the supports are equal, irreducibility can still hold as in the case where g and h are two Gaussian densities with distinct means, where the variance of h is no smaller than the variance of g (Scott et al., 2013).

The following corollary summarizes the above and states that irreducibility of G w.r.t. H is a sufficient condition for ν in (3) to be identifiable.

Corollary 1. *If $F = (1-\gamma)G + \gamma H$, and G is irreducible with respect to H, then $\gamma = \nu^*(F,H)$.*

Blanchard et al. (2010) studied an estimator $\widehat{\nu} = \widehat{\nu}(\widehat{F}, \widehat{H})$ of $\nu^*(F,H)$, where $\widehat{F}$ and $\widehat{H}$ denote the empirical distributions based on iid random samples from F and H. They show in Thm. 8 that $\widehat{\nu}$ is strongly universally consistent, i.e., for any F and H, $\widehat{\nu} \to \nu^*(F,H)$ in probability as the sample sized tend to ∞.[2] We will show that this estimator leads to consistent estimators of class probabilities. The estimator is discussed further in Sec. 6.1.

4 CLASS PROPORTION ESTIMATION

In this section we apply mixture proportion estimation to CPE. Let $P_1, \ldots, P_M$ be probability measures (distributions) on $(\mathcal{X}, \mathfrak{S})$.

4.1 Identifiability Conditions

As with mixture proportion estimation, class proportion estimation requires an identifiability condition.

(A) For all $i = 1, \ldots, M$, every element of $\text{conv}\{P_\ell : \ell \neq i\}$ is irreducible with respect to P_i.

[2] More precisely, Blanchard et al. (2010) use the notation $\pi = 1 - \nu$, and present a consistent estimator for π. Furthermore, they actually establish almost sure convergence. As noted by Scott et al. (2013), the statement of Thm. 8 of Blanchard et al. (2010) needs to be amended slightly (by constraining how the two sample sizes grow w.r.t. each other) for almost sure convergence to hold.

Here $\text{conv}\{Q_1, \ldots, Q_K\}$ denotes the set of convex combinations of $Q_1, \ldots, Q_K$, that is, the set of mixture distributions based on $Q_1, \ldots, Q_K$. To illuminate **(A)**, we introduce a second condition, where $\text{supp}(Q)$ denotes the support of distribution Q.

(B) For all $i = 1, \ldots, M$, $\text{supp}(P_i) \nsubseteq \cup_{\ell \neq i} \text{supp}(P_\ell)$.

(B) clearly implies **(A)** from the definition of irreducible.

We argue that **(B)** is a reasonable assumption in many real-world classification problems, and therefore so is **(A)**. In words, **(B)** means that for each class, there exist at least some instances, with positive probability of occurring (however small), that are always correctly classified by an optimal classifier. In other words, such instances could not possibly be mistaken for instances of another class. For example, consider handwritten digit recognition. Although various classes may have overlapping supports, each class has instances (corresponding to very clear handwriting, say) that could not possibly be mistaken for any other class.

4.2 Consistency in the Fully Observed Case

For now assume training samples from all M classes are observed. Under **(A)**, the proportions π_i are identifiable, and we propose to estimate them via

$$\widehat{\pi}_i = \widehat{\nu}(\widehat{P}_0, \widehat{P}_i) \tag{4}$$

for $i = 1, \ldots, M$, where $\widehat{\nu}$ is the estimator of Blanchard et al. (2010) discussed in the previous section.

Proposition 2. *Under* **(A)**, *for each* i, $\widehat{\pi}_i$ *converges to* π_i *in probability as* $\min\{n_0, n_i\} \to \infty$.

