Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)

Controlling for Unobserved Confounds in Classification Using Correlational Constraints

Virgile Landeiro, Aron Culotta
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616
vlandeir@hawk.iit.edu, aculotta@iit.edu

Abstract

As statistical classifiers become integrated into real-world applications, it is important to consider not only their accuracy but also their robustness to changes in the data distribution. In this paper, we consider the case where there is an unobserved confounding variable z that influences both the features $\mathbf{x}$ and the class variable y. When the influence of z changes from training to testing data, we find that the classifier accuracy can degrade rapidly. In our approach, we assume that we can predict the value of z at training time with some error. The prediction for z is then fed to Pearl's back-door adjustment to build our model. Because of the attenuation bias caused by measurement error in z, standard approaches to controlling for z are ineffective. In response, we propose a method to properly control for the influence of z by first estimating its relationship with the class variable y, then updating predictions for z to match that estimated relationship. By adjusting the influence of z, we show that we can build a model that exceeds competing baselines on accuracy as well as on robustness over a range of confounding relationships.

1 Introduction[1]

Statistical classifiers have become widely used to inform important decisions such as whether to approve a loan (Hand and Henley 1997), hire a job candidate (Miller 2015), or release a criminal defendant on bond (Monahan and Skeem 2016). Given the consequences of such decisions, it is critical that we can remove sources of systematic bias in classification algorithms. One important type of classifier bias arises from *confounding variables*. A confounder z is a variable that is correlated both with the input variables (or features) $\mathbf{x}$ and the target variable (or label) y of a classifier. When z is not included in the model, the true relationship between $\mathbf{x}$ and y can be improperly estimated; in the social sciences – and originally in econometrics – this is called omitted variable bias (King, Keohane, and Verba 1994).

Confounding variables can be particularly problematic in high-dimensional settings, such as text classification, where models may contain thousands or millions of parameters, making manual inspection of models impractical. The common use of text classification to derive variables in computational social science applications (Lazer et al. 2009) further adds to the urgency of the problem. In recent work (Landeiro and Culotta 2016), we proposed a text classification algorithm that used Pearl's back-door adjustment (Pearl 2003) to control for an *observed* confounding variable. It was found that this approach produces classifiers that are significantly more robust to shifts in the relationship between confounder z and class label y from training to testing data. However, an important limitation of this prior work is that it assumes that a training set is available in which every instance is annotated for both class label y and confounder z. This is problematic because there are many confounders we may want to control for (e.g., income, age, gender, race/ethnicity) that are often rarely available and difficult for humans to label, particularly in addition to the primary label y. A natural solution is to build statistical classifiers for confounders z, and use the predicted values of z to control for these confounders. However, the measurement error of z introduces *attenuation bias* in the back-door adjustment, resulting in classifiers that are still confounded by z.

In this paper, we present a classification algorithm based on Pearl's back-door adjustment to control for an *unobserved* confounding variable. Our approach assumes we have a preliminary classifier that can predict the value of the confounder z, and that we have an estimate of the error rate of this z-classifier. We offer two methods to adjust for the mislabeled z to improve the effectiveness of back-door adjustment. A straightforward approach is to remove training instances for which the confidence of the predicted label for z is too low. While we do find this approach can reduce attenuation bias, it must discard many training examples, degrading the y-classifier. Our second approach instead uses the error rate of the z-classifier to estimate the correlation between y and z in the training set. The assignment to z is then optimized to match this estimated correlation, while also maximizing classification accuracy. We compare our methods on a real-world text classification task: predicting the location of a Twitter user, based on their tweets and confounded by gender. The resulting model exhibits significant improvements in both accuracy and robustness, with some settings producing similar results as fully-observed back-door adjustment.

[1]For an expanded version of this paper, replication code, and data, please see http://arxiv.org/abs/1703.01671 and https://github.com/tapilab/icwsm-2017-confounds

2 Methods

In this section, we first review prior work using back-door adjustment to control for observed confounders in text classification. We then introduce two methods for applying back-door adjustments when the confounder is unobserved at training time and must instead be predicted by a separate classifier.

