Efficient Classification with Counterfactual Reasoning and Active Learning

Azhar Mohammed, Dang Nguyen, Bao Duong, Thin Nguyen

Applied Artificial Intelligence Institute (A^2I^2), Deakin University, Geelong, Australia {mohammedaz, d.nguyen, duongng, thin.nguyen}@deakin.edu.au

Abstract. Data augmentation is one of the most successful techniques to improve the classification accuracy of machine learning models in computer vision. However, applying data augmentation to tabular data is a challenging problem since it is hard to generate synthetic samples with labels. In this paper, we propose an efficient classifier with a novel data augmentation technique for tabular data. Our method called CCRAL combines causal reasoning to learn counterfactual samples for the original training samples and active learning to select useful counterfactual samples based on a region of uncertainty. By doing this, our method can maximize our model's generalization on the unseen testing data. We validate our method analytically, and compare with the standard baselines. Our experimental results highlight that CCRAL achieves significantly better performance than those of the baselines across several real-world tabular datasets in terms of accuracy and AUC. Data and source code are available at: https://github.com/nphdang/CCRAL.

Keywords: Data Augmentation \cdot Classification \cdot Counterfactual reasoning \cdot Active learning \cdot Tabular data

1 Introduction

Recently, machine learning has become one of the most successful tools for supporting decisions, and it has been applied widely to many real-world applications including face recognition [36], security systems [3], disease detection [22], or recommended systems [33]. Two core components of a machine learning tool are the algorithm and the data. The algorithm can be classified into two mainstreams, namely classification and clustering while the data can be in different formats, e.g. tabular or image.

When dealing with images in computer vision applications, machine learning models (or classifiers) often leverage data augmentation techniques to improve the classification accuracy [14]. The main idea is that given an image of 'dog', if we rotate or flip the image, then we still recognize the object in the image as a 'dog'. By doing this geometric transformation, the label of an image is unchanged but we can obtain different variants of the image, helping the machine learning classifier to observe more data and improve its generalization. In addition to

geometric transformation, other data augmentation techniques are mix-up [43] and cut-mix [38].

In spite of a great success in computer vision, applying data augmentation to tabular data is challenging. There are three main reasons. First, an image is typically invariant to a small modification, e.g. flip, zoom, or rotation whereas a small change for a record in tabular data can result in a totally different outcome. All features (i.e. pixels) in images are i.i.d (independent and identical distributed) whereas each feature in tabular data (e.g. Sex or Age) has different ranges of values. Finally, one transformation operator can be applied to all features in images whereas each feature in tabular data often requires a relevant transformation operator depending on the type of the feature (continuous, discrete or categorical).

Our method. We propose an efficient classification method with a new data augmentation technique for tabular data. Our method has two main steps. First, we use causal reasoning to learn counterfactual samples for the original training samples. Each counterfactual sample is a variant of an original sample whose all feature values are the same except the intervened feature. Since the counterfactual samples may have different outcomes from the original ones, we obtain their labels via a matching method. Second, we augment counterfactual samples to real samples to create a new training set to train the classifier. Since not all counterfactual samples are useful, we select the meaningful ones that potentially improve the classification performance using an active learning based method. Our active learning is an uncertain-based approach. It determines samples that are difficult to predict, then obtains their counterfactual version to enrich the training data. Using both real and counterfactual samples, our classifier improves its generalization, resulting in a better accuracy on unseen testing samples.

Our contribution. To summarize, we make the following contributions.

- 1. We propose **CCRAL** (*Classifier with Causal Reasoning and Active Learning*), a novel method for classification with data augmentation in tabular data. To the best of our knowledge, **CCRAL** is the first method that combines both causal reasoning and active learning to train a classifier with synthetic samples in tabular data.
- 2. We develop an efficient framework to generate synthetic data. It consists of two steps: (1) it creates counterfactual samples via sample matching and (2) it selects useful counterfactual samples via active learning.
- We demonstrate the benefits of our method on five real-world tabular datasets, where our method is significantly better than the standard classifier in both accuracy and AUC measures.

