Robust Covariate Shift Classification Using Multiple Feature Views

Anqi Liu **Hong Wang** **Brian D. Ziebart**
Department of Computer Science
University of Illinois at Chicago
Chicago, IL 60607
{aliu33, hwang207, bziebart}@uic.edu

Abstract

The covariate shift learning setting relaxes the widely-employed *independent and identically distributed* (IID) assumption by allowing different training and testing input distributions. Unfortunately, common methods for addressing covariate shift by trying to remove the bias between training and testing distributions using importance weighting often provide poor performance guarantees in theory and unreliable predictions with high variance in practice. Recently developed methods that construct a predictor that is inherently robust to the difficulties of learning under covariate shift are often too conservative when faced with high-dimensional learning tasks. We introduce a generalization of robust covariate shift classification that allows the influence of covariate shift to be limited to different feature-based views of the relationship between input variables and example labels. We demonstrate the benefits of this approach on classification under covariate shift tasks.

1 Introduction

The *independent and identically distributed* (IID) assumption employed widely across machine learning methods is quite restrictive. However, shift often occurs between the training distribution and testing distribution that makes models built on the IID assumption inappropriate. Specifically, the predictor minimizing (regularized) loss on the training samples provides no performance guarantees when applied to the testing distribution [14, 5]. Though nothing can be learned when the shift between training and testing data is arbitrary, under **covariate shift** only the distribution of inputs, $P_{\text{train}}(\mathbf{x})$ and $P_{\text{test}}(\mathbf{x})$, differ, while the conditional label distribution, $P(y|\mathbf{x})$, is the same under both the training and the testing distributions.

The most common methods for addressing covariate shift attempt to debias the training data by reweighting it by a density ratio, $P_{\text{test}}(\mathbf{x})/P_{\text{train}}(\mathbf{x})$. This approach tends to work well when the training and the testing distributions are fairly similar and large amounts of training samples are available. However, given limited amounts of training data and/or significant differences between the training and testing distributions, some of the density ratios for training examples can be extremely large. This leads to high-variance estimates that extrapolate heavily from scant amounts of training data and a lack of generalization guarantees for the resulting predictor [2, 3].

Recently developed robust covariate shift methods take a worst-case approach, constructing a predictor that (approximately) matches training data statistics, but is otherwise the most uncertain on the testing distribution [7, 1]. Unfortunately, these methods can be *too* robust, providing overly conservative predictions that are nearly uniform, particularly when the dimensionality of the input space is large—a situations that is problematic for importance weighted loss minimization as well. We seek a better balance between making generalization assumptions and robustness to uncertainty by introducing feature view-based generalization assumptions to the robust covariate shift approach.

2 Background

Covariate Shift: Under covariate shift, the training distribution and testing distribution share the same conditional label distribution, $P(y|\mathbf{x})$, but have differing distributions over inputs: $P_{\text{train}}(\mathbf{x}, y) = P_{\text{train}}(\mathbf{x})P(y|\mathbf{x})$ and $P_{\text{test}}(\mathbf{x}, y) = P_{\text{test}}(\mathbf{x})P(y|\mathbf{x})$. (The IID assumption [8] further imposes that $P_{\text{train}}(\mathbf{x}) = P_{\text{test}}(\mathbf{x})$.) Unfortunately, constructing a predictor with limited complexity that performs well on the training data does not guarantee good performance on the testing distribution.

Debiasing via Importance Weighting: The prevalent approach for addressing covariate shift attempts to remove the bias between the training and testing distributions by reweighting [10, 6, 13]:

$$\lim_{n\to\infty} \min_{\theta} \mathbb{E}_{(\mathbf{X},Y)\sim\tilde{P}^{(n)}_{\text{train}}} \left[\frac{P_{\text{test}}(\mathbf{X})}{P_{\text{train}}(\mathbf{X})} \text{loss}(\hat{f}_\theta(\mathbf{X}), Y) \right] = \min_{\theta} \mathbb{E}_{(\mathbf{X},Y)\sim P_{\text{test}}} \left[\text{loss}(\hat{f}_\theta(\mathbf{X}), Y) \right], \tag{1}$$

which asymptotically minimizes testing loss, so long as $P_{\text{test}}(\mathbf{x}) > 0 \implies P_{\text{train}}(\mathbf{x}) > 0$. Despite this guarantee, predictive performance can be poor when training from finite amounts of samples in both theory and practice. In fact, finite generalization bounds require finite second moments: $\mathbb{E}_{P_{\text{train}}(x)}[(P_{\text{test}}(X)/P_{\text{train}}(X))^2] < \infty$ [2], which is often not satisfied in practice.

