BayesBoost: Identifying and Handling Bias Using Synthetic Data Generators

Barbara Draghi BARBARA.DRAGHI@MHRA.GOV.UK
Zhenchen Wang ZHENCHEN.WANG@MHRA.GOV.UK
Puja Myles PUJA.MYLES@MHRA.GOV.UK
Medicine and Healthcare products Regulatory Agency, London UK

Allan Tucker ALLAN.TUCKER@BRUNEL.AC.UK
Brunel University London, UK

Editors: Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michał Woźniak and Shuo Wang.

Abstract

Advanced synthetic data generators can model sensitive personal datasets by creating simulated samples of data with realistic correlation structures and distributions, but with a greatly reduced risk of identifying individuals. This has huge potential in medicine where sensitive patient data can be simulated and shared, enabling the development and robust validation of new AI technologies for diagnosis and disease management. However, even when massive ground truth datasets are available (such as UK-NHS databases which contain patient records in the order of millions) there is a high risk that biases still exist which are carried over to the data generators. For example, certain cohorts of patients may be under-represented due to cultural sensitivities amongst some communities, or due to institutionalised procedures in data collection. The under-representation of groups is one of the forms in which bias can manifest itself in machine learning, and it is the one we investigate in this work.These factors may also lead to structurally missing data or incorrect correlations and distributions which will be mirrored in the synthetic data generated from biased ground truth datasets. In this paper, we explore methods to improve synthetic data generators by using probabilistic methods to firstly identify the difficult to predict data samples in ground truth data, and then to boost these types of data when generating synthetic samples. The paper explores attempts to create synthetic data that contain more realistic distributions and that lead to predictive models with better performance.

Keywords: Synthetic data generators, data bias, over-sampling, Bayesian network

1. Introduction

The use of synthetic data for the rapid development of AI systems in healthcare represents a great potential to avoid issues concerning patient privacy that have been highlighted in the General Data Protection Regulation (Goodman and Flaxman, 2017). This type of data captures the structure and distributions that are apparent in the real data sets whilst also preserving patient privacy and avoiding the risks of individual identification. Synthetic data generation occurs by applying techniques such as building generative models based on real data (Patki et al., 2016). In this case, models that capture the correct relationships and distributions are built, either hand-coded based upon expert knowledge or inferred from real

data using models such as Bayesian networks (BNs) (Young et al., 2009). These can then be used to generate synthetic data via sampling techniques. However, even using established techniques such as Bayesian networks for the generation of high-fidelity synthetic patient data (Tucker et al., 2020), and even when huge datasets are available, there is a high risk that biases still exist, which are carried over to the data generators. The presence of biases within data has proved to be a significant problem in applying AI techniques. Indeed, there is the risk of replicating and even amplifying human biases, particularly those affecting protected groups (Chodosh, 2018). Algorithmic bias can manifest in several ways with varying degrees of consequences for the subject group: biases in online recruitment tools (Hamilton, 2018), biases in word association (Hadhazy, 2017), biases in criminal justice decision making (Angwin et al., 2016) and others (Lee et al., 2019) Problems with biased training data have also led to accusations of prejudice in machine learning models ((Cossins, 2018), (Emspak, 2016)). The reason can be that data selected from a biased sample of the population leads to decisions that reflect the biases already in our society. As mentioned above, bias in machine learning can show up in several forms (Mehrabi et al., 2019). To directly solve the bias in the world is unrealistic, but what we can do is take measures to weed out bias from our data or models. In this work, bias is regarded as under-representation of groups of patients, whatever that reason may be. Similarly, the application of synthetic data generators on data in which this type of bias is present can lead to the generation of synthetic data in which specific cohorts of patients may be under-represented due to cultural sensitivities amongst some communities or standardized procedures in data collection. These factors may also lead to structurally missing data or incorrect correlations and distributions mirrored in the synthetic data generated from biased ground truth datasets. Datasets in medicine are often imbalanced, and the under-representation of certain patient groups reflects this imbalance. Existing approaches to deal with these imbalances include de-biasing methods, like Reweighing (Calders et al., 2009), Adversarial Debiasing (Zhang et al., 2018) and Reject option classification (Herbei and Wegkamp, 2006), and synthetic data generation approaches, including SMOTE (Chawla et al., 2002) and variants such as Adaptive Synthetic Sampling (AdaSyn) (He et al., 2008). The de-biasing methods aim to mitigate the bias prevalent in the training data to create an unbiased model regarding how it makes decisions based on specific sensitive attributes, whilst SMOTE and AdaSyn aim to re-balance the data on the class variables. In this paper, we explore a new technique, *BayesBoost*, that combines a Bayesian network synthetic data generator with a boosting approach. The aim is to identify under-represented samples in a ground truth dataset and then use the synthetic data to over-sample the under-represented cases and achieve a better overall distribution of overall features. The purpose of our work differs from the aim of the methods mentioned above since BayesBoost aims to create synthetic data that are more representative of the ground truth data population.