The rest of the paper is organized as follows. In Section 2, we briefly outline the fundamentals of data augmentation methods, the generation of counterfactual data, and active learning in the literature. We describe our proposed framework **CCRAL** with an algorithm and illustrate the *region of uncertainty* in Section 3. Our experimental settings, datasets, results are presented in Section 4, where we evaluate the performance of **CCRAL** and compare it with two existing methods. Finally, we conclude our work in Section 5.

2 Related Works

Data Augmentation: It is a process of augmenting newly generated data to the existing training set for improving the model's robustness. It can be performed by a minor alteration to the existing data. For example, in computer vision data augmentation is used to enhance deep learning models by *flipping*, color spacing, injecting noise, random erasing to reduce the bias in the classifier to favor more frequently presented training examples [18,11]. It can also be performed by generating synthetic data to act as a regularizer and reduce over-fitting while training machine learning models [37]. Some algorithms such as data wrapping, SMOTE and MaxUp modify real-world examples to create augmented datasets [4,8,17]. However, these methods are exclusively useful for either specific kinds of data. For example, image recognition dataset or to improve the performance of a particular algorithm like AGCN [38].

Counterfactual Augmented Data (CAD): Another popular method is to augment data is by using counterfactual reasoning to improve the generalization of the model. CAD can be generated by using existing machine learning algorithms by matching closely related samples within the training set, for example, POLYJUICE to generate text and counterfactual image generation for generating images by using generative adversarial networks [40,27]. Generating diverse sets of realistic counterfactuals has proven to improve the model's training efficiency and overall results [26]. For example, in classification problems, the models trained on CAD were not sensitive to spurious features unlike modified data [21,7]. While, in discrimination and fairness literature counterfactual data substitution and CAD helped to mitigate gender bias by replacing duplicate text and handling conditional discrimination respectively [25,44]. However, counterfactually augmented data does not always generalize better than unaugmented datasets of the same size and may also hurt the model's robustness [19]. There is a significant gap to explore on the quantity and quality of counterfactual data needed to be augmented on original dataset by an effective learning process such that, the model generalizes better and is robust across various environments.

Active learning: It is a process that learns by an interaction between oracle and learner agent, it resolves the problem of costly data labeling in the learning process to improve the obtained model by making it efficient [9,32]. It can also be implemented on existing classification and predictive algorithms to optimize a model's performance when compared with state-of-the-art methods [10]. For example, in classification problems, logistic regression yielded remarkably better results by implementing the simplest suggested active learning method [23,34,41]. There are lots of effective approaches such as margin-based methods [13] and uncertainty sampling-based methods to optimize this process [16,35]. By using the uncertainty sampling-based learning process we can measure how certain a probabilistic classifier's prediction is and, obtain counterfactual versions of uncertain samples from the region of uncertainty to improve the model's transportability and robustness.

3 Framework

3.1 Problem definition

Let f(x) be a classifier and $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$ be a dataset. Each $y_i \in \{0, 1\}$ is a binary true label. Given a sample $x_i \in \mathcal{D}$, $f(x_i)$ provides a probability (called predicted score) that x_i belongs to label 1 (i.e. $f(x_i) = P(y_i = 1 \mid x_i)$ and $f(x_i) \in [0, 1]$). We denote the predicted label of x_i as $\hat{y}_i \in \{0, 1\}$, where \hat{y}_i is the rounding of $f(x_i)$ (i.e. $\hat{y}_i = 1$ if $f(x_i) \geq 0.5$, otherwise $\hat{y}_i = 0$).

Definition 1. (Accuracy). We define accuracy as $P(\hat{y} = y)$, which means the percentage of samples in \mathcal{D} predicted correctly by f(x).

Problem statement. Given a training set $\mathcal{D}_{tr} = \{x_i, y_i\}_{i=1}^N$ and a hold-out test set $\mathcal{D}_{te} = \{x_i, y_i\}_{i=1}^M$, our goal is to learn a classifier f(x) using \mathcal{D}_{tr} such that f(x) maximizes its accuracy on \mathcal{D}_{te} . This is the traditional classification problem in machine learning [5].

3.2 Proposed method CCRAL

A typical way to solve the above problem is to train the classifier f(x) using the available samples in the training set \mathcal{D}_{tr} , which tries to minimize a loss function measuring the difference between the true labels y and the predicted labels \hat{y} . Although this approach is straightforward, it often does not achieve good results.

