Causal and Anti-causal Structure for Semi-supervised Learning

Yu Zhu
University of California, Santa Cruz

Abstract

Different machine learning methods can benefit from different casual structures. One of the methods is Semi-supervised Learning (SSL) where the models are trained with a large amount of unlabeled sample. Schölkopf et al.(2012) tried to prove an hypothesis that SSL will work under the anti-causal structure rather than the causal structure based on the empirical analysis. As for a possible extension, Schölkopf et al.(2012) considered the feature set with both cause and effect features, and proposed two algorithms exploring how to extract information from the conditional distribution from the unlabeled sample to improve the prediction under the SSL framework. We performed the simulation study to compare the proposed algorithms with the supervised baseline model as well as the coventional SSL methods.

1 Introduction

Causal structures can influence the machine learning (ML) results under different variable dependence scenarios, leading to applications in Covariate Shifts, Selection Bias, Transfer Learning, Semi-Supervised Learning (SSL), etc. Schölkopf et al.(2012) tried to explore how can different causal structures benefit the ML methods. Specifically, consider the input X and output Y, causal prediction indicates the causal direction of predicting effect Y from cause X, while anticausal prediction presents the inverse relationship that predicting cause Y from effect X. This paper (Schölkopf et al., 2012) mainly discussed the SSL case under the different causal structures: causal, anticausal/confounded and unclear, comparing with the supervised baseline methods.

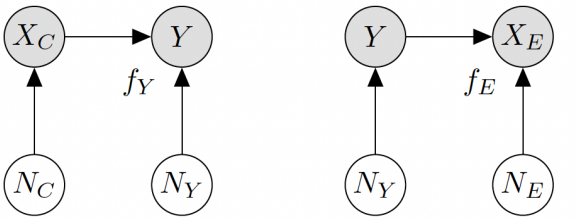
Intuitively, we expect that SSL would provide better predictions under anticausal structure since $P(X)$ and $P(Y|X)$ are dependent. Thus, additional information from $P(X)$ may lead to a more accurate inference of $P(Y|X)$. As for the causal structure, SSL may not have obvious influence on the prediction under the assumption of independence of mechanism and input, $P(X)$ and $P(Y|X)$ are independent. Under the empirical analysis, the results confirmed the hypothesis that SSL can help with anticausal structure.

Kügelgen et al. (2020) considered a possible extension of classification under the SSL framework with features containing both of the cause and effect (X_C and X_E). It can be seen as a generalization form of the previous study from Schölkopf et al.(2012) using the information of $P(X_E|X_C)$. For instance, we need to predict a disease from both risk factors and symptoms. We apply the two casually-motivated methods that proposed based on the information extraction from the conditional distribution $X_E|X_C$ from the unlabeled sample rather than the joint feature with simulation study in different settings. We want to explore whether the SSL still work under this causal structure, and also whether the exploit link between the conditional distribution is better than the joint distirbution.

2 SSL under Causal and Anti-causal Mechanisms

Under the SSL framework, the training data typically contains a small size of the labeled sample $(X^l, Y^l) = \{(x^i, y^i)\}_{i=1}^{n_l}$ as well as a large size of the unlabeled sample $X^u = \{x^i\}_{i=n_l+1}^{n_l+n_u}$. And we assume the labels are missing completely at random (MCAR). Our goal is to estimate $P(Y|X)$. Under supervised learning, we only train the model with the labeled sample. But if we still only use the labeled sample when we also have a large amount of unlabeled data, the information contained will be wasted and may be influential to the results. So the question becomes whether the additional unlabeled $\{X^u\}$ can improve estimating $P(Y|X)$. We know that $\{X^u\}$ can improve the estimation of $P(X)$, so if we have a link $P(X)$ and $P(Y|X)$, $P(Y|X)$ can be improved.

A lot of papers try to make additional assumptions to establish such link(Chapelle et al., 2010; Mey and Loog, 2019; Zhu, 2005). For example, the cluster assumption and the low-density assumption. But now we consider this problem under the causal relationship, and based on the Independent Causal Mechanisms (ICM) principle, $P(X_C)$ and $P(Y|X_C)$ are algorithmically independent. So intuitively, SSL may work under the anti-causal structure since $P(X_E)$ may contain information about $P(Y|X_E)$, while SSL will not work under the causal structure since $P(X_C)$ contains no information about $P(Y|X_C)$.