Adjusting for *observed* confounders

Suppose we wish to estimate the causal effect of a variable x on a variable y when a randomized controlled trial is not possible. If a sufficient set of confounding variables z is available, one can use the well-studied back-door adjustment equation (Pearl 2003) as follows: $p(y|do(x)) = \sum_z p(y|x,z) \times p(z)$. Notice $p(y|x) \neq p(y|\text{do}(x))$: this donotation is used in causal inference to indicate that an intervention has been made on x.

Text classifiers estimate the distribution $p(y|\mathbf{x})$, the probability of a class label y given a term vector $\mathbf{x}$, from labeled training data. To apply back-door adjustment to text classification, we assume that there is a confounder z that influences both the term vector $\mathbf{x}$ through $p(\mathbf{x}|z)$ as well as the target label through $p(y|z)$. Of course, the goal of text classification is not causal inference (a term vector $\mathbf{x}$ does not "cause" label y). However, we have found that by controlling for a confounder z, the resulting classifier is robust to cases in which the relationship between z and y changes between the training and testing data (Landeiro and Culotta 2016).

The approach works as follows: Given a typical training set $D = \{(\mathbf{x}_i, y_i)\}$, we augment the training set by including z as a feature for each instance: $D' = \{(\mathbf{x}_i, y_i, z_i)\}$. We then fit a classifier on D', resulting in $p(y|\mathbf{x}, z)$, and also estimate $p(z)$ by simply computing the observed frequencies of z in D'. At testing time, we apply the back-door adjustment equation above to classify new examples, which requires summing over z for each instance. By controlling for the effect of z, the resulting classifier is robust to the case where $p(y|z)$ changes from training to testing data.

In the experiments below, we consider the problem of predicting a user's location y based on the text of their tweets $\mathbf{x}$, confounded by the user's gender z. That is, in the training data, there exists a correlation between gender and location, but we want the classifier to ignore that correlation. When applying back-door adjustment to a logistic regression classifier, the result is that the magnitudes of coefficients for terms that correlate with gender are greatly reduced, thereby minimizing the effect of gender on the predictions.

Adjusting for *unobserved* confounders

In the previous approach, it was assumed that we had access to a training set $D = \{(\mathbf{x}_i, y_i, z_i)\}$; that is, each instance is annotated both for the label y and confounder z. This is a burdensome assumption, given that ultimately we will need to control for many possible confounders (e.g., gender, race/ethnicity, age, etc.). Because many of these confounders are unobserved and/or difficult to obtain, it is necessary to develop adjustment methods that can handle noise in the assignment to z in the training data.

Our proposed method assumes we have an (imperfect) classifier for z, trained on a secondary training set $D_z = \{(\mathbf{x}_i, z_i)\}$ — we call this the *preliminary study*, with the resulting *preliminary classifier* $p(z|\mathbf{x})$. This is combined with the dataset $D_y = \{(\mathbf{x}_i, y_i)\}$, used to train the primary classifier $p(y|\mathbf{x})$. The advantage of allowing for separate training sets D_y and D_z is that it is often easier to annotate z variables for some users than others; for example, Pennacchiotti and Popescu (2011) build training data for ethnicity classification by searching for online users that explicitly state their ethnicity in their user profiles. After training on D_z, the preliminary classifier is applied to D_y to augment it with predicted annotations for confounder z: $D = \{(\mathbf{x}_i, y_i, z'_i)\}_{i=1}^n$, where z'_i denotes the predicted value of z_i. A tempting approach is to simply apply back-door adjustment as usual to this dataset, ignoring the noise introduced by z'. However, the resulting classifier will no longer properly control for the confounder z because (1) the observed correlation between y and z' in the training data will underestimate the actual correlation, yielding reduced coefficients for the z features (*attenuation bias*), and therefore reducing the adjustment power of back-door adjustment; and (2) some training instances have mislabeled annotations for z, making it more difficult to detect which features in $\mathbf{x}$ correlate with z, thereby preventing back-door adjustment from reducing those coefficients. In the following two sections, we propose two methods to fix these issues.