Proof. WLOG assume $i = 1$. Now $P_0 = \pi_1 P_1 + (1 - \pi_1)Q$ where $Q \in \text{conv}\{P_\ell : \ell \neq 1\}$. Under **(A)**, $\nu^*(Q, P_1) = 0$, and therefore by Corollary 1, $\pi_1 = \nu^*(P_0, P_1)$. The result now follows by convergence in probability of $\widehat{\nu}(\widehat{P}_0, \widehat{P}_1)$ to $\nu^*(P_0, P_1)$. □

When $M = 2$, **(A)** says $\nu^*(P_1, P_2) = 0$ and $\nu^*(P_2, P_1) = 0$. This is the so-called *mutual irreducibility* assumption adopted by Scott et al. (2013) in the context of label noise. It turns out that when $M = 2$ we can consistently estimate the proportions under a weaker condition, namely, $P_1 \neq P_2$. To achieve this, we employ the following estimators:

$$\widehat{\pi}'_1 := \frac{1 - \widehat{\nu}(\widehat{P}_0, \widehat{P}_2)}{1 - \widehat{\nu}(\widehat{P}_1, \widehat{P}_2)}, \quad \widehat{\pi}'_2 := \frac{1 - \widehat{\nu}(\widehat{P}_0, \widehat{P}_1)}{1 - \widehat{\nu}(\widehat{P}_2, \widehat{P}_1)}.$$

The intuition is that in the binary case, even if **(A)** is violated, say $\nu^*(P_1, P_2) > 0$, we can use mixture proportion estimation to estimate $\nu^*(P_1, P_2)$, and rescale the estimates accordingly. Note that each of these modified estimators uses all three samples, and therefore this result does not generalize to the case where one class is unobserved.

Proposition 3. *If* $M = 2$ *and* $P_1 \neq P_2$, *then* $\widehat{\pi}'_1 \to \pi_1$ *in probability and* $\widehat{\pi}'_2 \to \pi_2$ *in probability, as* $\min\{n_0, n_1, n_2\} \to \infty$.

Proof. Consider estimation of π_1. Denote $\nu_{12} = \nu^*(P_1, P_2)$. By Proposition 1, there exists a unique distribution E_1 such that $P_1 = (1 - \nu_{12})E_1 + \nu_{12}P_2$ and $\nu(E_1, P_2) = 0$. Then

$$\begin{aligned} P_0 &= \pi_1[(1 - \nu_{12})E_1 + \nu_{12}P_2] + (1 - \pi_1)P_2 \\ &= \pi_1(1 - \nu_{12})E_1 + [\pi_1\nu_{12} + (1 - \pi_1)]P_2. \end{aligned}$$

Since $\nu(E_1, P_2) = 0$, by Corollary 1 we must have $\nu^*(P_0, P_2) = \pi_1\nu_{12} + (1 - \pi_1)$. Solving for π_1 yields $\pi_1 = \frac{1 - \nu^*(P_0, P_2)}{1 - \nu^*(P_1, P_2)}$. Since $P_1 \neq P_2$, the denominator is nonzero. The result now follows by consistency of $\widehat{\nu}$ and continuity of division. □

4.3 Consistent CPE with an Unobserved Class

The primary advantage of our approach to CPE is that it can consistently estimate all proportions, even π_M, when $n_M = 0$. The estimators $\widehat{\pi}_i$ of Eqn. (4) do not depend on $\widehat{P}_M$ when $i < M$, so they can remain the same in this setting. For $i = M$, we can just set $\widehat{\pi}_M := 1 - \sum_{i=1}^{M-1} \widehat{\pi}_i$. The following is an immediate consequence of the necessary condition $\sum_{i=1}^{M} \pi = 1$ and the consistency of $\widehat{\pi}_1, \ldots, \widehat{\pi}_{M-1}$.

Corollary 2. *Consider class proportion estimation where* $n_M = 0$. *Let* $\widehat{\pi}_i$ *be as in Eqn.* (4) *for* $i = 1, \ldots, M-1$, *and set* $\widehat{\pi}_M = 1 - \sum_{i=1}^{M-1} \widehat{\pi}_i$. *Under* **(A)**, *for each* $i = 1, \ldots, M$, $\widehat{\pi}_i$ *converges to* π_i *in probability as* $\min\{n_0, n_1, \ldots, n_{M-1}\} \to \infty$.

5 ANOMALY REJECTION

We now turn our attention to the design of a consistent discrimination rule for MCAR. In this setting, available data consist of iid random samples from $P_1, \ldots, P_{M-1}$ as in (1), and an iid random sample from P_0 as in (2). Data from P_M are not observed. Our goal is a discrimination rule $\widehat{f}$, constructed from the available data, whose risk converges to the Bayes risk as the various sample sizes tend to ∞. Note that previous work has not addressed this problem even in the case where all classes are observed (which still differs from standard classification because the test distribution has different class proportions).