Our method to solve the classification problem described in Section 3.1 is novel. Our main idea is that instead of using only training samples in \mathcal{D}_{tr} , we try to obtain more training samples, which is very helpful in improving the generalization of the classifier. When the classifier observes more training samples, it is more robust and its classification accuracy is often improved on *unseen* test samples. This process is often called *data augmentation*, which has become the state-of-the-art method to improve the performance of deep learning models in computer vision [37].

Our approach, called <u>Classifier with <u>Causal Reasoning and Active Learning</u> (**CCRAL**), has two main steps: (1) learning counterfactual samples using causal reasoning and (2) training a classifier with both real and counterfactual samples using active learning.</u>

Learning counterfactual samples. We are dealing with the classification task on tabular data. Typically, a tabular dataset includes a mix of different types of features. They can be continuous, binary, or categorical features. Following the standard approach in causal reasoning [39], given the training set \mathcal{D}_{tr} we select one binary feature T as the treatment feature. For example, the treatment feature can be Sex="male/female" or Marital_Status="single/married".

After determining the treatment feature T, we can obtain the counterfactual of any sample $x \in \mathcal{D}_{tr}$. Given a sample x_i , assume that its treatment feature has value 0 (i.e. $T_i = 0$), we then change the value of the treatment feature to

1. By doing this way, we now have the counterfactual sample \bar{x}_i of x_i , which is the same as x_i except that the treatment feature of \bar{x}_i has value 1 instead of 0.

Since \bar{x}_i is not a real sample, we do not have its label. To find the label \bar{y}_i of \bar{x}_i , we use the sample matching approach that computes the distance between \bar{x}_i and other samples $x' \in \mathcal{D}_{tr}$, and uses the label of the nearest sample as the label of \bar{x}_i [6]. The formulation is retrieved the label of \bar{x}_i is as follows:

$$\bar{y}_i = y(\operatorname*{argmin}_{x' \in \mathcal{D}_{tr}} d(\bar{x}_i, x')), \tag{1}$$

where $d(\bar{x}_i, x')$ is the function computing the distance between the counterfactual sample \bar{x}_i and a sample $x' \in \mathcal{D}_{tr}$. Any distance can be used, for example, Euclidean, cosine, or Manhattan distances. In our case, we use the Euclidean distance. The function $\underset{x' \in \mathcal{D}_{tr}}{\operatorname{argmin}} d(\bar{x}_i, x')$ returns the sample that is nearest to \bar{x}_i , and $y(x_i)$ is the function that returns the label of an sample $x_i \in \mathcal{D}_{tr}$.

Training classifier with real and counterfactual samples. Using Equation (1), we can generate the counterfactual version of any sample $x \in \mathcal{D}_{tr}$. The next question is how to use these counterfactual samples to improve the classification. Should we create the counterfactual counterpart for each sample, and augment them to the original training data to train the classifier? Using all counterfactual samples might not be a good solution. First, these counterfactual samples are unreal samples, they might add noises to the training data. Second, the quality of the labels of the counterfactual samples depend on how we compute the distance in Equation (1). Finally, in some cases, if there were not very similar samples with the counterfactual sample \bar{x}_i , then the label \bar{y}_i would be random.

To overcome the three above challenges when using the counterfactual samples as training data, we propose an *active learning* based method. We first train a classifier f(x) using samples x_i in the training data \mathcal{D}_{tr} . Once we have learned the classifier f(x), we use it to predict the score for each sample $x_i \in \mathcal{D}_{tr}$.

Since the classifier f(x) has been trained with \mathcal{D}_{tr} , f(x) predicts confidently the labels for most of the samples in \mathcal{D}_{tr} , where their predicted scores are close to 0 or 1. However, some samples are difficult to predict their outcomes, where their scores are close to the decision boundary (i.e. their scores are close to 0.5). We call these samples are uncertain samples.

To determine which samples are uncertain, we define an $uncertain\ region$ as follows:

$$0.5 - \alpha \le f(x) \le 0.5 + \alpha,\tag{2}$$

where α is the region margin, $0.5 - \alpha$ is the lower region margin, and $0.5 + \alpha$ is the upper region margin.

From Equation (2), if any training sample x_i whose predicted score $f(x_i)$ is in the uncertain region, then it will be the uncertain sample. Figure 1 illustrates the uncertain region and the uncertain samples.

