

# Dealing with Bias via Data Augmentation in Supervised Learning Scenarios

Vasileios Iosifidis<sup>1</sup> and Eirini Ntoutsis<sup>1</sup>

L3S Research Center, University of Hannover, Germany  
{iosifidis, ntoutsis}@l3s.de

**Abstract.** There is an increasing amount of work from different communities in data mining, machine learning, information retrieval, semantic web, and databases on bias discovery and discrimination-aware learning with the goal of developing not only good quality models but also models that account for fairness. In this work, we focus on supervised learning where biases towards certain attributes like race or gender might exist. We propose data augmentation techniques to correct for bias at the input/data layer. Our experiments with real world datasets show the potential of augmentation techniques for dealing with bias.

## 1 Introduction

Nowadays, decision making systems tend to become fully automated by replacing human judgment with algorithmic decisions that rely solely or to a great extent on data. Since decision making systems are data-driven, they can be applied in a wide variety of applications from recommendation systems and insurance ratings to medical diagnoses and decisions on whether a patient should receive treatment or not.

However, concerns have been raised as these algorithms may under-perform if trained on pre-existing biases which lay inside data distributions. These concerns led to anti-discrimination laws which try to prevent different treatment of individuals or groups based on specific attributes (e.g ethnicity, gender), also named protected attributes. Even without considering protected attributes in the learning process, algorithms can still be unfair towards specific individuals or groups. The reason can be explained by analyzing the data: in some cases, particular attributes, called *proxies*, can reveal the value of a protected attribute (e.g., attribute “wife” or “husband” can reveal the protected attribute “gender”).

One of the main reasons which causes discrimination in the classification process is the under-representation of protected groups. For instance, medical treatment data may lack observations of a specific disease misjudging ill patients as healthy. Under-represented groups tend to be highly misclassified compared to over-represented groups. In this work, we focus on improving the correctly classified instances of a protected group without degrading the overall classification performance. To this end, we propose data augmentation techniques to increase the representation of the (minority) protected group.

The rest of the paper is organized as follows: Section 2 gives an overview of the related work. Section 3 presents data augmentation techniques which have been used for dealing with class imbalance. In Section 4 we present our experimental analysis. Conclusions are given in Section 5.

## 2 Related work

There is an increasing amount of work on bias discovery and discrimination-aware learning [1]. Discrimination discovery methods try to spot discrimination for a query instance or a group of instances by investigating how model decisions vary between the query object and similar instances or the rest of the population.

Pre-processing approaches manipulate the data by either resampling to allow for a fair representation of minority classes or by manipulating the features to detect proxies to protected features [2–5]. In [2], authors propose a method in which instances that are closer to the decision border have higher probabilities to be oversampled than the rest of the instances. In [3], authors try to eliminate bias by switching class labels from most influential instances.

In-processing approaches reformulate the classification problem by incorporating in the optimization function the discrimination behavior of the model [6–8] and they optimize the model for both classification performance and fairness.

Post-processing approaches correct the output model for discrimination by either modifying its decision regions to allow for a fairer class representation or by taking into account both privacy and discrimination concerns during publishing. In a slightly different line of research, some methods build more human-interpretable models either by design [9] or through translation from complex models into simpler, more interpretable models. Also, there exist methods that explain model decisions for single instances or sets of instances.

## 3 Data augmentation techniques

Data augmentation refers to the process of data generation by using information from the training corpus. This process can increase the robustness of a model and prevent the model from overfitting.

For data augmentation we consider two alternatives: *Oversampling* and SMOTE [10].

*Oversampling* is a naive method which just duplicates instances of the minority group to incur balance. The selection of the instances to be duplicated is random.

On the other hand, SMOTE does not duplicate instances, rather it generates pseudo-instances in the neighborhoods of the minority group. The algorithm starts by taking each minority class instance and finding its  $k$ -nearest minority neighbors. Afterwards it randomly selects  $j$  of these neighbors and generates synthetic instances along the lines joining the minority sample and its  $j$  selected neighbors.