2. Method

In this section, we firstly define the developed methodology, describing the simulation of data biases and the BayesBoost approach. We then introduce the two datasets on which we test the approach. Finally, we describe the empirical analysis.

2.1. Methodology

The experiment can be broken down into three main sections. First, a data size reduction approach is proposed to generate as small a dataset as possible whilst maintaining data quality that is high enough for good model performance, thus reducing the inclusion of unnecessary data. Second, we move on to the simulation part of our work, which involves generating synthetic data containing biases. This section allows us to be sure to have data containing under-represented groups of patients. The third and final section of this work concerns BayesBoost: the identification and handling of data biases from a given dataset containing under-represented groups.

Simulation of data biases The first section of our work concerns a data size reduction approach. Given a dataset D_{GT} we identify binary classes (here representing the occurrence of disease) to be predicted and select the sample sizes to be extracted. Different percentage samples of the original dataset are investigated: 50%, 10%, and 1%. A random sample and a test set of the same size are extracted from D_{GT} for each desired size. After repeating the procedure n times for each dimension to be sampled, two tables reporting the AUC values calculated for the ROC and precision-recall curves for each iteration and each dimension are extracted. The variation in AUC values between the complete dataset and the samples for each iteration is adopted as a metric to choose the best size to sample. Therefore, the dimension leading to the minimal AUC values variation is used in the subsequent analyses of our approach. In fact, the goal is to reduce the size of the dataset while maintaining a data quality that is high enough for good model performance. Thus, as long as the ability to classify a target variable doesn't decrease, the extracted sample keeps most of the dataset's intrinsic characteristics. We chose to use the AUC as a metric since it represents a valid measure of classification performance (Bradley, 1997). After identifying the optimal sample size for D_{GT}, we move on to the simulation part of our work, which involves generating synthetic data containing biases. Bayesian networks (BNs) are used to model different samples of data. BNs are probabilistic models representing a set of stochastic variables with their respective dependencies and conditional distributions They facilitate the generation of random samples conditional on some evidence. Therefore, it is possible to generate random samples conditional on some evidence. This enables us to generate data with hand-coded biases in order to test our approach. After learning a Bayesian network from D_{GT}, we identify the numeric attributes we want to under-represent and the respective value beyond which we want to obtain the under-representation (e.g. if the chosen variable is age and the chosen value is 60, it means that subjects over 60 years of age will be under-represented in the sampled synthetic data). For categorical variables we explore two methods: One is to introduce within the network new conditional probabilities in the Bayesian network from which the data will be extracted. This approach is preferable if we want to have control over all levels of the variable. Another possible way is to proceed as with numeric variables, by choosing a value to under-represent (e.g. gender == F) and let the remaining percentage of data be composed of data with other values of the variable in question (e.g. gender == M, gender == I). This approach is preferable when the considered variables are binary, and therefore, once the value to be under-represented has been chosen, the only other remaining value is over-represented. Finally, the percentage of data to be under-represented is identified (e.g. 30% means that a data set will be generated containing 30%

of subjects having the characteristic chosen as the characteristic to be under-represented). After choosing these parameters, the data extraction from the Bayesian network is carried out using logic sampling (Henrion, 1988). This use of evidence to control the synthetic data generation enables us to produce data with the exact degree of under-sampled cases. In order to completely separate the biased data from the original synthetic data, a Bayesian network is learned based on the obtained dataset and, from this, a dataset of the desired size is extracted. This dataset, which from now on we will refer to as D_{Bias} the data set that, in our simulation, represents the original data set on which to apply the method for identifying and correcting data biases. The simulation of biases process is summarized in Figure 1.