To set notation, let Q denote the joint distribution of $(X, Y) \in \mathcal{X} \times \{1, \ldots, M\}$ such that the X-marginal of Q is P_0, the Y-marginal is given by the π_i, and the class-conditional distributions are P_i. For any classifier $f : \mathcal{X} \to \{1, \ldots, M\}$, denote the class-conditional error probabilities $R_i(f) := P_i(\{x : f(x) \neq i\})$, and the test-distribution risk $R(f) := Q(\{(x, y) : f(x) \neq y\}) = \sum_{i=1}^{M} \pi_i R_i(f)$. Let R^* denote the Bayes risk for distribution Q. Our goal is to construct a discrimination rule $\widehat{f}$ such that $R(\widehat{f}) \to R^*$ in probability as the sample sizes $n_0, n_1, \ldots, n_{M-1}$ tend to ∞.

To construct such a rule, we adapt a classic strategy from statistical learning theory (Devroye et al., 1996): empirical risk minimization (ERM) over a growing family of classifiers, also known as sieve estimation. This strategy relies upon VC theory, and since we are in a multiclass setting, we take the following generalization of VC dimension to multiclass. Define the (multiclass) VC dimension of a set of classifiers $\mathcal{F}$ to be the maximum conventional (two-class) VC dimension (Devroye et al., 1996) of the family of sets $\{x : f(x) \neq \ell\}_{f \in \mathcal{F}}$, over $\ell = 1, \ldots, M$.

As its name suggests, ERM also requires an estimate of the risk. We propose to estimate $R(f)$ by writing $R(f) = \sum_{i=1}^{M-1} \pi_i R_i(f) + \underline{R}_M(f)$, where $\underline{R}_M(f) := Q(\{(x, y) : f(x) \neq y, y = M\}) = \pi_M R_M(f)$, and estimating each term in this expression. For $i < M$, π_i is estimated by $\widehat{\pi}_i$ in Eqn. (4), and $R_i(f)$ is estimated by $\widehat{R}_i(f) := \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{1}_{\{f(X_j^i) \neq i\}}$. An estimate of $\underline{R}_M(f)$ is motivated as follows. Let $R_{iM}(f) := P_i(\{x : f(x) \neq M\})$ and observe that $R_{0M}(f) = \sum_{i=1}^{M-1} \pi_i R_{iM}(f) + \pi_M R_M(f)$. Then

$$\underline{R}_M(f) = R_{0M}(f) - \textstyle\sum_{i=1}^{M-1} \pi_i R_{iM}(f). \tag{5}$$

Plugging in $\widehat{R}_{iM}(f) := \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{1}_{\{f(X_j^i) \neq M\}}$ and our estimates for the π_i leads to the following estimator:

$$\widehat{\underline{R}}_M(f) = \widehat{R}_{0M}(f) - \textstyle\sum_{i=1}^{M-1} \widehat{\pi}_i \widehat{R}_{iM}(f). \tag{6}$$

Now set $\widehat{R}(f) := \sum_{i=1}^{M-1} \widehat{\pi}_i \widehat{R}_i(f) + \widehat{\underline{R}}_M(f)$.

We now define the ERM-based discrimination rule. Let $(\mathcal{F}_k)_{k \geq 1}$ be a sequence of VC classes with corresponding (multiclass) VC dimensions $V_k < \infty$. Let τ_k be any sequence of positive numbers tending to zero. Let $\widehat{f}_k$ be an approximate empirical risk minimizer, i.e., any classifier

$$\widehat{f}_k \in \Big\{ f \in \mathcal{F}_k \,:\, \widehat{R}(f) \leq \inf_{f' \in \mathcal{F}_k} \widehat{R}(f') + \tau_k \Big\}.$$

The introduction of τ_k lets us avoid assuming the existence of an empirical risk minimizer. Denote $\boldsymbol{n} := (n_0, n_1, \ldots, n_{M-1})$. We write $\boldsymbol{n} \to \infty$ to indicate $\min\{n_0, n_1, \ldots, n_{M-1}\} \to \infty$. Let $k(\boldsymbol{n})$ denote a sequence of positive integers indexed by $\boldsymbol{n}$. Finally, define the discrimination rule $\widehat{f} := \widehat{f}_{k(\boldsymbol{n})}$. Note that the sequences $(\mathcal{F}_k)_{k \geq 1}$ and $k(\boldsymbol{n})$ are user-specified and must grow in a certain way, indicated by the theory below, for $\widehat{f}$ to be consistent.