3 Robust Multiview Covariate Shift Classification

Covariate Shift Adversarial Game Formulation: Robust covariate shift classification [7] formulates the prediction task as an adversarial game between a predictor player choosing a conditional label distribution $\hat{P}(y|\mathbf{x})$ that minimizes the logarithmic loss and an adversarial player choosing $\check{P}(y|\mathbf{x})$ to maximize logarithmic loss:

$$\min_{\hat{P}} \max_{\check{P}\in\tilde{\Xi}_{\text{train}}} \mathbb{E}_{\mathbf{X}\sim P_{\text{test}},\check{Y}|\mathbf{X}\sim\check{P}} \left[-\log \hat{P}(\check{Y}|\mathbf{X}) \right], \qquad P(y|\mathbf{x}) \propto e^{\frac{P_{\text{train}}(\mathbf{x})}{P_{\text{test}}(\mathbf{x})}\theta\cdot\phi(\mathbf{x},y)}. \tag{2}$$

The adversary must choose a distribution that is similar to certain measured properties (features), e.g., $\mathbb{E}_{\mathbf{X}\sim\tilde{P},\check{Y}|\mathbf{X}\sim\check{P}}\left[\phi(\mathbf{X},\check{Y})\right] = \mathbb{E}_{(\mathbf{X},Y)\sim\tilde{P}}\left[\phi(\mathbf{X},Y)\right]$, of the training data. These are denoted by the convex set $\tilde{\Xi}_{\text{train}}$. The solution to this adversarial formulation has a parametric form with the density ratio appearing as: and moderates the uncertainty of the predictor to be larger for inputs that are relatively less likely in the training data. This formulation provides significant robustness guarantees: if $\tilde{\Xi}_{\text{train}}$ is constructed so that it contains the true conditional label distribution $P(y|\mathbf{x})$, then the adversarial game value upper bounds the loss on the testing distribution. Unfortunately, when the input space is high-dimensional this robustness guarantee can be *too* conservative, leading to predictions that are almost completely uncertain (i.e., uniform distributions).

View-based Feature Generalization: With multiple feature views, we denote the variables outside of view v as $\mathbf{x}_{-v}$. If we assume that view-based features partially generalize from the training distribution to the testing distribution - only the input variables outside of view v generalize to testing distribution, the right hand side of the constraints for those generalized views take the form of an importance weighting of view v's feature vector based on the non-view input variables, $\mathbf{x}_{-v}$: $\mathbb{E}_{(\mathbf{X},Y)\sim\tilde{P}_{\text{train}}}\left[\frac{P_{\text{test}}(\mathbf{X}_{-v}|\mathbf{X}_v)}{P_{\text{train}}(\mathbf{X}_{-v}|\mathbf{X}_v)}\phi_v(\mathbf{X}_v, Y)\right]$. Note that we need prior knowledge about features to make these reasonable assumptions. We use these partially reweighted features to formulate a new predictor for classification under covariate shift in Def. 1. Applying the above assumptions in the generalized formulation, This view-based robust classifier leverages partial generalization of features and can be regarded as a balance between the loss reweighting of Eq. (1) and the robust prediction of Eq. (2).

Robust Multiview Reformulation: Leveraging the view-based feature generalizations, we reformulate the adversarial game with $|\mathcal{V}_g|$ generalized and $|\mathcal{V}_o|$ non-generalized views of features.

Definition 1. *The* **robust multiview covariate shift classifier** *is the equilibrium of:*

$$\min_{\hat{P}} \max_{\check{P}} \mathbb{E}_{\mathbf{X}\sim P_{test},\check{Y}|\mathbf{X}\sim\check{P}} \left[-\log \hat{P}(\check{Y}|\mathbf{X}) \right], \qquad \textit{such that: } \forall v \in \mathcal{V}_g,$$

