Oversampling on Another Attribute for Image Classification

Kim Yeonsu* YEON17@UNIST.AC.KR
Ulsan National Institute of Science and Technology, Korea

Kim Junmok* JM0917@UNIST.AC.KR
Ulsan National Institute of Science and Technology, Korea

Sim Soyeon* TLATHDUS@UNIST.AC.KR
Ulsan National Institute of Science and Technology, Korea

Abstract

This paper investigates the effect of oversampling on another attribute which is not goal class for image classification. Machine learning predictive modeling on unbalanced datasets is one of the important challenges. In particular, the effect of oversampling on the predictive accuracy of other properties when applied to one property is not clearly understood.

This study investigates the effect of oversampling on the interaction between characteristics on the CelebA dataset. When oversampling is applied to a particular attribute, we want to analyze the change in machine learning prediction accuracy of another attribute. It is expected that this will clarify the effect of oversampling on other properties and provide new insights in addressing the imbalance in the dataset.

Keywords: Data Imbalance, Oversampling, Data Fairness

1. Introduction

Modern machine learning technologies are actively utilized in a variety of fields, which greatly improves the accuracy and efficiency of data analysis and predictive modeling. In particular, predictive modeling on imbalanced datasets is emerging as one of the key challenges.

Oversampling is one of the ways to alleviate the problem of data imbalance by increasing minority class data. However, the effect of oversampling on another property when applied to one property is not clearly identified.

This study investigates the impact of oversampling techniques on specific properties using the dataset whose attributes are imbalanced. In particular, we want to analyze the impact of oversampling on machine learning prediction accuracy in one attribute. It is expected that this will enhance our understanding of interactions between properties and provide new insights in addressing imbalances in the dataset.

In this paper, we analyze the impact of oversampling based on CelebA data to evaluate the performance of machine learning models and interaction between characteristics of data.

This paper is organized as follows: Section 2 introduces related works, Section 3 introduces methods. Then, Section 4 introduces results, Section 5 discusses conclusion and discussion in Section 6.

2. Related Works

A number of studies have covered various oversampling and undersampling techniques. Specifically, There are many studies about oversampling techniques on the structured dataset. Mohammed, Rawashdeh, and Abdullah (2020) (1) outlined these techniques and covered the experimental results, which explored the effectiveness of oversampling and undersampling techniques using binary classification datasets in Kaggle. Gosain and Sardana (2017) (2) addressed the class imbalance problem using different classifiers to compare different oversampling techniques such as SMOTE, ADASYN, Borderline-SMOTE, Safe-Level SMOTE, and more. Zhu, Lin, and Liu (2017) (3) focus on the Multiclass imbalance problem, which suggests how SMOM weights each neighboring direction to perform effective oversampling on a multi-class imbalance dataset, generating a composite data of stable and reliable minority classes.

In the case of image data, however, these methods are not commonly used. For image data, data augmentation technique is used broadly. For deep

* These authors contributed equally

learning, Connor and Taghi (2019) (4) surveyed the various data augmentation technique for image and applied it for oversampling the minor class. Perez and Wang (2017) (5) studied about the effectiveness of data augmentation in image classification. Data augmentation method is very effective for reducing overfitting and generalization error. Wong and Gatt (2016) (6) also used the data augmentation for oversampling.

These various studies are being conducted in the direction of proposing new approaches related to oversampling in multi-class imbalance datasets or improving existing methods. However, there are many studies on oversampling techniques, but studies on the effect of oversampling on the accuracy of one class are insufficient.

3. Method

In Experiments 1, 2, and 3, we observe the variations in prediction accuracy based on the oversampling of the CelebA dataset, as well as changes in prediction accuracy between male and female data.

Each experiment involves oversampling different features, as detailed in Table 1.

The model employed for gender prediction is ResNet-18, with the output dimension of the fully connected layer adjusted to Table 3 for binary classification. For a certain feature, we additionally observe the imbalance in gender-specific predictions based on oversampling ratios (Experiment 3).

4. Dataset

CelebA (CELEB Faces Attributes) is a facial image dataset, which is one of the popular person image datasets. The dataset contains facial images of 10,177 celebrities, each of which is annotated as approximately 200,000 facial landmarks. The landmarks represent specific points on a face, which help explain each person's facial features.