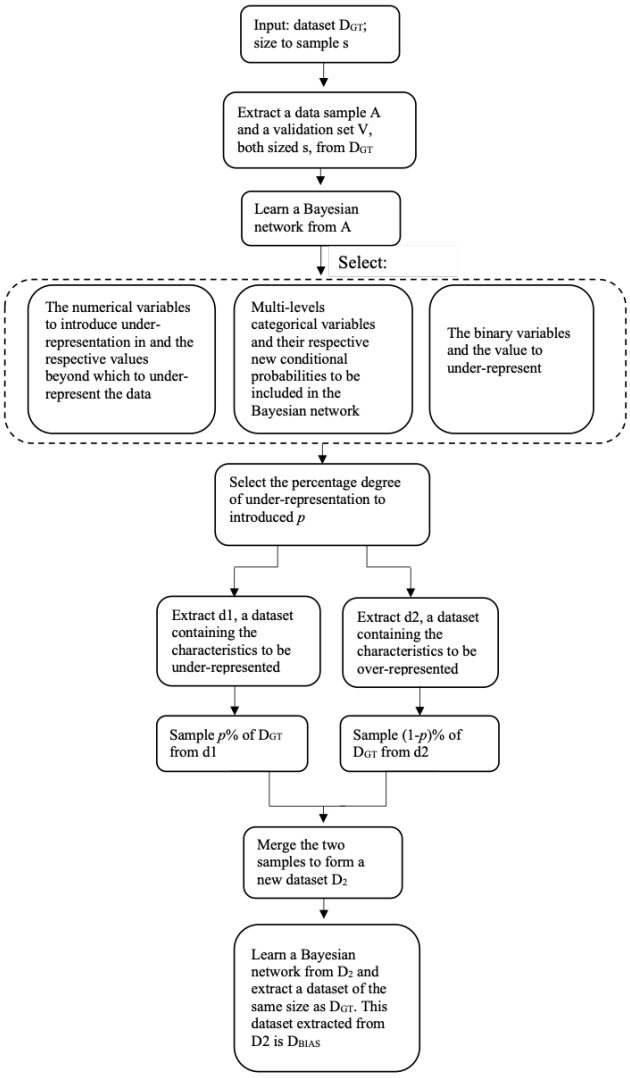

Figure 1: Simulation of biases process.

BayesBoost The third and final section of this work concerns BayesBoost: the developed approach for the identification and handling of data biases from a given dataset. In order to identify groups of under-represented data, the idea is to test a classifier, trained on D_{Bias}, in predicting a binary variable on a validation set extracted a priori from data A. If dataset D_{Bias} presents an under-representation of data groups, we expect to highlight them through a difficulty in classifying them. In particular, we choose to consider all the subjects that the classifier gets with uncertainty as difficult to classify. In our method, the uncertainty is a parameter that we consider to be each probability between some interval (here we choose the interval to be between 0.4 and 0.7, representing an uncertain binary classification). In literature, there are many measures used to evaluate fairness and/or bias in data sets, including Statistical Parity Difference, Equal Opportunity Difference and Average Odds Difference (Mehrabi et al., 2019). Although very useful, these metrics are model-oriented and oriented towards the identification of positive outcomes. Our approach does not aim to obtain a model in which different subgroups can have the same chance of being classified as positive.In fact, we only want to identify all those cases that are difficult to classify, regardless of whether they are positive or negative outcomes. The data classified with uncertainty within this interval forms a new dataset called D_{Unc}. Attributes are sorted based on differences between the distributions of D_{Bias} and D_{Unc}, framing the ordered set of variables named O. In order to generate m rows for each row of D_{Unc}, the idea is to sample new data from a Bayesian network, learned on dataset D_{Bias}, after assuming the values belonging to the data contained in D_{Unc} as evidence. Therefore, we expect the newly generated dataset to report subjects having characteristics that have been under-represented in D_{Bias}. At the end of this procedure, we will have a new data set that constitutes the new synthetic dataset D_{Boost} once added to D_{Bias}. It should be specified that the number m of rows to extract constitutes another parameter of our approach that will undoubtedly be subject to future optimizations. Three attempts are made: the first involves the extraction of m data for each row of D_{Unc} to obtain a dataset equal to half the D_{Bias} size. The second involves the extraction of m data to obtain a dataset with dimensions equal to D_{Bias}, while the third involves the extraction of m number of rows for each row of D_{Unc} in such a way as to obtain a dataset with dimensions equal to twice that of D_{Bias}. The full details of this entire process are documented fully in Algorithm 1.