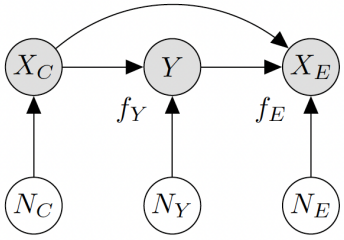


(a) Causal learning (b) Anticausal learning

3 SSL with Cause and Effect Features

It is quite common in real-world data that contains both cause features and effect features simultaneously. For example, in some medical data set, we aim to predict whether the patients have a specific disease. And the features like gender, age and BMI can be treated as causes, and symptoms like vomiting can be considered as the effect feature. Kügelgen et al. (2020) considered this more complicated classification problem under the SSL framework with features containing both the cause and effect (X_C and X_E). Kügelgen et al. (2020) mainly explored 2 questions: 1. Whether this causal structure can also lead to the success of the SSL, and 2. How to better exploit the information contained in the unlabeled samples. The causal model is shown in Figure 3 and can be presented in the form below:

$$X_C := N_C \tag{1}$$
$$Y := f_Y(X_C, N_Y) \tag{2}$$
$$X_E := f_E(Y, X_C, N_E) \tag{3}$$



Our goal is the estimation of $P(Y|X_C, X_E)$ with additional information of $P(X_C, X_E)$ from the unlabeled sample. If we refer back to the discussion of the SSL under causal and anti-causal mechanisms, subject to the Independent Causal Mechanisms (ICM) principle, $P(X_C)$ and $P(Y|X_C)$ are

algorithmically independent, and $P(X_C)$ contains no information about $P(Y|X_C)$ or $P(X_E|Y, X_C)$. And since we have $P(Y|X_C, X_E) \propto P(Y|X_C)P(X_E|Y, X_C)$, intuitively $P(X_C)$ will not be able to contribute to the estimation of $P(Y|X_C, X_E)$, and $P(X_E|X_C)$ contains all the information provided from the unlabeled sample. So instead of directly using the joint distribution $P(X_C, X_E)$, this paper is interested in extracting information from the conditional distribution $P(X_E|X_C)$. This assumption shows us a direction for solving question 2.

For this project, we only focus on the binary classification problem with the prediction of the unlabeled training sample.

$$Y := 1\{g(X_C) > U\} \tag{4}$$

$$X_E := Yf_1(X_C, N_E) + (1 - Y)f_0(X_C, N_E) \tag{5}$$

Kügelgen et al. (2020) modified the standard SSL assumptions discussed in the previous section such that they make advantage of potentially shared knowledge between $P(X_E|X_C)$ and $P(Y|X_C, X_E)$. The first refined assumption is called the conditional cluster assumption. We assume the points in the same cluster of $X_E|X_C$ share the same label Y, instead of (X_E, X_C). This assumption links the class labels to membership in clusters of $X_E|X_C$. So we tend to classify the unlabelled data based on whether X_E is better explained by $f_0(X_C)$ or $f_1(X_C)$. In this case, if we still apply the Support Vector Machine algorithm that tries to label the data based on the maximum margin, it can be easily confused and provides opposite predictions.

Another reformulated assumption is the low-conditional-density separation assumption. Originally we make assumption based on the joint feature set. But now we adjusted it to assume that class boundaries of $P(Y|X_C, X_E)$ should lie in regions where $P(X_E|X_C)$ is small

4 Algorithms

4.0.1 Semi-generative Model

We only model the informative part of the generative process: $P(Y, X_E|X_C)$. And it is called the semi-generative because it's not a completely generative model $P(Y, X_E, X_C)$ and not a discriminative model $P(Y|X_E, X_C)$.

$$argmax_\theta p(y^l, X_E^l|X_C^l; \theta) \sum_{y^u} p(y^u, X_E^u|X_C^u; \theta) \tag{6}$$

where $\theta = (\theta_Y, \theta_E)$
And we can minimize the negative log-likelihood (NLL), which decomposes for fixed labels into two distinct terms optimized independently for θ_Y and θ_E.