The CelebA dataset labels more than 40 facial features for each image. Labels describe the properties of each face (e.g., the shape of the eyes, the size of the nose, the presence of a smile, etc.), and this label information is used for various computer vision tasks such as face recognition, feature extraction, image processing, etc.

The CelebA dataset provides a wide range of applicability and diversity, making it widely used for research and algorithm development in fields such as facial image recognition and feature extraction. It can also be used for applications such as image creation and modification, style transformation, and face property modification.

5. Experiment

5.1. Experiment 1

5.1.1. Dataset setting

First, we created a model that trains us to predict genders with the original dataset (corresponding to 200,000), and we found an accuracy that determines the probability of matching female/male. When we randomly selected 10,000 data to see if they were bald or not, we found a 97:3 ratio. In other words, there were 219 bald data and 9781 no-bald data. We decided to match the bald data with the number of no-bald data as the goal of data oversampling. Therefore, the augmented dataset, bal_augmented, is 9636. Since image augmentation (oversampling), the total number of images is 19,636. Finally, the ratio of bald data to no-bald data is 49:51, with a training set of 15,709, a valid set of 1964, and a test set of 1963.

5.1.2. Experiment

For the original dataset, CelebA, we have prepared a model trained to determine the gender. ResNet-18 was used, and the learning rate was set to 0.001, the epoch number was set to 10, and the batch size was set to 256. The test account to match the gender for the dataset when oversampling with the corresponding model came out at 96.51% of the time.

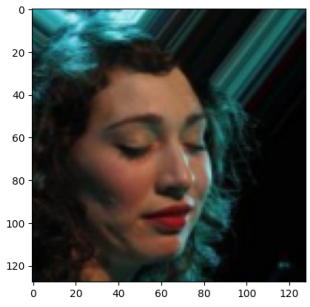

Figure 1: Example of Woman

For the example data above, the probability of not being bald is 99.97%, and the probability of being bald is 0.03

Experiment	Oversampled feature f	Minor label of f_{minor}	Proportion of $p(f = f_{\text{minor}})$	Gender ratio (Male:Female)
1	Bald	1	0.02281	0.99757:0.00243
2	Young	0	0.22106	0.68965:0.31035
3	Big Nose	1	0.23555	0.75078:0.24922

Table 1: Three distinct features, 'Bald,' 'Young,' and 'Big Nose,' exhibit imbalanced proportions within the original CelebA dataset including about 200,000 images , consistently demonstrating gender imbalances in the minority feature labels.

5.2. Experiment 2

5.2.1. Dataset Setting

The CelebA dataset has different proportions of the Old and Young attributes for Male and Female. Specifically, the Old ratio is higher among males, while the Young ratio is higher among females. To ensure accurate comparison, we adjusted the ratio of Old and Young in each gender to 3:1 and 1:3 and randomly sampled from entire CelebA dataset for training. As a result, we conducted experiments using datasets containing 6000 Old Male, 2000 Young Male, 2000 Old Female, and 6000 Young Female samples.

Additionally, we applied augmentation techniques to the images of Young Male and Old Female in train data, doubling their size. Consequently, the oversampled training dataset consists of 6000 Old Male, 4000 Young Male, 4000 Old Female, and 6000 Young Female samples. For the validation set during the training phase, the ratio was maintained at 3:1 and 1:3 similar to the original dataset. The test dataset comprises 1500 samples for each attribute to ensure equal representation of all attributes.

5.2.2. Experiment setting

We compared models trained using the original training dataset and the oversampled training dataset. The model used in the experiment is the ImageNet pretrained ResNet-18 model. We utilized SGD as the optimizer and employed cross-entropy loss as the loss function. There was no significant performance difference observed due to hyperparameter adjustments.

5.2.3. Gender Classification

Initially, we performed a task predicting the gender of images using both the original and oversampled training datasets. We then calculated the accuracy difference between images with Old and Young attributes within the Male and Female classes in the test dataset.

5.2.4. More Difficult Classification – Old

Furthermore, we wanted to evaluate performance on a slightly more challenging task than gender classification: Old classification. This task, compared to gender classification, has a significantly lower classification accuracy. Similarly, we conducted the task of predicting the Old class using both the original and oversampled training datasets. Then, we calculated the accuracy difference between images with Male and Female attributes within the Old and Young classes in the Test Dataset.