$$\mathbb{E}_{\mathbf{X}\sim\tilde{P}_{train},\check{Y}|\mathbf{X}\sim\check{P}} \left[\frac{P_{test}(\mathbf{X}_{-v}|\mathbf{X}_v)}{P_{train}(\mathbf{X}_{-v}|\mathbf{X}_v)}\phi_v(\mathbf{X}_v,\check{Y}) \right] = \mathbb{E}_{(\mathbf{X},Y)\sim\tilde{P}_{train}} \left[\frac{P_{test}(\mathbf{X}_{-v}|\mathbf{X}_v)}{P_{train}(\mathbf{X}_{-v}|\mathbf{X}_v)}\phi_v(\mathbf{X}_v,Y) \right],$$

$$\forall v' \in \mathcal{V}_o, \qquad \mathbb{E}_{\mathbf{X}\sim\tilde{P}_{train},\check{Y}|\mathbf{X}\sim\check{P}} \left[\phi_{v'}(\mathbf{X}_{v'},\check{Y}) \right] = \mathbb{E}_{(\mathbf{X},Y)\sim\tilde{P}_{train}} \left[\phi_{v'}(\mathbf{X}_{v'},Y) \right].$$

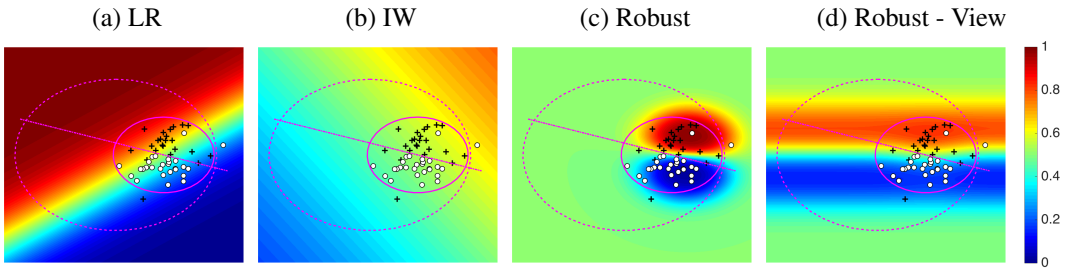
Theorem 1. *The robust multiview covariate shift classifier has the following parametric form:*

$$\hat{P}_\theta(y|\mathbf{x}) = e^{\sum_v \frac{P_{train}(\mathbf{x}_v)}{P_{test}(\mathbf{x}_v)}\theta_v \cdot \phi_v(\mathbf{x}_v, y) + \frac{P_{train}(\mathbf{x})}{P_{test}(\mathbf{x})} \sum_{v'} \theta_{v'} \cdot \phi_{v'}(\mathbf{x}_{v'}, y)} / Z(\mathbf{x}), \tag{3}$$

where view-specific density ratios, $P_{train}(\mathbf{x}_v)/P_{test}(\mathbf{x}_v)$ are applied on generalized views $\mathcal{V}_g$ and joint density ratios $P_{train}(\mathbf{x})/P_{test}(\mathbf{x})$ are applied on non-generalized views $\mathcal{V}_o$.

4 Synthetic Example

We consider a synthetic example with data sampled from two overlapping Gaussian distributions (X) and identical true decision boundary (Y). In 50 source and 100 target data points, 10% of the example are chosen uniformly at random (noise). We train four methods using source data points (shown in the figures, roughly within the smaller eclipses) and evaluate them on target data (not shown in the figures, roughly within larger eclipse). The colormap represents the testing conditional label distribution in the whole space. Logloss evaluated on the target data is listed below each figure.



logloss = 1.18 logloss = 0.916 logloss = 0.859 logloss = 0.729

Figure 1: Comparison of Logistic Regression(a), Importanct Weighting Logistic Regression(b), Robust Bias-Aware Prediction(c) and View-based Robust Bias-Aware Prediction(d).