2.2. Datasets

The developed approach is applied on synthetic datasets based on anonymised real primary care data (Wang et al., 2019) from the Clinical Practice Research Datalink (CPRD), a real-world research service supporting retrospective and prospective public health and clinical studies in the UK. It is jointly sponsored by the Medicines and Healthcare products Regulatory Agency and the National Institute for Health Research, as part of the Department of Health and Social Care (Wolf et al., 2019). First, the approach is applied to the CPRD Synthetic cardiovascular disease datasets (CVD) (Clinical Practice Research Datalink, 2020a), a dataset focusing on cardiovascular disease risk factors. The dataset covers 499,344 patients and 21 variables, including stroke or heart attack, smoking habits, region, age, chronic diseases, body mass index, systolic blood pressure and other cardiovascular disease risk factors. CVD is a mixed dataset because it contains both numeric

Algorithm 1: BayesBoost Algorithm: Pipeline for the identification and handling of biases

Input: a dataset containing biases D_{Bias} , an independent validation set V, binary class to predict C, range probabilities p_1 and p_2

1. Learn a Bayesian network b from D_{Bias} and fit the parameters of the Bayesian network b
2. Train a model on D_{Bias} and test it on V to predict values of C
3. Extract all the outcomes that the classifier get with uncertainty. Being p the outcome probability, we consider as uncertain all the cases where $p_1 < p < p_2$
4. Create the dataset D_{Unc} containing all the data from V that correspond to the uncertain outcomes.
5. For each factor variable, calculate the frequencies (i.e. the proportion of subjects belonging to each category in D_{Bias} and D_{Unc}) and compute the average of the absolute value of the difference between the frequencies of D_{Bias} and D_{Unc}.
6. For each numerical variable, calculate the mean of the absolute value of the difference between quartiles, medians and means of D_{Bias} and D_{Unc}.
7. Define the ordered set of variables O by excluding the one to be predicted and sorting the variables in descending order based on the difference between the distributions of D_{Bias} and D_{Unc} (the difference is calculated in 5 and 6).
8. Being U_{i} the i^{th} row of D_{Unc}, for each U_{i}:
    9. For each variable O_{j} contained in O, construct the evidence using the value of the same variable in U_{i}
    10. Using the constructed evidence, attempt to extract m samples from b by inferring the variable to be predicted
    11. If m rows are extracted, continue with the next row in D_{Unc}
    12. Otherwise, remove variables to use as evidence by starting with the last positioned within O
    13. If a variable o_{i} is not used as evidence, its value in U_{i} is copied within the new dataset.
14. The result is a new dataset that is added to D_{Bias}, producing a new dataset C.

and factor variables. Finally, the method is applied to the CPRD COVID-19 Synthetic datasets (Clinical Practice Research Datalink, 2020b), which focuses on patients presenting to primary care with symptoms indicative of COVID-19 (confirmed/suspected COVID-19) and control patients with negative COVID-19 test results. The dataset covers 779,546 patients and 47 variables, including age, age categories, gender, region, covid-19 diagnosis and covid-19 test results.

2.3. Experiments

Simulation of data biases As stated before, our experiment focuses on two synthetic data generator models focused on cardiovascular disease and covid-19. In order to simulate data biases, the idea is to introduce biases in variables that impact classification performances. To do that, we analysed the structure of the BNs learnt from each dataset, and the variables to consider for the biases simulation are identified based on direct relationships with the target variables. The selected variables for each dataset are specified in the following paragraphs.

CVD Synthetic Data Being CVD Synthetic data a mixed dataset, we can introduce biases in both numeric and categorical variables. In particular, we present two different bias protocols for this dataset. When applying bias protocol cvd_1, an under-representation is introduced for the variables age and body mass index (bmi), with a percentage of under-representation at 20%. Also, new conditional probabilities are introduced for the variables gender and chronic disease (ckidney). Bias protocol cvd_2 consider age, systolic blood pressure (sbp), gender, region and smoking.

Covid-19 Synthetic Data Considering the covid-19 synthetic data, we introduce biases in factor variables since the dataset shows only one numeric variable (age), which can be regarded by considering the correspondent factor attribute (age category, catAge). Three bias protocols are investigated for this dataset: bias protocol covid_1 introduces biases in the region, bias protocol covid_2 considers age categories, and bias protocol covid_3 introduces biases in both variables.