5.3. Experiment 3

5.3.1. Baseline model setting

For the task of gender classification from images, we employed the ResNet-18 model (torchvision.models.resnet18). The model's weights were initialized with pre-trained weights from ImageNet-1K dataset, achieving an accuracy of 69.758% on acc@1. To adapt the model for gender classification, we modified the output dimension of the last fully connected layer from 1000 to 2, and the weights of this layer were randomly initialized.

As a baseline model, we fine-tuned the partially pre-trained ResNet-18 model on the CelebA training dataset. However, to match the input dimensions of ResNet-18, we cropped the original CelebA dataset images.

The training process consisted of epoch=10, with a learning rate=0.001. We utilized the Stochastic Gradient Descent (SGD) optimizer with a momentum weight=0.9 (torch.optim.SGD). The chosen loss function for the task was the cross-entropy loss (torch.nn.CrossEntropyLoss).

Dataset	Male		Female	
	Old	Young	Old	Young
Original Train	6000	2000	2000	6000
Oversampled Train	6000	4000	4000	6000
Valid	1500	500	500	1500
Test	1500	1500	1500	1500

Table 2: Dataset Setting of Experiment 2

5.3.2. Experiment setting

To address the imbalance in certain features within the CelebA training dataset, oversampling was conducted. The selection of such features was informed by validating gender predictions on the validation dataset using the baseline model Table 3.

In Experiment 3, we generated training datasets with different random oversampling ratios specifically for data where 'Big_Nose' equals 1. The random oversampling ratios included 0.05, 0.1, 0.125, 0.15, 0.175, 0.2, and 0.225.

seven distinct models were trained by fine-tuning the baseline model with training datasets oversampled at varying ratios. The conditions applied during training, including epochs, learning rate, optimizer, and loss function, remained consistent with those used for training the baseline model.

6. Result

6.1. Result 1

For the original dataset and the oversampled dataset, the accuracy for man and woman were calculated respectively. For the oversampled result, the accuracy was 0.7272 for men and 0.0024 for women, and the result for the original dataset was 0.3717 for men and 0.0126 for women. The graph shows this as follows.

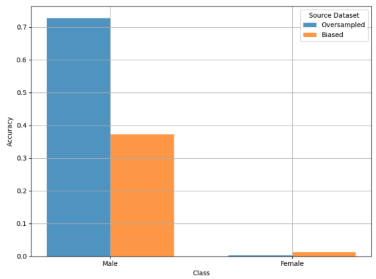

Figure 2: Accuracy of Male and Female Case

The oversampled data showed a much larger difference in accuracy, which we estimate is mostly due to the baldness data being male. This is because there are 4547 baldness data in the entire dataset, of which 4530 are male and bald.

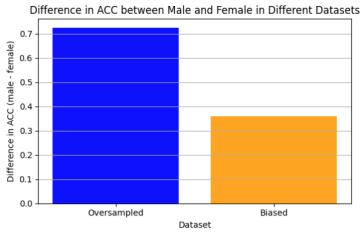

Figure 3: Difference in ACC between Male and Female in Different Datasets

6.2. Result 2

6.2.1. Gender Classification

In gender classification, difference in accuracy between old and young decreases in the male class about 1%p. In female class, however, the difference did not decrease. The accuracy of young male increased but accuracy of old female didn't.

6.2.2. More Difficult Classification - Old

Gender classification is very easy task with accuracy of about 96.4% in original dataset. However, old classification is hard task with accuracy of about 75.1% in original dataset. Therefore, we expected more extreme changes. In the experiment results, the difference in accuracy between Male and Female decreased by about 19%p in the Young class. This is a remarkable performance improvement. However, in the Old class, only a decrease of about 1%p was observed. Similar to before, while the accuracy significantly increased in Young Male, there was almost no improvement in performance for Old Female.