We consider two options when we generate pseudo instances. First option is to produce instances based on a given attribute. We force balance by populating the minority group for a specific attribute. Option two is to create pseudo instances based on a given attribute with respect to class. In detail, we generate instances from the under represented group of an attribute to deal with group’s class imbalance. For the former option we refer to this by `OverSample(attribute)` or `SMOTE(attribute)` while for the latter we refer to it as `OverSample w.r.t Class(attribute)` or `SMOTE w.r.t Class(attribute)`.

## 4 Experiments

The goal of the experiments is to evaluate the effect of augmentation on classifier’s performance and w.r.t fairness. We use Weka [11] for implementation and experimentation.

As our base classifiers we use Naive Bayes, for its simplicity and efficiency, and J48 for its ability to classify extremely fast unknown instances and for its robustness to outliers. For our experiments we use 10-fold-cross validation. We perform stratified sampling in order to split our datasets into training and testing sets, using 66% for training and 33% for testing. We use testing set only for evaluation purposes while we use augmentation only for training set to avoid overestimation of classifier’s performance. As evaluation metrics we employ AUC and F1 score. We prefer AUC since it reflects better the classifier’s performance, comparing to e.g., accuracy, when dealing with class imbalance [12].

Moreover, augmentation is performed multiple times, each time with random seeds for Oversampling and SMOTE. We report on the average scores of AUC, F1 and percentage of correctly classified instances.

#### 4.1 Datasets

**Census Income:** The Census Income dataset [13] contains 48.884 instances from 1994 and 1995 Current Population Surveys conducted by the U.S. Census Bureau. The data contains 41 demographic and employment related variables. The prediction task of this dataset is to decide if a person receives less than 50K per year or more. After we removed instances with missing values and duplicated entries we ended up with 45.175 instances. We noticed that attribute “race” is dominated by the category “White” while the other categories have lower counts. Since we focus on binary classification we grouped the rest categories as a new one which we call “Minorities”. In the end, 38.859 instances have race=“White” and 6.316 have race=“Minorities”. As positive class we consider instances whose income is less or equal to 50K and as negative when income is more than 50K.

**German Credit:** The German Credit dataset [13] consists of 1.000 instances. The prediction task of this dataset is to determine if it is risky or not to give credit for a person class “good”, “bad”. The dataset contains 20 attributes plus the class attribute but none of them explicitly refers to gender. Based on the attribute “Personal status and sex” (which contains values such as: “Male-divorced”, “Female-divorced-married”, “Male-single”, “Male-married” and “Female-single”), we derive a new attribute namely “gender” which contains 690 instances of gender=“male” and 310 instances of gender=“female”.

#### 4.2 Results on the Census Income Dataset

Table 1 describes the resulted cardinalities from each augmentation method and also the corresponding AUC and F1 scores of each classifier. It is clear that even though datasets have been modified their performance has not dropped.

Although, AUC and F1 scores are good indicators to help us monitor the performance of each classifier, they are not appropriate for depicting fairness among protected groups. Figures 1 and 2 demonstrate the percentages of correctly classified instances for each of our augmentation methods. In Figure 1, we show experiments which are contacted to “Gender” attribute, which we call “Gender case”, while in 2 to “Race” attribute namely “Race case”.

In “Gender” case, Figure 1, women have high classification error from the negative class. Smote does not have good impact in this case. The reason is that the number of women who are in the negative class is too low. Even though pseudo instances are

generated, positive class instances are mostly benefited. Smote creates pseudo instances based on the neighborhood which means pseudo instances of women who belong in negative class will have higher probability to include properties of women who belong in positive class. Despite this limitation, in this particular case, Smote can generate new information. Contrarily, oversampling overcomes this problem by adding weights to the already existing instances. We notice that oversampling, Smote wrt class and oversampling wrt class aid the female population to increase the true negative scores. Difference between SMOTE and SMOTE w.r.t class is visible.

By comparing classifiers, apparently they have same behavior when it comes to aiding minority class but J48 seems to be slightly more stable when contrasting male population in negative class. By observing Figure 2, we confirm that augmentation is beneficial as in “Gender” case. Classifiers exhibit same behavior as in former case.