BayesBoost After simulating biases, BayesBoost is applied to detect under-represented groups. Regarding the CVD data, three different variables are predicted: stroke and heart attacks, atrial fibrillation and type 2 diabetes. The target variable within the COVID-19 data is covid-19 diagnosis. For each attempt, the results of BayesBoost are three synthetic datasets: BB_{50}, BB_{100} and BB_{200}, depending on the degree of oversampling used in the BayesBoost Algorithm. BB_{50} results from the extraction of m data to boost the original data with an extra 50% of the size of D_{Bias}. BB_{100} results from the extraction of m data to boost the dataset by 100%. BB_{200} is the outcome when extracting m data to boost the dataset by 200%. In order to assess the efficacy of our approach, we generate synthetic datasets by applying SMOTE and Adaptive Synthetic Sampling (AdaSyn) to the dataset in which biases have been deliberately introduced. Then, we compare the two outcomes to those obtained by BayesBoost. In particular, the resulting five synthetic data sets are compared in predicting a binary variable by training a Naive Bayes classifier and test the models on the same independent test set. AUC values are estimated after computing the ROC and precision-recall curves for each prediction.

3. Results

We now explore the results from the two different datasets to see how BayesBoost performs. The outcome of this approach is a synthetic dataset which we expect to be more representative of reality (characteristics of the original dataset, D_{GT}, without added biases) in terms of underlying distributions and predictive performance. We compare the AUC values computed on the following datasets: the original dataset D_{GT}; the dataset containing data biases D_{Bias}; BB_{50}, BB_{100} and BB_{200}, three synthetic datasets obtained by applying the approach developed to D_{Bias} and varying the degree of oversampling; D_{Smote} the synthetic dataset obtained by applying SMOTE to D_{Bias}; D_{Adasyn} the synthetic dataset obtained applying AdaSyn to D_{Bias}. As stated before, this work aims to produce a synthetic dataset as close to reality as possible. Therefore, the distributions of some of the variables for which under-representations have been introduced are compared. The datasets producing the highest AUC values is chosen between BB_{50}, BB_{100} and BB_{200}, and represented when comparing distributions as D_{Boost}. Moreover, to highlight the ability of BayesBoost to identify those groups of data that have been under-represented, distributions are also reported of the data classified with uncertainty, D_{Unc}. We expect these distributions to reflect the under-represented cases. The following paragraphs report the most significant results obtained for each dataset by applying the bias protocols described above.

CVD Synthetic Data Regarding the CVD dataset, the AUC values computed for the ROC and precision-recall curves resulted from the bias protocol cvd_2 and the different predictions are reported in Table 1 and Table 2, respectively. Results obtained by applying the bias protocol cvd_1 are similar to the ones proposed here.

Table 1: The table shows the AUC values calculated for the ROC curves obtained by testing the models trained on the different datasets.

	Target	**Bias Protocol**	$\mathbf{D_{GT}}$	$\mathbf{D_{Bias}}$	$\mathbf{BB_{50}}$	$\mathbf{BB_{100}}$	$\mathbf{BB_{200}}$	$\mathbf{D_{Smote}}$	$\mathbf{D_{AdaSyn}}$
(a)	strokeha	cvd_2	0.85	0.82	0.83	**0.84**	**0.84**	0.82	0.81
(b)	af	cvd_2	0.87	0.85	**0.87**	**0.87**	**0.87**	0.84	0.84
(c)	type2	cvd_2	0.84	0.83	**0.83**	0.82	0.81	0.82	0.82

The distributions of the results obtained when applying (a) are reported in Figure 2, Figure 3 and Figure 4, where distributions of the variables age, gender and smoking are reported. Results obtained from attempts (b) and (c) are similar to the ones reported here. These figures show the distributions of some of the variables in which under-representations have been introduced. Comparing the distributions of D_{Unc} and D_{Bias}, the main conclusion is that BayesBoost correctly identifies the under-represented groups. Furthermore, the distribution of the D_{Bias} data shows how BayesBoost overcomes the biases highlighted from D_{Bias} and capture by D_{Unc} , which identifies from D_{Bias} which subjects need to be sample boosted. As can be seen in Figure 2, the distribution of uncertain data mainly identifies subjects aged over 75, who are the ones who have been under-represented in D_{Bias}. Our approach, SMOTE and AdaSyn, produce data with realistic distributions by increasing all

Table 2: The table shows the AUC values calculated for the P-R curves obtained by testing the models trained on the different datasets.