Experiment	Oversampled feature f	Minor label of f_{minor}	Proportion of $p(f = f_{\text{minor}})$	Probability of misclassification	
				Minor feature f	Major feature f
1	Bald	1	0.02068	0.00487	0.01758
2	Young	0	0.25343	0.02224	0.01564
3	Big_Nose	1	0.2488	0.01742	0.01699

Table 3: Predictions of the baseline model on CelebA validation dataset. The table presents the results of the baseline model's predictions on the CelebA validation dataset, where y_{label} represents the actual gender, and y_{pred} denotes the gender predicted by the model.

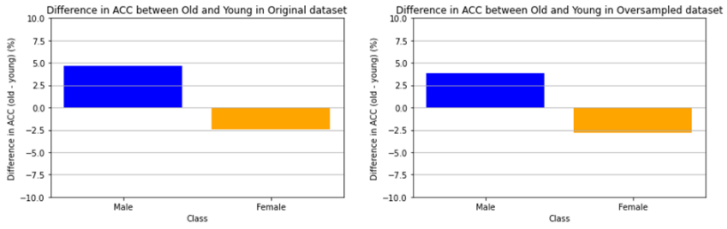

Figure 4: Difference in accuracy between Old and Young attribute for gender classification

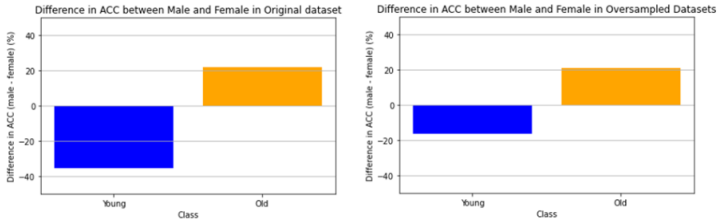

Figure 5: Difference in accuracy between Male and Female attribute for old classification

6.3. Result 3

6.3.1. Comparison of model performance based on oversampling in CelebA training dataset

Table 4 presents the results of comparing the prediction accuracy on the CelebA test set under different oversampling conditions for the 'Big_Nose' feature. When oversampling was not applied, the accuracy was 0.9778, which is 0.0019 higher than the baseline accuracy of 0.9759. The highest accuracy of 0.9779 was achieved when the oversampling ratio in the training dataset was 0.1. However, Table 5 reveals that, contrary to expectations, oversampling did not consistently lead to higher prediction accuracy compared to the scenario where oversampling was not applied.

6.3.2. Differences in prediction Accuracy between Male and Female in Some Cases

Next, the differences in prediction accuracy between male and female were compared for the scenarios of no oversampling, oversampling ratio of 0.1, and oversampling ratio of 0.225 in Table 5.

When the oversampling ratio was 0.1, the difference in predictive accuracy between male and female on the test dataset was 0.0143, which was the smallest, and simultaneously, the predictive accuracy was the highest at 0.9779. However, Table 5 indicates that when the oversampling ratio was 0.1, the predictive accuracy for females was 0.0935, which is 0.0003 lower than the baseline accuracy of 0.9838 for females, suggesting a partial trade-off in predictive performance. This observation could be attributed to the probability of misclassification based on the presence of the

	Baseline	Oversample						
Oversampling rate		0.05	0.1	0.125	0.15	0.175	0.2	0.225
Test accuracy	0.9759	0.9776	0.9779	0.9776	0.9762	0.9776	0.9773	0.9776

Table 4: Comparison of prediction accuracy on CelebA test dataset for different oversampling ratios of 'Big_Nose' feature.

	Rate	Total	Female	Male	Difference
Baseline		0.9759	0.9838	0.9633	0.0205
	0.1	0.9779	0.9835	0.9692	0.0143
	0.225	0.9776	0.9844	0.9667	0.0177

Table 5: Differences in prediction Accuracy between Male and Female on CelebA test dataset for different oversampling ratios of 'Big_Nose' Feature

'Big_Nose' feature, which, as seen in Table 3, did not show significant differences compared to the 'Bald' and 'Young' features.

6.3.3. Comparison of Gender Distribution in Misclassifications on CelebA Test Dataset

Furthermore, we compared the proportion of males and females in misclassifications on the CelebA test dataset for the scenarios of no oversampling, oversampling ratio of 0.1, and oversampling ratio of 0.225 in Table 6

When the oversampling ratio was 0.1, the difference in the proportion of males and females in misclassifications on the test dataset was the smallest, with a value of 0.1530. This suggests that appropriate oversampling ratios can help mitigate the disproportionate contribution of a specific gender to misclassified data.