Thresholding on confidence of z predictions Our first approach is fairly simple; its objective is to directly reduce the number of mislabeled annotations in z'. Our preliminary model produces the value z'_i (the prediction of the true confounder z_i) as well as $p(z_i = z'_i|\mathbf{x}_i)$ (the confidence of the prediction; i.e., the posterior distribution over z). We use these posteriors to remove predictions with low confidence. By setting a threshold $\epsilon \in [0.5, 1]$, we filter the original dataset $D = \{\mathbf{x}_i, y_i, z'_i\}$ by keeping an instance i only if it satisfies $p(z_i = z'_i|\mathbf{x}_i) \geq \epsilon$. With this smaller set of training instances, we run back-door adjustment without modification. However, one important drawback of this method is that we remove instances from our training dataset.

Correlation matching The above approach aims to reduce errors in z', and as a side effect improves the estimate of $r(y, z)$, the correlation between y and z. In this section we propose an approach that directly tries to improve the estimate of $r(y, z)$ while also reducing errors in z. Let $r' = r(y, z')$ be the observed correlation between y and z', and let $r = r(y, z)$ be the true (unobservable) correlation between y and z in the training data for y, $D = \{\mathbf{x}_i, y_i, z'_i\}$. Our proposed approach builds on the insight of Francis, Coats, and Gibson (1999), who show that r' can be estimated from r using the variances of y and z as well as the variances of the errors in y and z:

$$r' = r\sqrt{\frac{1}{(1+\frac{V_{ey}}{V_y})(1+\frac{V_{ez}}{V_z})}} \Rightarrow r \quad = r' \times \sqrt{1+\frac{V_{ez}}{V_z}} \tag{1}$$

where V_z is the variance of z, and V_{ez} is the variance of error on z, and analogously for V_y, V_{ey}. Since in our setting y

is observed, we can set $V_{ey} = 0$ and solve for r. Thus, the factor by which r' underestimates r is proportional to the ratio of the variance of the error in z to the variance of z. We can estimate the terms V_z and V_{ez} using cross-validation on the preliminary training data $D_z = \{(\mathbf{x}_i, z_i)\}$. Plugging these estimates into Equation 1 enables us to estimate the true correlation between y and z in the target training data D. We will refer to this estimated correlation as $\hat{r}$. Now, let $\mathbf{Z}$ be the set of all possible assignments to z in the training set D (i.e., if z is a binary variable and $|D| = n$, then $|\mathbf{Z}| = 2^n$). Let $\mathbf{z}^j = \{z_1^j \dots z_n^j\} \in \mathbf{Z}$ be a vector of assignments to z, and let $r'(\mathbf{z}^j)$ indicate the correlation $r(\mathbf{z}^j, y)$. Then our objective is to choose an assignment from $\mathbf{Z}$ to minimize $r'(\mathbf{z}^j) - \hat{r}$, while still maximizing the probability of that assignment according to the preliminary classifier for z. We can write this objective as follows:

$$\mathbf{z}^* \leftarrow \arg\max_{\mathbf{z}^j \in \mathbf{Z}} \left(\frac{1}{n} \sum_{z_i^j \in \mathbf{z}^j} p(z_i = z_i^j | \mathbf{x}_i) \right) - |\hat{r} - r'(\mathbf{z}^j)| \quad (2)$$

Thus, we search for an optimal assignment $\mathbf{z}^*$ that maximizes the average posterior of the predicted z value, while minimizing the difference between the estimated correlation $\hat{r}$ and the observed correlation $r'(\mathbf{z}^j)$. This optimization problem can be approached in several ways. We implement a greedy hill-climbing algorithm that iterates through the values in z' sorted by confidence and flips the value if it reduces $|r - r'|$. The advantage of this approach is that it not only produces assignments to z that better align with the expected correlation $\hat{r}$, but it also results in more accurate assignments to z. The latter is possible because we are using prior knowledge about the relationship between z and y to assign values of z when the classifier is uncertain. As with the thresholding approach of the previous section, once the new assignments to z are found, back-door adjustment is run without modification.

3 Experiments

We conducted text classification experiments in which the relationship between the confounder z and the class variable y varies between the training and testing set. We consider the scenario in which we directly control the discrepancy between training and testing. Thus, we can determine how well a confounder has been controlled for by measuring how robust the method performs across a range of discrepancy levels. We denote $r_{train}(y, z)$ (respectively $r_{test}(y, z)$) as the correlation between y and z in the training set (resp. testing set). We also denote $\delta_{yz} = r_{train}(y, z) - r_{test}(y, z)$, the discrepancy between the training and test set.