	Target	Bias Protocol	D_{GT}	D_{Bias}	BB_{50}	BB_{100}	BB_{200}	D_{Smote}	D_{AdaSyn}
(a)	strokeha	cvd_2	0.33	0.24	0.26	0.28	**0.29**	0.24	0.23
(b)	af	cvd_2	0.18	0.13	0.16	**0.17**	**0.17**	0.13	0.13
(c)	type2	cvd_2	0.24	0.18	**0.19**	**0.19**	**0.19**	0.18	0.18

those data that had been under-represented. Figure 3 shows how, after under-representing female subjects, the distribution of D_{Unc} highlights a high percentage of females. Furthermore, the result obtained by the BayesBoost approach presents distributions closer to reality, increasing the number of female subjects who have been under-represented in D_{Bias}. On the contrary, DSmote and DAdaSyn maintain a distribution similar to D_{Bias}. Figure 4 shows the distribution of the smoking variable, where level 0 indicates non-smoker patients, level 1 refers to ex-smokers, level 2 to light smokers, 3 to moderate smokers and 4 to heavy smokers. Once again, we can see how the D_{Unc} data distribution identifies the data that have been under-represented in D_{Bias}. Moreover, we can see how the D_{Boost} gets closer to the real data distribution, unlike DSmote and DAdaSyn, which derive from the realistic data distribution.

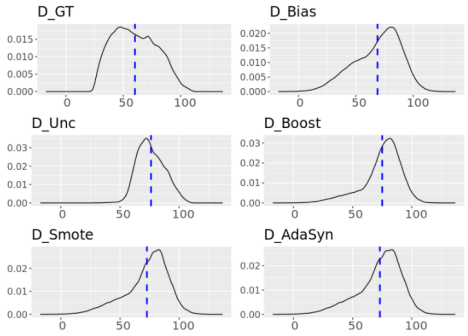

Figure 2: Age density distributions of the data obtained by attempt (a).

Covid-19 Synthetic Data This section analyses the outcomes obtained by applying the bias protocol covid_3 and predicting the covid diagnosis, but similar results are obtained by applying bias protocols covid_1 and covid_2. The AUC computed for the ROC and precision-recall curves are reported in Table 3 and Table 4. Comparisons between the distributions of the obtained outcomes are shown in Figure 5.

Considering Figure 5(a), it is evident that our approach reduces the number of data belonging to class 1 to obtain a distribution of data that is as close to reality as possible, while the other approaches show distributions that are closer to D_{Bias}. Considering Figure 5(b) comparing the distribution of the age category in datasets D_{Bias} and D_{Unc}, we can see that

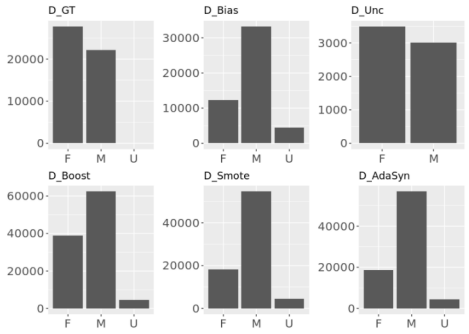

Figure 3: Gender distributions of the data obtained by attempt (a). The x-axis shows the values, and the y-axis represents the frequencies.

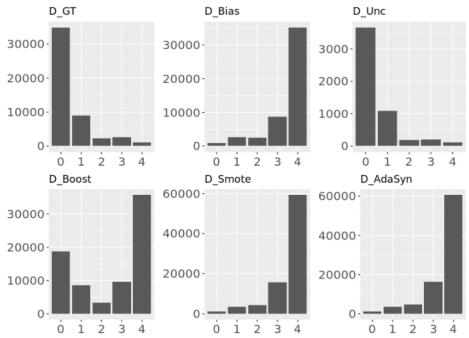

Figure 4: Smoking distributions of the data obtained by attempt (a). The x-axis shows the values, and the y-axis represents the frequencies.

Table 3: The table shows the AUC values calculated for the ROC curves obtained by testing the models trained on the different datasets. The target variable is the covid diagnosis.