We see from the figures that the true decision boundary (the tilted line) could not be recovered by any of the methods using the limited data points. In fact, this is why covariate shift problems are so challenging, even though the assumption $P_{\text{train}}(y|x) = P_{\text{test}}(y|x)$ holds. LR makes very certain predictions while IW, with reweighted source data, provides a less abrupt decision boundary but remains very certain towards the corners of the input space. The robust method restricts the certain prediction regions only to areas with enough source data to support the prediction. The rest of the target distribution space is covered with uniform predictions. It achieves better target logloss by being more conservative. But the question remains: could we leverage more information from the source data? We get our answer from the last model: our robust view-based model. In this two dimensional space, the view-based source and target feature distribution is much closer in the vertical dimension ($\mathbf{x}_2$) than in the horizontal dimension ($\mathbf{x}_1$). We can therefore assume the source vertical feature dimensions can generalize to target ones in our generalized robust covariate shift classifier. This corresponds with a parametric form as $\hat{P}_\theta(y|\mathbf{x}) \propto e^{\frac{P_{\text{train}}(\mathbf{x})}{P_{\text{test}}(\mathbf{x})}\theta_1 \cdot \phi_v(\mathbf{x}_1, y) + \frac{P_{\text{train}}(\mathbf{x}_2)}{P_{\text{test}}(\mathbf{x}_2)}\theta_2 \cdot \phi_v(\mathbf{x}_2, y)}$. It maintains uncertainty in areas with little data (the top and bottom area in the input space), but gives more meaningful predictions in areas where the method expects the data could provide reasonable extrapolations.

5 Experiments

We conduct experiments on real dataset and investigate the performance of our multiview approach in this section. We chose four datasets from the UCI repository [9, 11]. In order to create covariate shift, we synthetically generate 30 separate experiments in each dataset by drawing 100 source samples and 100 target data samples from it. Note that we normalize the data to the range $[0, 1]$ beforehand. In the UCI experiments, we regard each feature dimension as a specific view for simplification.

We evaluate the multiview robust covariate shift approach and three other methods: **Multiview robust bias aware classifier (Robust - View)** utilize the robust covariate shift classification framework applying multiview feature generalization assumptions as in Definition 1; **Robust bias aware**

classifier (Robust) adversarially minimizes the target distribution logloss, using the parametric form as Eq.2; **Logistic regression (LR)** maximizes the conditional log likelihood on training data, $\max_\theta \mathbb{E}_{P_{\text{train}}(x)P(y|x)} [\log P_\theta(Y|X)] - \lambda\|\theta\|_2$, where $\hat{P}_\theta(y|x) = \frac{\exp(\theta \cdot \Phi(x,y))}{\sum_{y' \in \mathcal{Y}} \exp(\theta \cdot \Phi(x,y'))}$, ignoring the covariate shift of the problem setting entirely; and **Importance weighting method (IW)** maximizes the conditional target data likelihood as estimated using importance weighting with the density ratio, $\max_\theta \mathbb{E}_{P_{\text{train}}(x)P(y|x)} \left[\frac{P_{\text{test}}(x)}{P_{\text{train}}(x)} (\log P_\theta(Y|X)) \right] - \lambda\|\theta\|_2$.

Density Estimator, Generalization Criterion and Model Selection: To estimate density ratios as $\frac{P_{\text{test}}(\mathbf{X}_{-v}|\mathbf{X}_v)}{P_{\text{train}}(\mathbf{X}_{-v}|\mathbf{X}_v)}$, we analyze this ratio according to Bayes' rule: $\frac{P_{\text{test}}(\mathbf{X}_{-v}|\mathbf{X}_v)}{P_{\text{train}}(\mathbf{X}_{-v}|\mathbf{X}_v)} = \frac{P(\text{test}|\mathbf{X})}{P(\text{train}|\mathbf{X})} \cdot \frac{P(\text{train}|\mathbf{X}_v)}{P(\text{test}|\mathbf{X}_v)}$. We construct two logistic regression models to estimate each ratio for $\mathbf{X}$ and $\mathbf{X}_v$ respectively. We use L_2 regularization for both logistic regression models in density estimation for both robust and importance weighted methods and set $\lambda = D_2[1 + (2+\sqrt{2})\sqrt{\ln(1/\sigma)})]/\sqrt{2m}$, where D_2 is the l_2 diameter of the feature space. Because the loss of the estimated conditional label distribution will be bounded as $L(\hat{\theta}) \leq L(\theta^*) + ||\theta^*||_2 D_2[\sqrt{2} + 2(1+\sqrt{2})\sqrt{\ln(1/\sigma)}]/\sqrt{m}$, with probability at least $1-\sigma$ [4]. Note that since we normalize features, $D_2 \leq 1$ and σ is set to 0.05. We use first order features for density estimation in our experiment, which is enough in most cases.

