Optimizing Data Augmentation vs. Racial Bias in Heart Disease Classification

Asmi Kumar (G), Rayna Arora (UG)

October 28, 2023

1 Project Responsibilities

Rayna Arora (UG):

1. Process dataset into a format that can be passed into the models
2. Implement logistic regression and decision tree models
3. Implement random oversampling and undersampling data augmentation techniques
4. Analyze, based on various metrics, how much the dataset should be augmented and which technique should be used in order for the accuracy discrepancies across race to be statistically insignificant and to minimize the overall accuracy sacrificed.
5. Prepare the project presentation and final proposal

Asmi Kumar (G):

1. Implement random forest and k-nearest neighbors models
2. Implement the SMOTE data augmentation technique
3. Write the code to process accuracy measures from each model run and calculate ANOVA scores
4. Analyze, based on various metrics, how much the dataset should be augmented and which technique should be used in order for the accuracy discrepancies across race to be statistically insignificant and to minimize the overall accuracy sacrificed
5. Prepare the project presentation and final proposal

2 Thesis

In the context of utilizing machine learning for health predictions or classifications, certain racial groups might be underrepresented, leading to potential misdiagnosis or underdiagnosis. To ensure equitable accuracy across all racial groups, what is the minimum necessary augmentation of the dataset required to render the variations in predictive accuracy statistically insignificant?

3 Introduction

There has been much discussion in class about disparities in health diagnoses among various racial or gender groups, attributed to unequal data collection [1]. These inequities may also stem from societal biases [2], impacting disease risk estimations and causing delays in necessary treatments. Research explores dataset augmentation to address distribution disparities and improve accuracy across races. One study [3] investigates two techniques: oversampling (duplicating examples from the minority class) and SMOTE (generating synthetic samples via linear interpolation). While equalizing training examples across classes had varied effects, excessive augmentation can lead to overfitting. Thus, a study to determine the optimal augmentation level is proposed to minimize discrepancies across races and preserve overall accuracy.

4 Experiments

We will use the heart disease prediction dataset from Kaggle, which includes demographic and health information for 319795 patients and includes six different races. We plan to first try various models

such as logistic regression, decision tree, random forest, and k-nearest neighbors on this dataset. We will choose the model(s) that yield the highest accuracy.

In this dataset, Caucasians make up the vast majority, with the next largest population being only $\frac{1}{5}$ as large. When augmenting our data, we will make it so that each of the minority race populations each make up the same percentage of the dataset, and incrementally increase this percentage until all races make up the same proportion. Our baseline will be the initial accuracy for each race with our chosen model(s) on the entire dataset (no data augmentations). We plan to apply the following data augmentation techniques:

- Random oversampling: Randomly selecting examples from the minority class(es), with replacement, and adding them to the training dataset
- Random undersampling: Randomly selecting examples from the majority class(es) and deleting them from the training dataset
- Other undersampling techniques like NearMiss: A variation on the naive random undersampling approach, which includes implementing a set of heuristic rules in order to select samples
- SMOTE data augmentation: An oversampling technique where synthetic data samples are generated based on linear interpolation between original examples of minority class(es)

We plan to run a one-way ANOVA in order to evaluate whether the difference in accuracies for each race is statistically significant. For each of our dataset adjustments per augmentation technique, we will need to run the model 30 times in order to generate enough data for statistical testing with ANOVA.

For each data augmentation technique and data distribution combination, we will train and test our model 30 times in order to find a mean accuracy measure for each race. Thus, we anticipate having multiple arrays (one for each race) that each contain accuracy measures during a particular run (e.g., total accuracy, F1 score, AUC). For example, within a particular data augmentation technique, `caucasian[0]` will contain the 30 accuracy measures among Caucasians resulting from training our model on the first level of data augmentation, and `caucasian[1]` will contain the 30 accuracy measures among Caucasians resulting from training our model after we augmented the data more. To assess the similarity between racial groups, we will conduct one-way ANOVA tests.

5 Evaluation

We will plot the percentage of non-Caucasian representation against the ANOVA statistic for each augmentation technique. We anticipate that aligning proportions across different races more closely will improve the ANOVA statistic, signifying reduced differences in accuracy measures [4]. Yet, excessive data augmentation might lead to overfitting, potentially reducing accuracy for minority classes. Additionally, we'll present plots showing the impact of data augmentation on overall accuracy, as it's crucial for scientists to balance both overall accuracy and equity.

6 Potential societal and ethical risks

When attempting to reduce bias across one protected class, such as race, the question of why other protected classes are not corrected for arises. For example, this model may inequally represent those from certain income backgrounds and thus make inaccurate predictions on these groups. They may still be unfairly impacted even after the model is corrected for race, so it is important to consider the broader context of intersectionality and the interplay of various societal factors. Another social and ethical concern when attempting to address bias in models is the potential for unintended consequences. When making corrections to mitigate bias, there is a risk of introducing new biases or exacerbating existing ones. Striking the right balance between fairness and accuracy is a complex challenge, and model developers must carefully consider the trade-offs and potential unintended consequences to ensure that their efforts genuinely lead to fairer and more equitable outcomes.

References

[1] Alina Baciu, Yamrot Negussie, Amy Geller, James N Weinstein, Engineering National Academies of Sciences, Medicine, et al. The state of health disparities in the united states. In *Communities in action: Pathways to health equity*. National Academies Press (US), 2017.

[2] Ziad Obermeyer, Brennan Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[3] Vasileios Iosifidis and Eirini Ntoutsi. Dealing with bias via data augmentation in supervised learning scenarios. 2018.

[4] R. Bevans. One-way anova — when and how to use it (with examples). https://www.scribbr.com/statistics/one-way-anova/, June 22 2023. Accessed on October 26, 2023.