Since the classifier f(x) is very confused about the label of uncertain samples. It could be useful if we used their counterfactual version for the training process. Let $\mathcal{U} = \{x_1, x_2, ..., x_n\}$ be the set of uncertain samples. Following the process

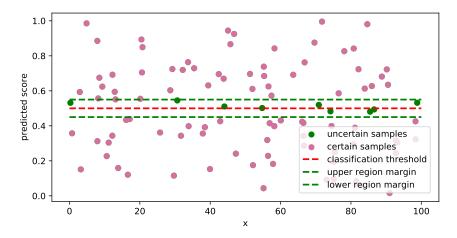


Fig. 1. Illustration of uncertain region and uncertain samples. Uncertain samples are indicated by green circles while the uncertain region is formed by two dashed green lines, the upper region margin and the lower region margin.

in Section 3.2, we learn counterfactual version for each sample $x_i \in \mathcal{U}$. We then augment these counterfactual samples $\bar{\mathcal{U}} = \{\bar{x}_1, \bar{x}_2, ..., \bar{x}_n\}$ to the original training set \mathcal{D}_{tr} i.e. we have the new training set $\mathcal{D}'_{tr} = \mathcal{D}_{tr} \cup \bar{\mathcal{U}}$. Finally, we train the classifier f(x) again with the new training set \mathcal{D}'_{tr} .

Since the region margin α has values being in the range of [0,0.5], we use a grid search (or Bayesian optimization [28]) to find the α that derives the best classifier f(x) measured on a validation set \mathcal{D}_{va} . In particular, at each search iteration, we expand the uncertain region by increasing the value of α , and obtain more uncertain samples. We then find the counterfactual counterparts of these uncertain samples. Finally, we train the classifier f(x) with real training samples along with the counterfactual samples and measure its accuracy on a validation set. The final classifier is the classifier whose accuracy is highest on the validation set, and this final classifier will be evaluated on the hold-out test set.

Algorithm 1 summarizes our method CCRAL.

4 Experiments and Discussions

We conduct extensive experiments on five real-world tabular datasets to evaluate the classification performance (accuracy and AUC) of our method **CCRAL**, comparing it with two strong baselines.

4.1 Datasets

To create an environment for comprehending counterfactual reasoning involved in our method **CCRAL**, we choose five real-world tabular datasets that have at least one binary feature that intrigues one's causal thinking. These datasets

Algorithm 1: The proposed CCRAL algorithm.

```
Input: \mathcal{D} = \{x_i, y_i\}_{i=1}^N: training set, K: # of iterations
 1 split \mathcal{D} into a (smaller) training set \mathcal{D}_{tr} and a validate set \mathcal{D}_{va};
 2 define a grid of margins [\alpha_1, \alpha_2, ..., \alpha_K];
 3 train a classifier f(x) on \mathcal{D}_{tr};
 4 select a binary feature T as the treatment feature;
 5 for each sample x_i \in \mathcal{D}_{tr} do
          generate its counterfactual sample \bar{x}_i by changing the value of the
            treatment feature of x_i;
          compute its counterfactual label \bar{y}_i = y(\operatorname{argmin}_{x' \in \mathcal{D}_{tx}} d(\bar{x}_i, x')) (see
 7
           Equation (1);
 8 use f(x) to predict a score f(x_i) for each sample x_i \in \mathcal{D}_{tr};
 9 for k = 1, 2, ..., K do
10
          find U_k = \{x_1, x_2, ..., x_n\}, where x_i is an uncertain sample i.e.
           0.5 - \alpha_k \le f(x_i) \le 0.5 + \alpha_k (see Equation (2));
          generate new training data \mathcal{D}_{tr}^k = \mathcal{D}_{tr} \cup \bar{\mathcal{U}}^k where \bar{\mathcal{U}}^k = \{\bar{x}_1, \bar{x}_2, ..., \bar{x}_n\} is
11
           the counterfactual of \mathcal{U}^k;
          train f_k(x) on \mathcal{D}_{tr}^k;
12
          evaluate accuracy acc_k of f_k(x) on \mathcal{D}_{va};
14 return the best classifier f_{k^*}(x), where k^* = \operatorname{argmax}_k acc_k;
```

were often used to evaluate fairness-aware and causal inference machine learning algorithms [15,42,29].