7. Conclusion

Based on the analysis about Experiment 1 conducted, the oversampling technique significantly impacted the model's accuracy in predicting genders, particularly due to the augmentation of baldness data. This augmentation led to a considerable imbalance, where the majority of bald data aligned with male samples. Consequently, the oversampled dataset showcased a substantial difference in accuracy between genders, predominantly influenced by the augmented male baldness data. This emphasizes the crucial role of data augmentation methods in altering model performances, especially when specific attributes, like baldness, significantly correlate with gender identification.

In Experiment 2, We used oversampling techniques to address unfair learning resulting from the imbalance in other attributes when classifying one attribute between Gender and Old attributes. However, while there was reasonable fairness improvement for specific attributes, it was not observed for the other. This implies that adjusting data ratios for only one attribute has limitations in enhancing fairness.

The results from Experiment 3 demonstrate that an appropriate oversampling ratio can be beneficial in (1) increasing predictive accuracy and (2) mitigating the disproportionate contribution of a specific gender to misclassified data.

8. Discussion

The findings from Experiment 1 underscore the significant influence of oversampling, particularly the augmentation of baldness data, on gender prediction accuracy. This augmentation introduced a notable imbalance, primarily aligning baldness with male samples. Consequently, the oversampled dataset exhibited substantial gender-based accuracy differences, primarily driven by the augmented male baldness data. This highlights the dangerousness of data augmentation methods, especially when attributes like baldness strongly correlate with gender identification, in reshaping model performance.

Moving to Experiment 2, oversampling showcased varying impacts on gender and age attribute classifications within the CelebA dataset. While it notably

	Rate	Female	Male	Difference
Baseline		0.3814 (37)	0.6186 (60)	0.2372
	0.1	0.4235 (36)	0.5765 (49)	0.153
	0.225	0.3696 (34)	0.6304 (58)	0.2608

Table 6: Proportion of Males and Females in overall misclassifications on CelebA test dataset for different oversampling ratios of 'Big_Nose' feature. (The values in parentheses next to the ratios represent the corresponding data counts.)

reduced accuracy disparities between Old and Young attributes in Male samples, its effects were less consistent across Female samples. Although Gender classification tasks exhibited significant improvements, particularly in Young Male attributes, the enhancements were less pronounced in the Old Female attributes. These results emphasize the nuanced effects of oversampling across different attribute classifications, underscoring the need for tailored approaches to optimize predictive accuracy across diverse attributes in gender and age classification tasks.

Furthermore, Experiment 3 highlighted the potential benefits of using an appropriate oversampling ratio, indicating its potential to improve predictive accuracy while mitigating disproportionate contributions of specific genders to misclassified data. This suggests that careful consideration and optimization of oversampling strategies can play a crucial role in refining model performance and addressing imbalances within datasets, particularly in multi-attribute classification tasks like gender and age prediction.

References

[1] Mohammed, R., Rawashdeh, J., Abdullah, M. (2020, April). Machine learning with oversampling and undersampling techniques: overview study and experimental results. In 2020 11th international conference on information and communication systems (ICICS) (pp. 243-248). IEEE.

[2] Gosain, A., Sardana, S. (2017, September). Handling class imbalance problem using oversampling techniques: A review. In 2017 international conference on advances in computing, communications and informatics (ICACCI) (pp. 79-85). IEEE.

[3] Zhu, T., Lin, Y., Liu, Y. (2017). Synthetic minority oversampling technique for multiclass imbalance problems. Pattern Recognition, 72, 327-340.

[4] Shorten, C., Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. J Big Data 6, 60 (2019). https://doi.org/10.1186/s40537-019-0197-0

[5] Perez, L., Wang, J. (2017). The Effectiveness of Data Augmentation in Image Classification using Deep Learning. https://doi.org/10.48550/arXiv.1712.04621

[6] S. C. Wong, A. Gatt, V. Stamatescu and M. D. McDonnell, "Understanding Data Augmentation for Classification: When to Warp?," 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Gold Coast, QLD, Australia, 2016, pp. 1-6, doi: 10.1109/DICTA.2016.7797091.