Analysis of this discrimination rule hinges on uniform control of the deviation $|R(f) - \widehat{R}(f)|$ over $\mathcal{F}_{k(\boldsymbol{n})}$ as $\boldsymbol{n} \to \infty$. The following result establishes this property. In the proof, the error deviance is decomposed in such a way that uniform control follows from the multiclass VC extension and consistency of the class proportion estimators. The proof of this and the next result are found in the supplemental material.

Proposition 4. *Assume* **(A)** *holds and suppose* $k(\boldsymbol{n}) \to \infty$ *as* $\boldsymbol{n} \to \infty$ *such that*

$$\frac{V_{k(\boldsymbol{n})} \log n_i}{n_i} \to 0, \tag{7}$$

for $0 \leq i \leq M - 1$. *Then*

$$\sup_{f \in \mathcal{F}_{k(\boldsymbol{n})}} |R(f) - \widehat{R}(f)| \to 0$$

in probability as $\boldsymbol{n} \to \infty$.

So that arbitrary classifiers can be accurately approximated, we choose $(\mathcal{F}_k)_{k \geq 1}$ satisfying the following universal approximation property: For any joint distribution Q on $\mathcal{X} \times \{1, \ldots, M\}$,

$$\lim_{k \to \infty} \inf_{f \in \mathcal{F}_k} R(f) = R^*$$

where R^* is the Bayes error corresponding to Q. Devroye et al. (1996) give examples of families of VC classes that satisfy the above approximation property. We can now state the main result of this section.

Theorem 1. *Assume* **(A)** *holds and that* $(\mathcal{F}_k)_{k \geq 1}$ *is chosen to satisfy the universal approximation property above. Further suppose* $k(\boldsymbol{n})$ *is chosen such that as* $\boldsymbol{n} \to \infty$, $k(\boldsymbol{n}) \to \infty$ *and* (7) *holds for* $0 \leq i \leq M - 1$. *Then* $R(\widehat{f}) \to R^*$ *in probability.*

Although we have focused on the probability of error as a performance measure, it would not be difficult to adapt this result to any other performance measure that is a continuous function of the class proportions π_i and class-conditional errors R_i, such as a cost-sensitive Bayes risk or the minmax error.

6 IMPLEMENTATION AND EXPERIMENTS

In this section we introduce a practical algorithm for mixture proportion estimation (MPE) and use it to implement the proposed CPE methodology. We then

compare our method to existing methods for CPE on a variety of binary and multiclass data sets. We consider two experimental settings. In the first setting, we adopt the assumption that the unlabeled test data do not contain an anomalous class. This is the assumption adopted by competing methods and, not surprisingly, we find that they outperform our own approach, which allows for the existence of an anomalous class in the test data. In the second group of experiments, the test data contain an anomalous class, and our approach vastly outperforms the competitors in this scenario.

For a fairer head-to-head comparison with existing methods, we introduce two additional class proportion estimators based on MPE that make the same assumptions as competing methods (namely, that there is not an anomalous class in the test data). We compare these to existing methods under the first experimental setting and find they are competitive, which offers experimental validation of the MPE-based framework.

A thorough experimental investigation of MCAR is beyond the scope of this work. The discrimination rule we introduce for MCAR could be implemented for various VC classes such as histograms or decision trees, but other methods would also be worthy of exploration, such as those based on convex surrogate losses.

6.1 Practical Algorithm for MPE

As discussed in Scott et al. (2013), Theorem 6 of Blanchard et al. (2010) tells us $\nu^* = \nu^*(F, H)$ is related to the optimal Receiver Operating Characteristic (ROC) that arises when the distribution H is viewed as the null hypothesis and F as the alternative. This optimal ROC is the function[3]

$$p(\alpha) := \sup_{C \subset \mathcal{X}} F(C) \quad \text{s.t. } H(C) \leq \alpha.$$

This function gives the optimal detection probability of a binary classifier constrained to have false alarm rate no more than α, where C here represents a subset of $\mathcal{X}$ that predicts the class of F.