Table 1: Census-income: Overall results of augmentation methods

	Class		Sensitive Attributes				Naive Bayes		J48	
	Positive	Negative	Male	Female	White	Minor.	AUC	F1 Sc.	AUC	F1 Sc.
Training Set (no augmentation)	22,648	7,468	20,348	9,768	25,915	4,201	<b>0.902</b>	0.814	<b>0.893</b>	0.848
SMOTE(Race)	43,664	8,166	35,274	16,556	25,915	25,915	0.897	0.809	<b>0.896</b>	0.85
SMOTE(Gender)	32,996	7,700	20,348	20,348	36,028	4,668	0.898	0.817	0.894	0.849
SMOTE(Class)	22,648	22,648	35,346	9,950	41,016	4,280	<b>0.902</b>	0.826	0.888	0.845
SMOTE w.r.t Class(Race)	22,648	10,328	23,128	9,848	25,915	7,061	<b>0.902</b>	0.819	0.888	0.851
SMOTE w.r.t Class(Gender)	22,648	15,053	20,348	17,353	33,500	4,201	0.895	0.817	0.892	0.849
OverSample(Race)	40,674	11,156	32,409	19,421	25,915	25,915	0.901	0.817	0.863	0.845
OverSample(Gender)	31,571	9,125	20,348	20,348	34,492	6,204	0.901	0.816	0.864	0.844
OverSample(Class)	22,648	22,648	33,311	11,985	39,735	5,561	<b>0.902</b>	0.825	0.812	0.824
OverSample w.r.t Class(Race)	22,648	10,999	23,139	10,508	25,915	7,732	0.901	0.812	0.885	0.847
OverSample w.r.t Class(Gender)	22,648	16,145	20,348	18,445	33,491	5,302	0.886	0.806	0.870	0.841

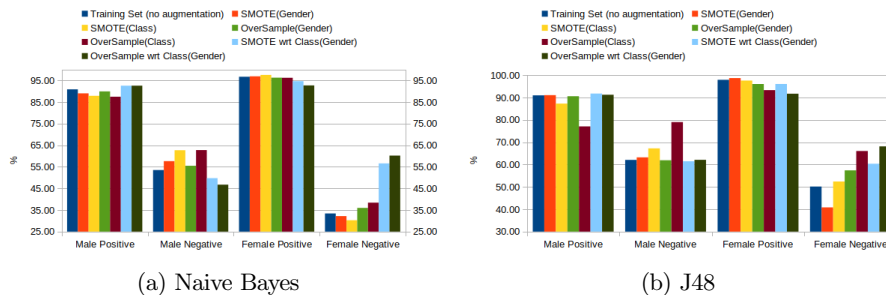


Fig. 1: Census-Income: “Gender case” study

### 4.3 Results on the German Credit Dataset

In Table 2, we report the results of each augmentation method for German census. In some cases, augmentation is slightly better than original training dataset. In this dataset, in contrast to census income dataset, only “gender” attribute can be considered

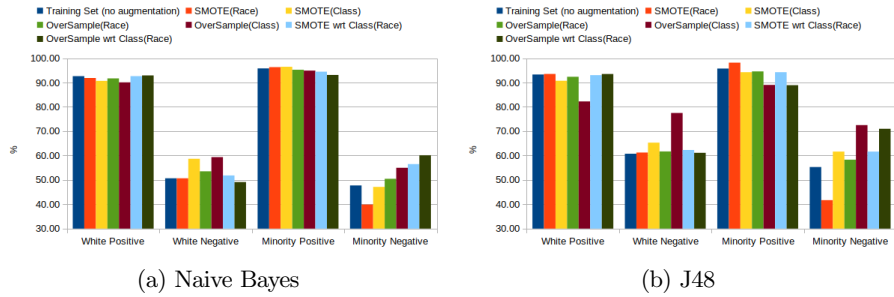


Fig. 2: Census-Income: “Race case” study

as protected. We investigate this case study and depict the correctly classified instances in Figure 3.