We evaluate the KL-divergence of training distribution $P_{\text{train}}(\mathbf{x}_v)$ and testing distribution $P_{\text{test}}(\mathbf{x}_v)$ after density estimation to determine whether we should assume the generalization of each view, i.e. $v \in \mathcal{V}_o$ or $v \in \mathcal{V}_g$. We use the threshold of 0.1, that is if $K < 0.1$, we think $P_{\text{train}}(\mathbf{x}_v)$ is similar enough with $P_{\text{test}}(\mathbf{x}_v)$ and $v \in \mathcal{V}_g$, otherwise, $v \in \mathcal{V}_o$. We include both training and testing inputs in the computation of KL-divergence. $K = \sum_{\mathbf{x}_v \in \mathbf{x}_{\text{train}}} P_{\text{train}}(\mathbf{x}_v) \log(P_{\text{train}}(\mathbf{x}_v)/P_{\text{test}}(\mathbf{x}_v)) + \sum_{\mathbf{x}_v \in \mathbf{x}_{\text{test}}} P_{\text{test}}(\mathbf{x}_v) \log(P_{\text{test}}(\mathbf{x}_v)/P_{\text{train}}(\mathbf{x}_v))$. In practice, we could rely on both data observation and expert knowledge to choose the generalization criterion.

For each method, the regularization parameter λ are chosen using 5-fold cross validation, or importance weighted cross validation (IWCV) on a parameter range $\lambda \in [2^{-16}, 2^{-12}, 2^{-8}, 2^{-4}, 1]$. Here the traditional cross validation is applied on LR, while IWCV is applied on all the other methods. Note that the traditional cross validation process is not correct anymore in the covariate shift setting where the source marginal data distribution of $P(X)$ is different from the target distribution [12].

Result Analysis: We compare logloss and accuracy of each method in Table 1. We denote significantly best result under paired t-test with significance level 0.05 in *bold*. If the best model cannot be differentiated by the paired t -test, all of them are *bolded*. We can see from the tables that Robust-View outperforms all other methods in most datasets. Even though the logloss of our method in `Seed` and accuracy in `Vehicle` is worse, they are still comparable with the other methods.

Table 1: Average Logloss and Accuracy Comparison

Dataset	Logloss				Accuracy			
	Robust-View	**Robust**	**LR**	**IW**	**Robust-View**	**Robust**	**LR**	**IW**
`Seed`	**1.039**	**1.105**	1.385	1.299	**0.734**	0.618	0.555	0.560
`Vertebral`	**0.577**	0.830	0.811	0.810	**0.860**	0.795	0.795	0.791
`Vehicle`	**1.68**	1.82	2.82	2.59	**0.498**	**0.441**	0.433	**0.465**
`Spam`	**0.853**	1.804	1.981	0.969	**0.680**	0.533	0.534	0.531

6 Conclusions and Future Work

Covariate shift classification is an important and difficult task for machine learning in non-stationary environments when the target labels are not available. We propose a multiview robust covariate shift classification framework that is flexible enough to make different feature generalization assumptions for multiview features and still preserve robustness. We use a synthetic example and UCI biased datasets to demonstrate the model performance of multiview covariate shift classification. In future work, we will investigate this framework in real multi-view covariate shift data and continue exploring theory and algorithms for making both robust and accurate predictions under covariate shift.

References

[1] Xiangli Chen, Mathew Monfort, Anqi Liu, and Brian D Ziebart. Robust covariate shift regression. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 1270–1279, 2016.

[2] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems*, pages 442–450, 2010.

[3] Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *Proceedings of the International Conference on Algorithmic Learning Theory*, pages 38–53, 2008.

[4] Miroslav Dudík and Robert E. Schapire. Maximum entropy distribution estimation with generalized regularization. In *Learning Theory*, pages 123–138. Springer Berlin Heidelberg, 2006.

[5] Wei Fan, Ian Davidson, Bianca Zadrozny, and Philip S. Yu. An improved categorization of classifier's sensitivity on sample selection bias. In *Proc. of the IEEE International Conference on Data Mining*, pages 605–608, 2005.

[6] Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Schölkopf. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, pages 601–608, 2006.

[7] Anqi Liu and Brian D. Ziebart. Robust classification under sample selection bias. In *Advances in Neural Information Processing Systems*, pages 37–45, 2014.

[8] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

[9] D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz. UCI repository of machine learning databases, 1998.

[10] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[11] J P Siebert. Vehicle recognition using rule based methods. Technical report, Mar 1987.

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

[13] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul V. Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems*, pages 1433–1440, 2008.

[14] Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the International Conference on Machine Learning*, pages 903–910. ACM, 2004.