$\mathbf{D_{GT}}$	$\mathbf{D_{Bias}}$	$\mathbf{BB_{50}}$	$\mathbf{BB_{100}}$	$\mathbf{BB_{200}}$	$\mathbf{D_{Smote}}$	$\mathbf{D_{AdaSyn}}$
0.7	0.69	0.71	0.72	**0.73**	0.7	0.7

Table 4: The table shows the AUC values calculated for the P-R curves obtained by testing the models trained on the different datasets. The target variable is the covid diagnosis.

$\mathbf{D_{GT}}$	$\mathbf{D_{Bias}}$	$\mathbf{BB_{50}}$	$\mathbf{BB_{100}}$	$\mathbf{BB_{200}}$	$\mathbf{D_{Smote}}$	$\mathbf{D_{AdaSyn}}$
0.21	0.2	0.22	0.23	**0.25**	0.21	0.2

our approach correctly identifies the data that has been under-represented in D_{Bias}. It is also evident that our approach produces a dataset more similar to reality than the datasets produced with SMOTE or AdaSyn, which, as we can see, do not introduce improvements to the distribution of the variable that has been deliberately under-represented. In conclusion, Figure 5 shows that BayesBoost identifies the under-represented groups in D_{Bias} highlighting them in D_{Unc} as groups that need to be sample boosted. The result is an over-sampling of these groups, leading the resulting dataset D_{Bias} to overcome those biases.

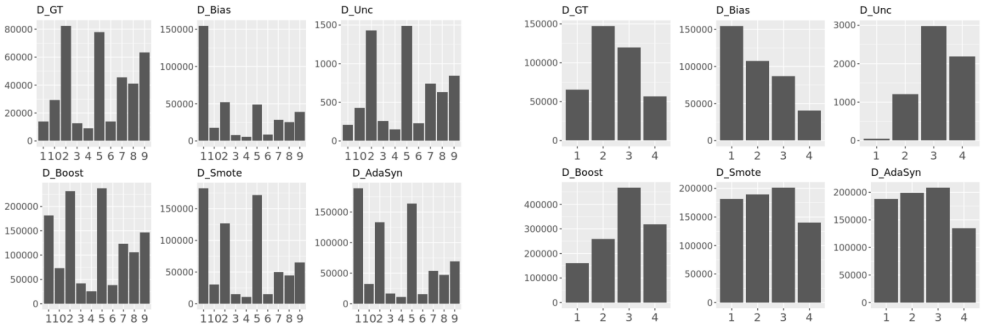

Figure 5: 5 (a) represents the region distribution; (b) represents the age categories distribution. The x-axes show the values, and the y-axes represent the frequencies.

4. Conclusion

BayesBoost is a novel approach that aims to detect and overcome biases within data. Being able to identify under-represented cohorts of patients represents a powerful technique, especially when considering synthetic data generators. The use of rare event detection and correction approaches can be essential for a synthetic dataset service like that used at the Clinical Practice Research Datalink in the UK. The advantage of having access to synthetic data is that it can be used instead of real patient data for complex statistical analyses and machine learning and artificial intelligence (AI) research applications. Indeed, detecting and correct biases in the ground truth datasets will avoid structurally missing data or incorrect correlations and distributions to be mirrored in the synthetic data generated from

biased ground truth datasets. Methods for detecting rare events include synthetic sampling approaches like SMOTE and AdaSyn. As stated before, these techniques aim to balance the classification problem. Therefore, they aim to balance the target variable, even if it results in deviating from the correct distribution of the data. Figure 6 illustrates this point by plotting the resulting distributions of a target variable for one of our simulations on covid data where we inject bias and try to infer the original distributions. The results obtained by applying SMOTE and AdaSyn show the resulting distribution in the new data balances the classification problem (the target variable is balanced between the two levels), despite obtaining data distributions that deviate from the original unbiased data, D_{GT}. On the contrary, the data distributions obtained through BayesBoost resemble the original distribution D_{GT}. The reason is that, as we can see from the distribution of data D_{Bias} and data D_{Unc}, the target variable does not show variations after introducing biases in the data. BayesBoost approach does not detect biases for this variable since they are not present and, consequently, does not attempt to balance that variable. The results obtained through the application of SMOTE and AdaSyn showed us comparable AUC values, both for the results obtained on the CVD data and for the covid-19 data. The reported results show how the datasets resulting from the application of BayesBoost lead to better AUC values than those obtained with SMOTE and AdaSyn. Moreover, there are cases in which the classification performance carried out using the D_{Boost} even exceeds the D_{GT}. Additionally, BayesBoost shows an excellent ability to identify under-represented groups within data and correct them, as we can see considering the obtained data distributions.