4 Results

To validate our approach, we use the dataset from Landeiro and Culotta (2016), where the task is to predict the location of a Twitter user given her tweets confounded by gender. In the expanded version of this paper, we also experiment on a new dataset in which the task is to predict whether a Twitter user smokes based on their tweets, again confounded by gender. This second dataset yields similar core results in our experiments. Below, $F1_z$ denotes the F1 score for the z classifier on the preliminary study, and $F1_y$ denotes the F1 score for the y classifier on the primary study.

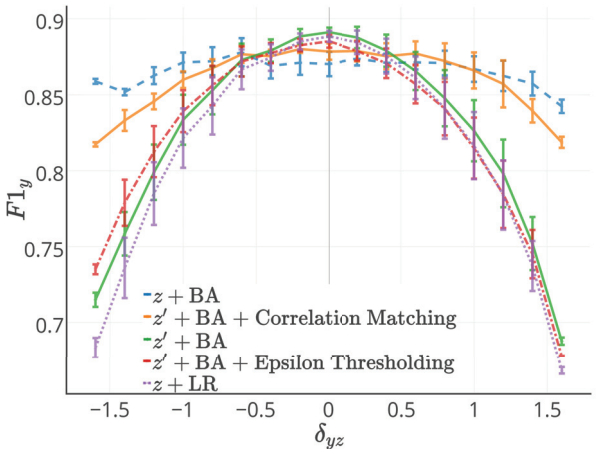


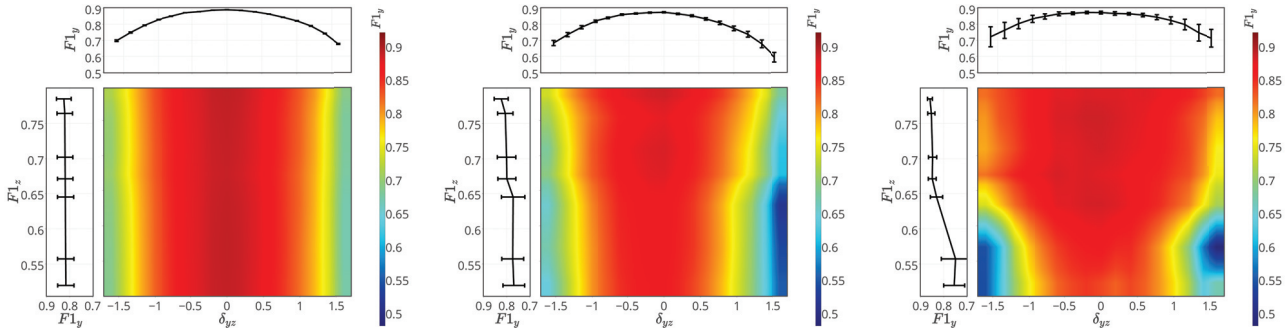
Figure 1: Effect of controlling for confounders on classifier robustness.

Effects of correlation adjustments on $F1_z$

When using ϵ **thresholding**, $r_\epsilon(y, z)$ approaches the true $r(y, z)$ as ϵ increases, leading to improved performance on our external study. However, it takes a high value of ϵ to get a correct approximation of the true association between y and z, meaning that we need to discard many training examples from our preliminary study to approximate $r(y, z)$. Using **correlation matching method**, we obtain similar or better results for $F1_z$, but we do not discard instances from the dataset.

Effects of correlation adjustments on $F1_y$

Fixed $\mathbf{F1_z = 0.784}$: As our primary result, we report the $F1_y$ obtained by different correlation adjustment methods across a range of shifts in the discrepancy between training and testing. For the Twitter dataset, the best performance we get in the preliminary study is $F1_z = 0.784$. We then compare testing $F1_y$ as $r_{train}(y, z)$ and $r_{test}(y, z)$ vary. The results are shown in Figure 1. Without any attempt to address measurement error in z', backdoor-adjustment is only marginally more robust than Logistic Regression (z'+BA vs. z+LR). When using ϵ thresholding, the performance is slightly improved in the extreme cases but only by a few points at most. However, when using the correlation matching method, we improve $F1_y$ by 10 to 15 points in the most extreme cases. For comparison, the figure also shows the fully observed case (z+BA), which uses back-door adjustment on the *true* values of z, to serve as an empirical upper-bound. We can see that correlation matching is comparable to the fully observed case, even with a 20% error rate on z. These results show that by getting a better estimate of the association between y and z, we can reduce attenuation bias and improve the robustness of our classifier, even though our observation of z is noisy.