Table 1 shows characteristics of each dataset along with the selected treatment feature and the respective outcome.

Table 1. Characteristics of five tabular datasets. We denote N: the number of samples, M: the number of features, T: the treatment feature, and y: the class feature.

Dataset	N	M	T	T=1	T = 0	y	y = 1	y = 0
german	1,000	20	Sex	"male"	"female"	Credit	"good"	"bad"
bank	4,521	14	Marriage	"married"	"single"	Subscription	"yes"	"no"
twins	4,821	52	Weight	"heavier"	"lighter"	Mortality	"alive"	"death"
compas	4,010	10	Sex	"male"	"female"	Rearrested	"no"	"yes"
adult	30,162	13	Sex	"male"	"female"	Income	">50K"	"<50K"

german: this dataset describes each individual's credit score whether she/he has a good or bad credit score [12]. It has 1,000 samples and 20 features. We use Sex as the treatment feature.

bank: this dataset is about direct marketing campaigns of individuals for term deposit subscriptions. The outcome of this data is whether a person is subscribed or not depending upon the marketing and duration campaigned. Marriage is the treatment feature in this dataset.

twins: this dataset consists of around 5,000 records of twin's birth collected during the period of 1989-1991 in the U.S. [1]. It is a popular benchmark dataset in causality researches [24]. The outcome corresponds to the mortality of each twin's during the first year of birth. We choose twins of the same gender to replicate the counterfactual. The treatment feature is the twin's weight.

compas: this dataset includes a collection of data in Broward country, Florida about the use of the COMPAS risk assessment tool and has the data regarding felonies and charges on the degree of the arrest [2]. This dataset has the treatment feature Sex with an outcome of getting rearrested within two years.

adult: this dataset is the collection of individual data of their income recorded during the 1994 U.S census [20]. The outcome is a person's income. If the income is greater than \$50K, then it is labeled as "1". Otherwise it is "0". This dataset has 30,162 samples and 13 features. We select Sex as the treatment feature.

4.2 Baselines and evaluation

We compare our method **CCARL** with two strong baselines.

- 1. Standard: this method uses available training samples to train a classifier.
- 2. Counterfactual: this method uses the counterfactual samples of all original training samples in the training process. In other words, it fixes $\alpha = 0.5$ in Equation (2).

For a fair comparison, we measure the accuracy and AUC of each method on the same hold-out test set. We also use the same classifier for all methods, namely the Support Vector Machine (SVM) with the linear kernel and C=1 for the regularization. Note that other machine learning classifiers can be used with our method. We use the default search range [0,0.5] for α , and set the number of iterations K=10. We evaluate methods on each dataset in five times with different train-test data splits, and report the averaged accuracy and AUC.

4.3 Results

Figure 2 shows the accuracy of each method on five datasets. It can be seen that our method **CCRAL** is much better than the standard classifier on all datasets. On *german* (a small dataset), Standard achieves only 61.0% whereas **CCRAL** achieves 70.0%, resulting in 9% better. On *adult* (a very large dataset), the accuracy of Standard is 79.28% compared to 82.82% of our **CCRAL**. On this dataset, our method achieves around 3% gains over the standard classifier.

Compared to the Counterfactual method, **CCRAL** is comparable on three datasets *bank*, *twins*, and *adult* while it is much better on two datasets *german* and *compas*. This shows that using all counterfactual samples in the training process was not a good solution since they might add noise to the training data, as we discussed in Section 3.2. Our method which uses active learning to select useful counterfactual samples is a more efficient approach to train the classifier.

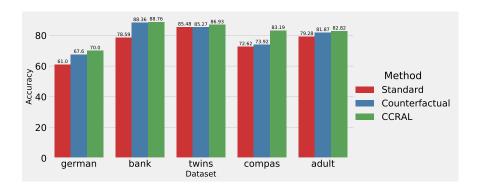


Fig. 2. The averaged classification accuracy of two baselines Standard, Counterfactual, and our method CCRAL on each dataset.