$$NLL(\theta|X_C, Y, X_E) := -logp(Y, X_E|X_C; \theta) \tag{7}$$

$$= -logp(Y|X_C; \theta_Y) - logp(X_E|Y, X_C; \theta_E) \tag{8}$$

we first initialize the parameter θ_Y and θ_E based on the local optimum of the NLL of the two parametric models $P(Y|X_C)$ and $p(X_E|Y, X_C)$. And we iteratively compute the expected label given the current parameters, and then minimising the NLL w.r.t. to the parameters keeping the labels fixed until the algorithm is converged. Here we define the convergence based on whether the current iteration of Y equals to the previous iteration.

4.0.2 Conditional Self-learning

We extract information from $P(X_E|X_C)$ instead of propagating labels based on similarities between points in the joint feature space (X_C, X_E). And we assume an additive noise model:

$$f_i(X_C, N_E) = f_i(X_C) + N_{E,i}, i = 0, 1 \tag{9}$$

To ensure the one-to-one function from X_C to X_E for each label, we also need to make the assumption for the noise to be zero mean and unimodal.

We first initialise $\hat{f}_0$ and $\hat{f}_1$ from labeled data by regressing X_E on X_C. Then, we use the fitted f_0 and f_1 to predict the unlabeled sample. The class corresponds to the smallest prediction error will be applied. and we use it to update the corresponding $\hat{f}_i$. We keep repeat this procedure until all initially unlabeled points are labeled.

Algorithm 1: EM-like algorithm for fitting a semi-generative model by maximum likelihood

Input: labelled data $(\mathbf{X}_C^l, \mathbf{y}^l, \mathbf{X}_E^l)$; unlabelled data $(\mathbf{X}_C^u, \mathbf{X}_E^u)$; parametric models $p(y|\mathbf{x}_C; \boldsymbol{\theta}_Y)$ and $p(\mathbf{x}_E|\mathbf{x}_C, y; \boldsymbol{\theta}_E)$

Output: fitted labels $\mathbf{y}^u$; estimates $\hat{\boldsymbol{\theta}}_Y, \hat{\boldsymbol{\theta}}_E$

$t \leftarrow 0$
$\hat{\boldsymbol{\theta}}_Y^{(0)} \leftarrow \arg\min \text{NLL}(\boldsymbol{\theta}_Y|\mathbf{y}^l)$
$\hat{\boldsymbol{\theta}}_E^{(0)} \leftarrow \arg\min \text{NLL}(\boldsymbol{\theta}_E|\mathbf{y}^l)$
while *not converged* **do**
  $\mathbf{y}^{(t)} \leftarrow \mathbb{I}\{p(\mathbf{y}|\mathbf{X}_C^u, \mathbf{X}_E^u; \boldsymbol{\theta}_Y^{(t)}, \boldsymbol{\theta}_E^{(t)}) > 0.5\}$
  $\hat{\boldsymbol{\theta}}_Y^{(t+1)} \leftarrow \arg\min \text{NLL}(\boldsymbol{\theta}_Y|\mathbf{y}^l, \mathbf{y}^{(t)})$
  $\hat{\boldsymbol{\theta}}_E^{(t+1)} \leftarrow \arg\min \text{NLL}(\boldsymbol{\theta}_E|\mathbf{y}^l, \mathbf{y}^{(t)})$
  $t \leftarrow t + 1$
end
return $\mathbf{y}^{(t-1)}$, $\boldsymbol{\theta}_Y^{(t)}$, $\boldsymbol{\theta}_E^{(t)}$

Algorithm 2: Conditional self-learning

Input: labelled data $(\mathbf{X}_C^l, \mathbf{y}^l, \mathbf{X}_E^l)$; unlabelled data; $(\mathbf{X}_C^u, \mathbf{X}_E^u)$; `regress()` method