As shown in Blanchard et al. (2010); Scott et al. (2013), $\nu^* = \frac{dp}{d\alpha}\Big|_{\alpha=1^-}$, the slope of the optimal ROC evaluated at the right endpoint where the false positive rate becomes 1. The estimator $\widehat{\nu}$ studied in Blanchard et al. (2010) implements this principle, but relies on distribution free confidence intervals (to achieve universal consistency), and thus tends to be too conservative in practice.

[3]Technically, if the function is not concave, the optimal ROC is the smallest concave function that upper bounds $p(\alpha)$.

Therefore we introduce a more practical implementation of the above principle for MPE, and apply it to CPE. Given random samples $\widehat{F}$ and $\widehat{H}$ from F and H, we treat these as training classes for a binary classification problem, and train a kernel logistic regression (KLR) classifier using a Gaussian kernel. We then vary the threshold on the KLR posterior class probability to generate an empirical version of the optimal ROC, and obtain $\widehat{\nu}$ by estimating the slope of this empirical ROC at its right endpoint. Note that the choice to use KLR is simply for convenience, and any binary classifier capable of producing an ROC, such as cost-sensitive SVMs, could be used instead.

Since the empirical ROC may be noisy at its right endpoint, we fit a curve to the empirical ROC and take the right endpoint slope of the fitted curve to be our proportion estimate. Lloyd (2000) provides two regression models for ROCs, and we augment them both to include an extra linear term in an attempt to better model the linear behavior seen towards the right end of the ROC.

In particular, for a given ROC, let α denote the false positive rate, $p(\alpha)$ the corresponding detection rate, and $f(\alpha)$ the model for $p(\alpha)$. Our regression models are:

$$f_{\gamma,\Delta}(\alpha) = (1-\gamma)Q(Q^{-1}(\alpha) + \Delta) + \gamma\alpha. \quad (8)$$

$$f_{\gamma,\Delta,\mu}(\alpha) = (1-\gamma)(1 + \Delta(\alpha^{-\mu} - 1))^{-\frac{1}{\mu}} + \gamma\alpha. \quad (9)$$

where Q is the standard normal CDF, Δ controls ROC quality, μ is an asymmetry parameter, and γ is the slope of the added linear component. See Lloyd (2000) for more insight into the form of these models.

Since the domain and range of the ROC are probabilities, we fit the models by minimizing the binomial deviance between the empirical ROC given by $\widehat{\alpha}_j$ and $\widehat{p}_j$, where $j = 1, \ldots, n$ indexes sample points along the empirical ROC, and the model $f(\widehat{\alpha})$ as given by Eqns. (8) or (9):

$$B_f(\widehat{\alpha}, \widehat{p}) = -2\sum_{j=1}^{n} \widehat{p}_j \log(f(\widehat{\alpha}_j)) + (1-\widehat{p}_j)\log(1-f(\widehat{\alpha}_j))$$

The right-endpoint slope of the model as a function of the fitted parameters is γ in the case of (8) and $(1-\gamma)\Delta + \gamma$ in the case of (9).

6.2 New MPE-based Algorithms for CPE

We apply the above algorithm to CPE following the framework of Sec. 4, so that $\widehat{\pi}_i := \widehat{\nu}(\widehat{P}_0, \widehat{P}_i)$, where recall $\widehat{P}_0$ and $\widehat{P}_i$ represent the data drawn from the unlabeled test distribution and training class i respectively. In the first set of experiments, there are M

observed training classes, and our method allows for the existence of an $(M+1)st$ class, estimating $\widehat{\pi}_{M+1} = 1 - \sum_{i=1}^{M} \widehat{\pi}_i$. In the second set of experiments, there are $M-1$ training classes, and the anomalous class proportion π_M is estimated as $\widehat{\pi}_M = 1 - \sum_{i=1}^{M-1} \widehat{\pi}_i$. We found the model from Eqn. (9) performed best. In the results we denote this CPE method as *MPE-Incomplete* since it assumes incomplete knowledge of the classes.