Table 2: German-credit: Overall results of augmentations methods

	Class		Sensitive Attr.		Naive Bayes		J48	
	Positive	Negative	Male	Female	AUC	F1 Sc.	AUC	F1 Sc.
Training set (no augmentation)	466	200	461	205	0.770	0.718	0.686	0.699
SMOTE(Class)	466	466	688	244	0.756	0.693	0.657	0.661
SMOTE(Gender)	689	233	461	461	0.756	0.693	0.657	0.661
SMOTE w.r.t Class(Gender)	466	253	461	258	0.752	0.720	0.667	0.691
Oversample(Class)	466	466	628	304	<b>0.772</b>	0.716	0.652	0.670
OverSample(Gender)	613	314	461	466	0.763	0.718	0.645	0.675
OverSample w.r.t Class(Gender)	466	253	461	258	0.759	0.713	<b>0.695</b>	0.699

By examining Figure 3, it is visible that negative class oversampling helps female minority group the most. Using Smote on women’s group has a negative impact on correctly classified instances for the negative class. Since majority of women belong to the positive class this behavior is expected. Same happens when oversampling is applied. In addition, as we have already noticed in census income dataset, augmentation w.r.t class is beneficial without effecting significantly the overall performance of a classifier.

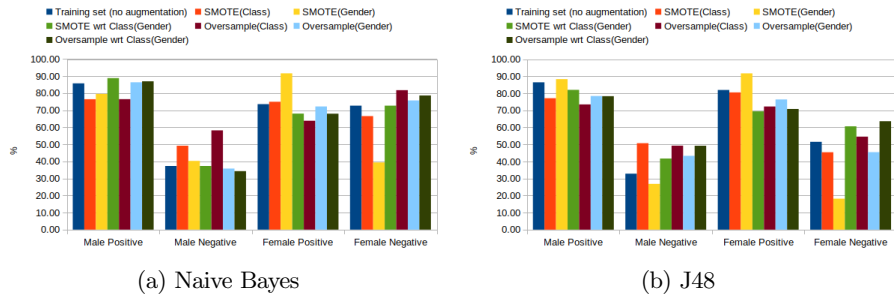


Fig. 3: German-Credit: “Gender case” study

## 5 Conclusions

In this work, we deal with bias towards certain attributes. Our approach to eliminate bias is to generate pseudo instances in order to enhance minority groups. We experiment on two real world datasets: Census income and German Credit. The gained insights from the analysis can tell us that data augmentation can reduce classification error for discriminated groups. Furthermore, even though different classifiers do not perform equally good, they exhibit positive results when data augmentation takes place. By and large, data augmentation is useful when dealing with classification bias.

## Acknowledgment

The work was partially funded by the European Commission for the ERC Advanced Grant ALEXANDRIA under grant No. 339233 and by the German Research Foundation (DFG) project OSCAR (Opinion Stream Classification with Ensembles and Active learners) No. 317686254.

## References

1. S. Hajian, F. Bonchi, and C. Castillo, “Algorithmic bias: From discrimination discovery to fairness-aware data mining,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016.
2. F. Kamiran and T. Calders, “Classification with no discrimination by preferential sampling,” in *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*. Citeseer, 2010.
3. —, “Classifying without discriminating,” in *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on*. IEEE, 2009.
4. —, “Data preprocessing techniques for classification without discrimination,” *Knowledge and Information Systems*, 2012.
5. B. T. Luong, S. Ruggieri, and F. Turini, “k-mn as an implementation of situation testing for discrimination discovery and prevention,” in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011.
6. F. Kamiran, T. Calders, and M. Pechenizkiy, “Discrimination aware decision tree learning,” in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*.
7. M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, “Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment,” in *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017.
8. —, “Fairness constraints: Mechanisms for fair classification,” *arXiv preprint arXiv:1507.05259*, 2017.
9. J. Zeng, B. Ustun, and C. Rudin, “Interpretable classification models for recidivism prediction,” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2017.
10. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, 2002.
11. I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.
12. C. X. Ling, J. Huang, and H. Zhang, “Auc: a better measure than accuracy in comparing learning algorithms,” in *Conference of the canadian society for computational studies of intelligence*. Springer, 2003.
13. C. J. Merz and P. M. Murphy, “{UCI} repository of machine learning databases,” 1998.