(a) No adjustment. (b) Thresholding at $\epsilon = 0.75$. (c) Correlation matching.

Figure 2: Experimental results for back-door adjustment with an *unobserved* confounding variable in the location/gender dataset.

Variable $\mathbf{F1_z}$: We showed in the previous section that when we use our preliminary study with $F1_z = 0.784$, we can build a robust classifier using the correlation matching method combined with back-door adjustment. Additionally, we wish to see how sensitive the correlation adjustment methods are to the quality of $F1_z$. To do so, we increasingly add noise to the dataset used to train the preliminary classifier ($D_z = \{\mathbf{x}_i, \mathbf{z}_i\}$) to make $F1_z$ decrease. Because we want to visualize $F1_y$ against two variables ($F1_z$ and δ_{yz}), we visualize the results in a heatmap. Figure 2(a) shows the results for back-door adjustment when we use the predictions of the preliminary study but we do not try to fix the mislabeled values in z'. Figures 2(b) and 2(c) respectively show the results when we use ϵ thresholding with $\epsilon = 0.75$ and correlation matching. Similar to Figure 1, ϵ thresholding only brings small improvement to no adjustment at all. Furthermore, as $F1_z$ decreases, correlation adjustment with ϵ thresholding performs worse than when we are not doing any correlation adjustment. Clearly, the ϵ thresholding method is more sensitive to the quality of the preliminary study than the other methods.

The correlation matching method outperforms the other methods in robustness and $F1_y$ when $F1_z \geq 0.645$, as we can see by the wider range of red in Figure 2(c). This method is also sensitive to the quality of the preliminary study as we can see that $F1_y$ decreases with $F1_z$. Recall that in this data the best $F1_z$ is 0.784 — the trends suggest that correlation matching would continue to outperform baselines as $F1_z$ continues to increase.

5 Conclusion

In this paper, we have proposed two methods of using back-door adjustment to control for an unobserved confounder. Using a real-life dataset extracted from Twitter, we have found that correlation matching on the predicted confounder can recover the underlying correlation $r(y, z)$ and perform comparably to back-door adjustment with an observed confounder. We also showed that ϵ thresholding can be used to slightly improve the predictions compared to logistic regression, though it can harm accuracy when too many training instances are discarded. We showed that correlation matching provides a way to adjust for an unobserved confounder and outperform plain back-door adjustment as long as $F1_z > 0.65$. In future work, we will consider hybrid methods that combine ϵ thresholding and correlation matching to increase robustness as $F1_z$ decreases.

Acknowledgments

This research was funded in part by the National Science Foundation under awards #IIS-1526674 and #IIS-1618244.

References

Francis, D. P.; Coats, A. J.; and Gibson, D. G. 1999. How high can a correlation coefficient be? effects of limited reproducibility of common cardiological measures. *International journal of cardiology* 69(2):185–189.

Hand, D. J., and Henley, W. E. 1997. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 160(3):523–541.

King, G.; Keohane, R. O.; and Verba, S. 1994. *Designing social inquiry: Scientific inference in qualitative research*. Princeton university press.

Landeiro, V., and Culotta, A. 2016. Robust text classification in the presence of confounding bias. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Lazer, D.; Pentland, A. S.; Adamic, L.; Aral, S.; Barabasi, A. L.; Brewer, D.; Christakis, N.; Contractor, N.; Fowler, J.; Gutmann, M.; et al. 2009. Life in the network: the coming age of computational social science. *Science (New York, NY)* 323(5915):721.

Miller, C. C. 2015. Can an algorithm hire better than a human? *The New York Times* 25.

Monahan, J., and Skeem, J. L. 2016. Risk assessment in criminal sentencing. *Annual Review of Clinical Psychology* 12:489–513.

Pearl, J. 2003. Causality: models, reasoning and inference. *Econometric Theory* 19:675–685.

Pennacchiotti, M., and Popescu, A.-M. 2011. A machine learning approach to twitter user classification. *ICWSM* 11(1):281–288.