In the fully observed case (the first experimental setting), we showed in Sec. 4.2 that our approach consistently estimates the true class proportions. However, due to estimation error the estimates $\widehat{\pi}_1, \ldots, \widehat{\pi}_M$ do not sum to one, as they should in this setting. Therefore, for a fairer comparison with existing methods, we also introduce two extensions of MPE-based CPE that, like previous methods, do not support an anomalous class in the test data, but do perform better when all classes are observed.

The first extension is to simply project the vector of estimated proportions onto the probability simplex Δ^M. In the results, we denote this projected estimate as *MPE-Projected*.

The second extension forms M empirical ROCs based on the distributions (P_0, P_i), $i = 1, \ldots, M$, and fits all ROC curves simultaneously while constraining the estimated class proportions to sum to one. We use the model from Eqn. (8) since the slope at the right endpoint is simply γ. Letting f be Eqn. (8), and B_f the binomial deviance given above, we solve

$$\underset{\gamma_i, \Delta_i}{\text{minimize}} \quad \sum_{i=1}^{M} B_f(\widehat{\alpha}^i, \widehat{p}^i), \text{ subject to} \sum_{i=1}^{M} \gamma_i = 1$$

where $(\widehat{\alpha}^i, \widehat{p}^i)$ is the empirical ROC based on $\widehat{P}_0$ and $\widehat{P}_i$. This extension is denoted *MPE-Joint*.

6.3 Evaluation

Recall that we consider two experimental settings. In the first, all training classes are observed, while in the second, the Mth class is not observed.

We compare against several approaches noted in the related work section. We denote the methods by Latinne et al. (2001), Titterington (1983), and Du Plessis and Sugiyama (2012) as EM, L^2 Distance, and KL-Divergence[4], respectively. Since the EM algorithm requires posterior class probabilities, we use kernel logistic regression in both the EM algorithm and our method. Finally, we compare against a simple baseline estimate defined as the proportions of the labels predicted by a KLR classifier on the test data.

Our experiments were conducted on 13 well-known binary data sets and 5 multiclass data sets. Each data set was permuted 10 times and performance was computed by averaging over permutations. To measure performance we use the ℓ_1-norm between the estimated class proportion vector and the vector of true class proportions. For each data set and permutation, we manually set the class proportion of the Mth class to range over the following set of values: {1%, 10%, 20%, ..., 90%, 99%}. In the binary case, the positive class proportion was taken to be the Mth class ($M = 2$). In the multiclass case, the largest class in the original data set was taken to be the Mth class. The size of both the training set and testing set were kept constant over all proportions. As a result, as the M-th class grows the remaining classes shrink proportionately.

In the first experimental setting, the Mth class is observed. Under the assumption that all classes are observed, and to fairly compare to the other methods, in this scenario we discard the estimate of the $(M+1)st$ class proportion for the *MPE-Incomplete* method. Table 1 reports the ℓ_1-norm performance measure means and standard deviations, where the average is taken over permutation and varied class proportion. Fig. 1 shows the performance of each method, averaged over the binary data sets, as a function of the artificially modified class proportion.

The results show that the *MPE-Projected* and *MPE-Joint* extensions are comparable to the best performing algorithms in the binary case, and achieve the best performance on a few data sets. In some multiclass data sets the baseline error is low indicating the classes are highly separable. The EM algorithm often performed well but had high variance. The L^2 Distance method performed consistently well and best overall. The *MPE-Incomplete* method does not assume the test distribution contains only training classes, yet, it still performs reasonably well. Using a Wilcoxon signed rank test, we found the mean performances (across data set and varied proportion) of the algorithms were significantly different at the 5% level, except the *MPE-Projected*, *MPE-Joint*, and EM methods in the binary case were mutually insignificant from each other.

In the second experimental setting, the Mth class is not available to the various algorithms. Since competing methods do not natively support this scenario, we allow them to estimate the class proportions of classes they have observed and set their estimate of the anomalous class proportion to zero. Predictably,

[4]Due to computational constraints, we limited the input to the KL-Divergence method to 1000 training and 1000 testing examples, and were not able to use it in the multiclass setting.

Table 1: Comparison of mean performances with standard deviations, taken over all data permutations and resampled proportions.