Output: fitted labels $\mathbf{y}^u$; functions $\hat{f}_0$, $\hat{f}_1$

$t \leftarrow 0$
while *unlabelled data left* **do**
  for $i = 0, 1$ **do**
    $\hat{f}_i^{(t)} \leftarrow \texttt{regress}(\mathbf{X}_{E,i}^l, \mathbf{X}_{C,i}^l)$
    $\mathbf{r}_i \leftarrow ||\mathbf{X}_E^u - \hat{f}_i^{(t)}(\mathbf{X}_C^u)||^2$
  end
  $(i, j) \leftarrow \arg\min\{\mathbf{r}_{i,j} : i = 0, 1; j = 1, ..., n_u\}$
  $y^{n_l+j} \leftarrow i$
  $\mathbf{X}_{E,i}^l, \mathbf{X}_{C,i}^l \leftarrow \texttt{append}(\mathbf{x}_E^{n_l+j}, \mathbf{x}_C^{n_l+j})$
  $t \leftarrow t + 1$
end
return $\mathbf{y}^u$, $\hat{f}_0^{(t-1)}$, $\hat{f}_1^{(t-1)}$

5 Simulation Study

We perform the simulation study to further compare the performance of the semi-generative model and conditional self-learning with the following baseline methods:

- Supervised Logistic Regression (SLR): only use the labeled data and ignore the causal structure.
- Transductive Support Vector Machine (T-SVM) with linear and RBF kernels: conventional SSL methods.

Based on the form of the causal model, We draw $X_C \in \mathbb{R}^{d_C}$ from a mixture of m d_C-dimensional Gaussian distribution. Y is generated given X_C following a Bernoulli distribution with a probability $\sigma(a'X_C + b)$. And X_E is generated based on the class-dependent additive noise model given X_C.

$$Y := 1\{\sigma(a'X_C + b) > N_Y\} \tag{10}$$
$$X_E := \{f_i(X_C) + D_i N_E\} 1\{Y = i\} \tag{11}$$

with $N_Y \sim U[0,1]$, $N_E \sim N_{d_E}(0, I)$, $a \in \mathbb{R}^{d_C}$, $D_i \in \mathbb{R}^{d_E \times d_E}$, $\sigma(x)$ is the sigmoid function, $i = 0, 1$.

We introduce different types of additive noise models: linear or non-linear, and under low/high dimensional feature space. We decide to consider the following three cases:

- Case I: Linear additive noise model with one-dimensional feature space
- Case II: Linear additive noise model with high dimensional feature space
- Case III: Non-linear additive noise model with high dimensional feature space

We generate the data in each case with 10 labeled data and different numbers of unlabeled data: 20, 200, 1000 and 2000. Thus we can explore the influence of the missing proportion on the models. And we repeat the data generation process for 100 iterations. For each iteration, we compute the accuracy of each model. In the end, we compare the mean and the standard deviation of the accuracy among the models.

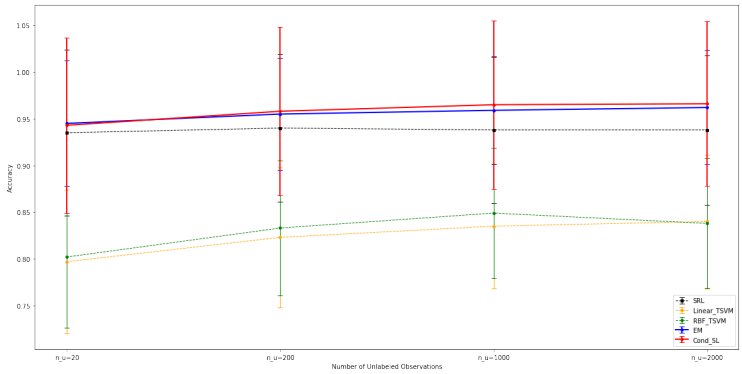
5.0.1 Case I

For Case I, the linear form of additive noise model can be presented as

$$f_i(X_C) = A_i{}' X_C + b_i \tag{12}$$

As for the one dimensional feature, set up $d_C = d_E = 1$. X_C is draw from a 3 mixture of normal distribution. Due to the low dimensional structure as well as the linear form of additive noise model, we use the linear regression for $P(X_E|X_C, y; \theta_E)$ in the semi-generative model and $P(X_E^l|X_C^l)$ in the conditional self-learning. The parameter settings are as below:

- feature dimensions: $d_C = d_E = 1$
- $X_C \sim \sum_{m=1}^{3} \omega_m N(\mu_{C_m}, \sigma^2_{C_m})$: $\omega = [0.3, 0.4, 0.3]$, $\mu_C = [-5, 0, 5]$, $\sigma_C = [0.5, 0.5, 0.5]$
- Y: $a = 0.5, b = 0$
- X_E: $A_0 = A_1 = 1, b_0 = -b_1 = 2, D_0 = D_1 = 0.25$

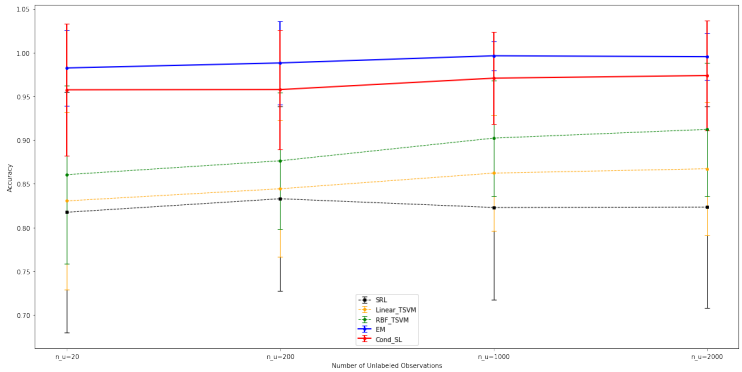


The two causally-motivated methods outperform the supervised logistic regression and two T-SVM methods. And there's no significant difference between them. And Two T-SVM methods perform even worse than the supervised logistic regression. It's under expectation because if we consider the linear additive noise model, conditional cluster assumption consider to classify X_E based on whether it's better explained by $f_0(X_C)$ or $f_1(X_C)$. However, conventional SSL methods such as SVM try to do the classification based on the maximum margin, which might lead to around 50% error.

5.0.2 Case II

For Case II, we apply the same linear form of additive noise model as Case I. And we set up $d_C = d_E = 10$ for the high dimensional feature setting. We draw X_C from a 2 mixture of multivariate normal distribution. To better fit the high-dimensional features, we apply the ridge regression with default penalty $\lambda = 1$ for $P(X_E|X_C, y; \theta_E)$ in the semi-generative model and $P(X_E^l|X_C^l)$ in the conditional self-learning. The parameter settings are as below:

- feature dimensions: $d_C = d_E = 10$
- $X_C \sim \sum_{m=1}^{2} \omega_m MVN_{d_C}(\mu_{C_m}, \Sigma_{C_m})$: $\omega = [0.5, 0.5]$, $\mu_{C_1} = [-\frac{1}{2}, -\frac{2}{2}, ..., -\frac{d_C}{2}]$, $\mu_{C_2} = [\frac{1}{2}, \frac{2}{2}, ..., \frac{d_C}{2}]$, $\Sigma_{C_1} = \Sigma_{C_2} = diag_{d_C}(0.5)$
- Y: $a = 0.5, b = 0$
- X_E: $A_0 = -A_1 = 0.5J_{10}$, $b_0 = b_1 = 0$, $D_0 = D_1 = diag_{d_E}(0.25)$



Under this high-dimensional case, two causally-motivated methods still outperform the other three. Most of the SSL methods perform better than the low-dimensional Case I, and overwhelm the supervised logistic regression.