We also report the AUC of each method in Figure 3. Our **CCRAL** is the best method, where it significantly outperforms two baselines Standard and Counterfactual. **CCRAL** always outperforms the standard classifier by a large margin across all datasets. Compared to Counterfactual, our method shows a great improvement, where it achieves 3-9% gains over Counterfactual. Again, this suggests that using active learning to select useful counterfactual samples is a much better strategy than using all counterfactual samples for training the classifier.

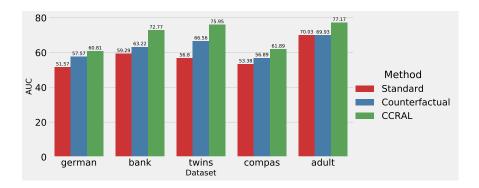


Fig. 3. The averaged AUC of two baselines Standard, Counterfactual, and our method CCRAL on each dataset.

5 Conclusion

In this paper, we have introduced an efficient classifier (named **CCRAL**) with a novel data augmentation technique for tabular datasets. We generate counterfactual data by flipping the binary value of the treatment feature of original

training samples, and obtain their labels by using a matching method. We use active learning to select useful counterfactual samples based on a region of uncertainty depending on the predicted scores of the original training samples. We augment selected counterfactual samples to the set of original training samples to train the classifier. We demonstrate the efficacy of **CCRAL** on five standard real-world tabular datasets. The obtained results show that **CCRAL** generalizes better and is more robust towards unseen testing samples, where it significantly outperforms other methods. Our approach can be conceptually extended to other types of data such as sequences [31] and graphs [30].

Acknowledgment

This research is partly supported by NHMRC Ideas Grant GNT2002234.

References

- Almond, D., Chay, K.Y., Lee, D.S.: The costs of low birth weight. The Quarterly Journal of Economics 120(3), 1031–1083 (2005)
- Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. ProPublica 23(2016), 139–159 (2016)
- Apruzzese, G., Colajanni, M., Ferretti, L., Marchetti, M.: Addressing adversarial attacks against security systems bsed on machine learning. In: Proceedings of the International Conference on Cyber Conflict. vol. 900, pp. 1–18 (2019)
- Baird, H.S.: Document image defect models. In: Structured Document Image Analysis, pp. 546–556. Springer (1992)
- 5. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
- Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D.X., Chickering, D.M., Portugaly, E., Ray, D., Simard, P., Snelson, E.: Counterfactual reasoning and learning systems: The example of computational advertising. Journal of Machine Learning Research 14(11) (2013)
- Chang, C.H., Adam, G.A., Goldenberg, A.: Towards robust classification model by counterfactual and invariant data generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15212–15221 (2021)
- Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321–357 (2002)
- 9. Cohn, D.A., Ghahramani, Z., Jordan, M.I.: Active learning with statistical models. Journal of Artificial Intelligence Research 4, 129–145 (1996)
- Collet, T., Pietquin, O.: Active learning for classification: An optimistic approach. In: Proceedings of the Symposium on Adaptive Dynamic Programming and Reinforcement Learning. pp. 1–8 (2014)
- 11. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
- Dua, D., Graff, C.: UCI machine learning repository (2019), http://archive.ics.uci.edu/ml
- 13. Ducoffe, M., Precioso, F.: Adversarial active learning for deep networks: a margin based approach. arXiv preprint arXiv:1802.09841 (2018)