Data set (M)	MPE-Incomplete	MPE-Projected	MPE-Joint	EM-KLR	L^2 Dist.	KL-Diverg.	baseline
All Binary	.188 ± .20	.131 ± .17	.140 ± .20	.145 ± .21	**.104 ± .12**	.155 ± .17	.270 ± .39
All Multiclass	.143 ± .08	.137 ± .09	.114 ± .07	.098 ± .14	.109 ± .08	n/a	**.097 ± .10**
Australian (2)	.169 ± .12	.132 ± .13	.094 ± .07	.096 ± .08	**.077 ± .06**	.164 ± .14	.179 ± .12
Banana (2)	.045 ± .04	.030 ± .04	.019 ± .02	**.016 ± .02**	.128 ± .08	.296 ± .22	.117 ± .07
Breast-cancer (2)	.535 ± .20	.312 ± .24	.488 ± .32	.442 ± .35	**.234 ± .17**	.235 ± .19	.875 ± .58
Diabetes (2)	.221 ± .10	.152 ± .11	.201 ± .17	.133 ± .12	**.112 ± .09**	.182 ± .18	.393 ± .29
German (2)	.307 ± .15	.188 ± .17	.219 ± .18	.211 ± .17	**.146 ± .10**	.180 ± .13	.645 ± .47
Image (2)	.086 ± .06	.066 ± .06	.044 ± .04	**.020 ± .02**	.083 ± .07	.134 ± .11	.053 ± .04
Ionosphere (2)	.217 ± .17	.176 ± .17	.129 ± .11	**.052 ± .04**	.125 ± .10	.140 ± .12	.098 ± .08
Ringnorm (2)	.023 ± .03	.018 ± .03	**.010 ± .01**	.165 ± .20	.014 ± .01	.022 ± .01	.018 ± .01
Saheart (2)	.406 ± .20	.283 ± .22	.364 ± .27	.222 ± .19	**.184 ± .15**	.225 ± .18	.552 ± .39
Splice (2)	.088 ± .07	.073 ± .07	**.049 ± .05**	.050 ± .03	.050 ± .04	.080 ± .06	.105 ± .06
Thyroid (2)	.265 ± .19	.204 ± .20	**.153 ± .13**	.183 ± .28	.163 ± .17	.300 ± .25	.339 ± .54
Twonorm (2)	.022 ± .02	.018 ± .01	**.010 ± .01**	.269 ± .21	.010 ± .01	.023 ± .01	.025 ± .01
Waveform (2)	.063 ± .04	.045 ± .03	.043 ± .03	.028 ± .02	**.019 ± .02**	.036 ± .03	.113 ± .07
SensIT (3)	.189 ± .08	.140 ± .09	.169 ± .08	.340 ± .16	**.104 ± .06**	n/a	.210 ± .12
DNA (3)	.080 ± .04	.074 ± .04	.048 ± .03	**.025 ± .02**	.062 ± .03	n/a	.055 ± .02
Opportunity (4)	.154 ± .07	.158 ± .08	.116 ± .05	**.067 ± .04**	.156 ± .14	n/a	.136 ± .09
SatImage (6)	.109 ± .06	.115 ± .08	.085 ± .04	**.031 ± .01**	.083 ± .04	n/a	.059 ± .02
Segment (7)	.183 ± .08	.196 ± .11	.152 ± .07	.027 ± .01	.139 ± .05	n/a	**.025 ± .02**

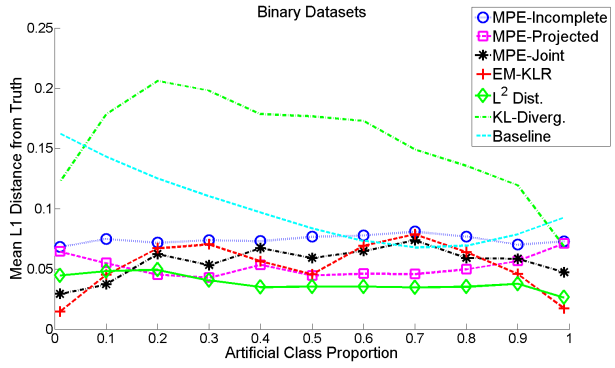

Figure 1: Mean performance over all permutations and binary data sets as manipulated class proportion changes.