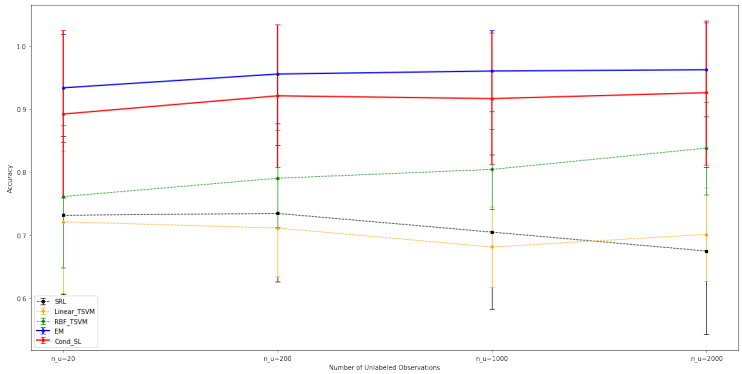
5.0.3 Case III

For Case III, we establish the non-linear form of additive noise model as

$$f_0(X_C) = A_0{}'X_C + sin(X_C) + b_0 \tag{13}$$
$$f_1(X_C) = A_1{}'X_C + cos(X_C) + b_1 \tag{14}$$

The rest of the settings are exact same as Case II. Consider the non-linearity of the additive noise model, we apply the ridge regression with default penalty $\lambda = 1$ for $P(X_E|X_C, y; \theta_E)$ in the semi-generative model and the kernel ridge regression with penalty $\lambda = 1$ for $P(X_E^l|X_C^l)$ in the conditional self-learning. The parameter settings are as below:

- feature dimensions: $d_C = d_E = 10$
- $X_C \sim \sum_{m=1}^{2} \omega_m MVN_{d_C}(\mu_{C_m}, \Sigma_{C_m})$: $\omega = [0.5, 0.5]$, $\mu_{C_1} = [-\frac{1}{2}, -\frac{2}{2}, ..., -\frac{d_C}{2}]$, $\mu_{C_2} = [\frac{1}{2}, \frac{2}{2}, ..., \frac{d_C}{2}]$, $\Sigma_{C_1} = \Sigma_{C_2} = diag_{d_C}(0.5)$
- Y: $a = 0.5, b = 0$
- X_E: $A_0 = -A_1 = 0.5J_{10}$, $b_0 = b_1 = 0$, $D_0 = D_1 = diag_{d_E}(0.25)$



It seems that under the non-linear additive noise model, all the methods perform worse compared with Case II. EM-like approach still provides stable performance. And T-SVM with RBF kernel can work better compared with linear T-SVM and supervised logistic regression.

Overall, the semi-generative model and the conditional self-learning outperform the supervised logistic regression baseline as well as the other SSL T-SVM methods. Specifically, the semi-generative model shows the stable performance under all the scenarios, and if we apply with some other regression functions and tune the parameters, both of the semi-generative model and the conditional self-learning may show even higher prediction accuracy. Besides, the influence to the number of unlabeled sample is not as significant as expected. As the number of unlabeled sample increase from 20 to 200 and then to 1000, most of the SSL methods tends to show the increase in the prediction accuracy. So under this causal structure that contains both cause and effect features, more additional unlabeled sample can indeed improve the performance of the SSL methods to some degree.

6 Conclusions

Based on the simulation study, we can find out that the exploration of the conditional distribution $X_E|X_C$ instead of the joint distribution of (X_E, X_C) can help improve the classification performance. Besides, the increase of unlabeled data can slightly contribute to higher accuracy for SSL methods (T-SVM and proposed methods) with both cause and effect features. And the proposed two models are also quite flexible. If we need to apply them to the unstructured data such as natural images or text, we can use GANs or VAEs to model the additive noise functions. This could also be connected to the domain adaptation: the proposed approaches are robust to changes in $P(X_C)$. $P(Y|X_C)$ and $P(X_E, Y|X_C)$ remain stable. We can further explore it under the domain adaptation framework.

7 References

Schölkopf, Bernhard, et al. "On causal and anticausal learning." arXiv preprint arXiv:1206.6471 (2012).

von Kügelgen, J., Mey, A., Loog, M., Schölkopf, B. (2019). Semi-Supervised Learning, Causality and the Conditional Cluster Assumption. arXiv. https://doi.org/10.48550/arXiv.1905.12081

Hoyer, P. O., D. Janzing, J. M. Mooij, J. Peters, and B. Sch¨olkopf (2009). "Nonlinear causal discovery with additive noise models". In: Advances in Neural Information Processing Systems 21, pp. 689– 696.