- Fawaz, H.I., Forestier, G., Weber, J., Idoumghar, L., Muller, P.A.: Data augmentation using synthetic data for time series classification with deep residual networks. arXiv preprint arXiv:1808.02455 (2018)
- Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 329–338 (2019)
- Gal, Y., Islam, R., Ghahramani, Z.: Deep Bayesian active learning with image data. In: Proceedings of the International Conference on Machine Learning. pp. 1183–1192 (2017)
- 17. Gong, C., Ren, T., Ye, M., Liu, Q.: Maxup: Lightweight adversarial training with data augmentation improves neural network training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2474–2483 (2021)
- 18. Hernández-García, A., König, P.: Data augmentation instead of explicit regularization. arXiv preprint arXiv:1806.03852 (2018)
- Huang, W., Liu, H., Bowman, S.R.: Counterfactually-augmented SNLI training data does not yield better generalization than unaugmented data. arXiv preprint arXiv:2010.04762 (2020)
- Kamiran, F., Karim, A., Zhang, X.: Decision theory for discrimination-aware classification. In: Proceedings of the IEEE International Conference on Data Mining. pp. 924–929 (2012)
- Kaushik, D., Hovy, E., Lipton, Z.C.: Learning the difference that makes a difference with counterfactually-augmented data. arXiv preprint arXiv:1909.12434 (2019)
- Kumar, V.B., Kumar, S.S., Saboo, V.: Dermatological disease detection using image processing and machine learning. In: Proceedings of the International Conference on Artificial Intelligence and Pattern Recognition. pp. 1–6 (2016)
- Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 3–12 (1994)
- 24. Louizos, C., Shalit, U., Mooij, J., Sontag, D., Zemel, R., Welling, M.: Causal effect inference with deep latent-variable models. arXiv preprint arXiv:1705.08821 (2017)
- Maudslay, R.H., Gonen, H., Cotterell, R., Teufel, S.: It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. arXiv preprint arXiv:1909.00871 (2019)
- Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 607–617 (2020)
- 27. Neal, L., Olson, M., Fern, X., Wong, W.K., Li, F.: Open set learning with counter-factual images. In: Proceedings of the European Conference on Computer Vision. pp. 613–628 (2018)
- Nguyen, D., Gupta, S., Rana, S., Shilton, A., Venkatesh, S.: Bayesian optimization for categorical and category-specific continuous inputs. In: AAAI. vol. 34, pp. 5256– 5263 (2020)
- Nguyen, D., Gupta, S., Rana, S., Shilton, A., Venkatesh, S.: Fairness Improvement for Black-box Classifiers with Gaussian Process. Information Sciences 576, 542–556 (2021)
- 30. Nguyen, D., Luo, W., Nguyen, T., Venkatesh, S., Phung, D.: Learning graph representation via frequent subgraphs. In: SDM. pp. 306–314. SIAM (2018)
- 31. Nguyen, D., Luo, W., Nguyen, T., Venkatesh, S., Phung, D.: Sqn2Vec: Learning sequence representation via sequential patterns with a gap constraint. In: ECML-PKDD. vol. 11052, pp. 569–584. Springer (2018)

- 32. Nissim, N., Boland, M.R., Tatonetti, N.P., Elovici, Y., Hripcsak, G., Shahar, Y., Moskovitch, R.: Improving condition severity classification with an efficient active learning based framework. Journal of Biomedical Informatics **61**, 44–54 (2016)
- Portugal, I., Alencar, P., Cowan, D.: The use of machine learning algorithms in recommender systems: A systematic review. Expert Systems with Applications 97, 205–227 (2018)
- 34. Settles, B.: Active Learning. Morgan & Claypool Publishers (2012)
- 35. Settles, B., Craven, M., Ray, S.: Multiple-instance active learning. Advances in Neural Information Processing Systems **20**, 1289–1296 (2007)
- Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K.: Accessorize to a crime: Real
 and stealthy attacks on state-of-the-art face recognition. In: Proceedings of the
 ACM SIGSAC Conference on Computer and Communications Security. pp. 1528–
 1540 (2016)
- 37. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of Big Data **6**(1), 1–48 (2019)
- Walawalkar, D., Shen, Z., Liu, Z., Savvides, M.: Attentive cutmix: An enhanced data augmentation approach for deep learning based image classification. arXiv preprint arXiv:2003.13048 (2020)
- Wang, J., Mueller, K.: The visual causality analyst: An interactive interface for causal reasoning. Transactions on Visualization and Computer Graphics 22(1), 230–239 (2015)
- Wu, T., Ribeiro, M.T., Heer, J., Weld, D.S.: POLYJUICE: Generating counterfactuals for explaining, evaluating, and improving models. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (2021)
- 41. Yang, Y., Loog, M.: A benchmark and comparison of active learning for logistic regression. Pattern Recognition 83, 401–415 (2018)
- Zafar, M.B., Valera, I., Rodriguez, M.G., Gummadi, K.P., Weller, A.: From parity to preference-based notions of fairness in classification. arXiv preprint arXiv:1707.00010 (2017)
- Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
- 44. Žliobaite, I., Kamiran, F., Calders, T.: Handling conditional discrimination. In: Proceedings of the IEEE International Conference on Data Mining. pp. 992–1001 (2011)