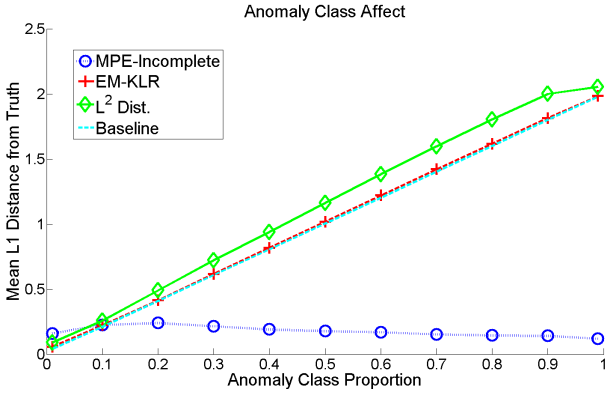

Figure 2: Mean performance over all permutations and multiclass data sets as anomaly class proportion changes.

as shown in Fig. 2, the performances of competing methods (averaged over data sets) rise linearly as the anomalous class proportion grows. The *MPE-Incomplete* method, in contrast, adapts to the anomalous class.

In the supplemental material, additional details of the experiments are reported. We also describe a method that successfully estimates confidence intervals on the π_i, with experimental results.

7 CONCLUSION

This work has demonstrated, both theoretically and experimentally, that *mixture proportion estimation* can be successfully applied to the problem of class proportion estimation. Unlike existing methods for CPE, our approach is able to accurately estimate the proportion of an anomalous class in the unlabeled test data. This feature of our method facilitates error estimation with respect to the test distribution, which forms the basis of a consistent discrimination rule for multiclass anomaly rejection. These approaches based on MPE are, to our knowledge, the first viable solutions to these two fundamental domain adaptation problems.

Acknowledgements

C. Scott was supported in part by NSF Grants 0953135, 1047871, and 1217880.

References

P. Hall. On the non-parametric estimation of mixture proportions. *Journal of the Royal Statistical Society*, 43(2):147–156, 1981.

D. M. Titterington. Minimum distance non-parametric estimation of mixture proportions. *Journal of the Royal Statistical Society*, 45(1):37–46, 1983.

P. Hall and X.-H. Zhou. Nonparametric estimation of component distributions in a multivariate mixture. *Annals of Statistics*, 31(1):201–224, 2003.

P. Latinne, M. Saerens, and C. Decaestecker. Adjusting the outputs of a classier to new a priori probabilities may signicantly improve classication accuracy: Evidence from a multi-class problem in remote sensing. In C. Sammut and A. H. Hoffmann, editors, *Proc. 18th Int. Conf. on Machine Learning*, pages 298–305, 2001.

M. C. Du Plessis and M. Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. In J. Langford and J. Pineau, editors, *Proc. 29th Int. Conf. on Machine Learning*, pages 823–830, 2012.

C. K. Chow. On optimum error and reject trade-off. *IEEE Transactions on Information Theory*, 16:41–46, 1970.

M. Palatucci, D. Pomerleau, G. E Hinton, and T. Mitchell. Zero-shot learning with semantic output codes. In *Advances in neural information processing systems*, pages 1410–1418, 2009.

N. Görnitz, M. Kloft, K. Rieck, and U. Brefeld. Toward supervised anomaly detection. *J. Artif. Intell. Res.(JAIR)*, 46:235–262, 2013.

G. Blanchard, G. Lee, and C. Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11:2973–3009, 2010.

C. Scott, G. Blanchard, and G. Handy. Classification with asymmetric label noise: Consistency and maximal denoising. In *Proc. 2013 Conference on Learning Theory, JMLR W&CP 30*, pages 489–511, 2013.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

C. Lloyd. Regression models for convex ROC curves. *Biometrics*, 56(3):862–867, September 2000.

D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, G. Trster, P. Lukowicz, G. Pirkl, D. Bannach, A. Ferscha, J. Doppler, C. Holzmann, M. Kurz, G. Holl, R. Chavarriaga, H. Sagha, H. Bayati, and J. Mill n. Collecting complex activity data sets in highly rich networked sensor environments. In *Proc. 7th Int. Conf. on Networked Sensing Systems*, 2010.

M. Duarte and Y. H. Hu. Vehicle classification in distributed sensor networks. *Journal of Parallel and Distributed Computing*, 64(7):826–838, 2004.