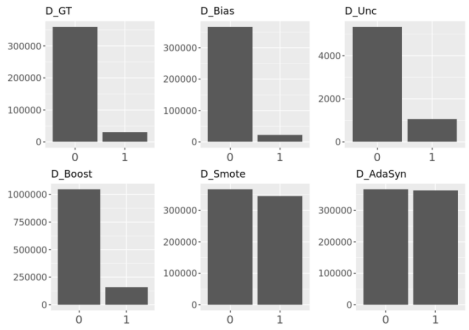

Figure 6: Distribution of the covid diagnosis variable obtained when applying our approach on the Covid-19 Synthetic data.

5. Further works

There are still a number of issues with BayesBoost that need to be explored. In particular, the correct parameterization (such as to what extent the under-represented samples need to be boosted). What is more, the approach has only been applied when classifying binary variables, but future development will undoubtedly consider multiclass classification problems. Likewise, it will be possible to investigate new datasets and, in particular, new diseases.

Acknowledgments

This work has been funded by a grant from NHSX.

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. propublica, may 23, 2016, 2016.

Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357, 2002.

Sara Chodosh. Courts use algorithms to help determine sentencing, but random people get the same results. *Popular Science*, 2018.

Clinical Practice Research Datalink. Cprd cardiovascular disease synthetic dataset (version 2020.06.001) [data set], 2020a. https://doi.org/10.11581/YK6N-B652.

Clinical Practice Research Datalink. Cprd covid-19 symptoms and risk factors synthetic dataset april 2021 (version 2021.04.001) [data set], 2020b. https://doi.org/10.48329/fbjh-es87.

Daniel Cossins. Discriminating algorithms: 5 times ai showed prejudice. *New Scientist*, 12, 2018.

Jesse Emspak. How a machine learns prejudice. *Scientific American, December*, 29:2016, 2016.

Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a 'right to explanation'. *AI Magazine*, 38(3):50–57, Oct 2017. ISSN 0738-4602. doi: 10.1609/aimag.v38i3.2741. URL http://dx.doi.org/10.1609/aimag.v38i3.2741.

Adam Hadhazy. Biased bots: Artificial-intelligence systems echo human prejudices. *Princeton University, April*, 18, 2017.

Isobel Asher Hamilton. Why it's totally unsurprising that amazon's recruitment ai was biased against women. *Retrieved from Business Insider: https://www. businessinsider. com/amazon-ai-biased-against-women-no-surprise-sandra-wachter-2018-10*, 2018.

Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. IEEE, 2008.

Max Henrion. Propagating uncertainty in bayesian networks by probabilistic logic sampling. In *Machine intelligence and pattern recognition*, volume 5, pages 149–163. Elsevier, 1988.

Radu Herbei and Marten H. Wegkamp. Classification with reject option. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 34(4):709–721, 2006. ISSN 03195724. URL http://www.jstor.org/stable/20445230.

Nicol Turner Lee, Paul Resnick, and Genie Barton. Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. *Brookings Institute: Washington, DC, USA*, 2019.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *CoRR*, abs/1908.09635, 2019. URL http://arxiv.org/abs/1908.09635.

Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410, 2016. doi: 10.1109/DSAA.2016.49.

Allan Tucker, Zhenchen Wang, Ylenia Rotalinti, and Puja Myles. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ digital medicine*, 3(1):1–13, 2020.

Zhenchen Wang, Puja Myles, and Allan Tucker. Generating and evaluating synthetic uk primary care data: Preserving data utility & patient privacy. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 126–131. IEEE, 2019.

Achim Wolf, Daniel Dedman, Jennifer Campbell, Helen Booth, Darren Lunn, Jennifer Chapman, and Puja Myles. Data resource profile: clinical practice research datalink (cprd) aurum. *International journal of epidemiology*, 48(6):1740–1740g, 2019.

Jim Young, Patrick Graham, and Richard Penny. Using bayesian networks to create synthetic data. *Journal of Official Statistics*, 25(4):549, 2009.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pages 335–340, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278779. URL https://doi.org/10.1145/3278721.3278779.