UPPSALA
UNIVERSITET

MASTER THESIS SPECIFICATION

Mitigating Representation Bias in Extremely Imbalanced Data via Conditional Generative Models

Student:
Chenglong Li
chenglong.li.4328@student.uu.se

Supervisor:
Andrey Shternshis
andrey.shternshis@it.uu.se

Subject Reviewer:
Lovisa Eriksson
lovisa.eriksson@it.uu.se

Department of Information Technology

December 13, 2024

Contents

1 Title

The primary objective of this thesis project is to alleviate representation bias in datasets with extreme class imbalance. The approach involves leveraging conditional generative models, such as Conditional Variational Autoencoders (cVAE) and Conditional Generative Adversarial Networks (cGAN), among others. Hence, the thesis is titled **Mitigating Representation Bias in Extremely Imbalanced Data via Conditional Generative Models**.

2 Abstract

Addressing representation bias in extremely imbalanced datasets is crucial in fields like medical diagnostics and fraud detection, where accurately identifying minority classes is essential for decision-making. Although conditional generative models have recently been widely used for generating new images, their application to solving the problem of extreme class imbalance remains underexplored, particularly in the context of data augmentation and generating minority class instances. The key to addressing dataset imbalance using conditional generative models lies in generating representative minority class samples, thereby increasing their data volume, reducing model bias towards majority classes, and enhancing the classifier's ability to recognize minority classes. This thesis aims to demonstrate the effectiveness of the model in balancing datasets and improving classification accuracy through experiments using Conditional Variational Autoencoders (cVAE) on MNIST and CIFAR-10 datasets.

3 Background

In real-world problems, data collection often faces the issue of class imbalance, where certain classes have significantly fewer samples than others. For instance, in medical diagnostics [1], patients requiring special attention are often much fewer than regular patients; in insurance fraud detection [2], the number of normal customers is more than that of fraudulent customers. Similar situations are also observed in banking risk analysis and manufacturing defect detection. In such unbalanced datasets, differences in sample size can lead to suboptimal performance of traditional machine learning models, often biasing them towards the majority class and making it difficult to identify the minority class. This bias can lead to discrimination against minority groups, and the cost of misclassifying minority group samples is often significantly

higher than that of misclassifying majority group samples. Therefore, special attention must be given to the identification of minority groups.

To address the challenges faced by machine learning models when learning from imbalanced datasets, several effective methods have been developed. Firstly, in the data preprocessing stage, different sampling techniques, such as over-sampling [3] and under-sampling [4], can be employed to obtain a relatively balanced subset. Over-sampling involves increasing the number of minority class samples while under-sampling reduces the number of majority class samples to achieve balance. Secondly, data augmentation techniques, such as data transformation, Conditional Variational Autoencoders (cVAE) [5], Conditional Generative Adversarial Networks (cGAN) [6], or Diffusion Models [7], can be used to generate new samples, thereby increasing the number of minority class samples. In addition, during model training, different class weights can be assigned to emphasize the importance of minority class samples, thereby mitigating bias. Finally, ensemble learning methods [8] are also widely used to alleviate the issues arising from imbalanced datasets, as ensemble models combine the predictions of multiple classifiers, enhancing the recognition capability for minority class samples. By combining these approaches, the model's performance on unbalanced datasets can be improved, leading to better identification of minority samples.

4 Description of Tasks

When collecting data from different classes, imbalances often occur, leading to unfair data distribution among classes. During machine learning model training, this imbalance may result in unfair outcomes, such as the model focusing disproportionately on the majority class and discriminating against the minority class. The main goal of this project is to use conditional generative models to generate new instances for the minority class, balance the dataset, and reduce bias against underrepresented classes, thereby mitigating these unfair effects.

Tasks of this project are:

- Test the suitability of conditional generative models for over-sampling in extremely imbalanced datasets.
- Propose techniques to enhance model performance in generating new samples.
- Verify if the generated samples improve classification accuracy.
- Compare with state-of-the-art methods for over-sampling.

5 Work Description

This thesis has two main goals. The first goal is to generate samples for minority classes using a generative model. The second goal is to propose methods to improve the performance of these models. This thesis will focus on conditional generative models, particularly Conditional Variational Autoencoders (cVAE). In addition, other high-performing variations of cVAE will be explored to enhance the model's effectiveness.

To understand how data imbalance affects the performance of generative models, we will use the imbalance ratio (IR) [8] as a measurement. The cVAE will be trained multiple times to evaluate the quality of the generated samples, using evaluation metrics such as Fréchet Inception Distance (FID) and Inception Score (IS). To better visualize the generated data, experiments will be conducted on the MNIST and CIFAR-10 datasets.

The ultimate goal of this thesis is to mitigate the negative impact of imbalanced datasets on machine learning models. To achieve this, baseline classifiers will be trained on both the balanced dataset created by our model and the original imbalanced dataset. By comparing the classification results, we aim to assess the effectiveness of conditional generative models in addressing the issues caused by data imbalance.

UU will provide a workplace and sufficient computational resources(GPUs). The project will be implemented using Python.

6 Relevant Courses

- 1MS041 - Introduction to Data Science
- 1RT700 - Statistical Machine Learning
- 1MS047 - Theoretical Foundations for Data Science
- 1MS049 - Computer-Intensive Statistics and Applications
- 1DL508 - Project in Data Science

7 Delimitations

The objective of this project is to employ conditional generative models to generate samples for minority classes, thereby balancing the original dataset. The evalua-

tion will involve training classification models on the balanced dataset to assess the performance of the conditional generative models. The primary focus is not on enhancing the classification models' predictive capabilities but rather on evaluating the generative models' effectiveness. The evaluation metrics will emphasize the extent of dataset balance and the quality of the generated data, providing a comprehensive evaluation of the conditional generative model's ability to address class imbalance issues.

8 Time Plan

The project starts on January 19, 2025, and will take approximately 100 days to complete. Below is the estimated time required for each section:

- Literature Review and Model Selection (19/01/2025 - 01/02/2025)
  - Read Papers (10 days)
    - Dataset (MNIST, CIFAR-10)
    - Conditional Generative Model
    - Classifer (NN, SVM)
    - Evaluation metrics
  - Write - Related Work (3 days)
- Data Preparation and Pre-processing (02/02/2025 - 15/02/2025)
  - Data Pre-processing: Generate some datasets of imbalance levels (10 days)
  - Write - Experiments (Introduction to Dataset) (3 days)
- Model Learning (16/02/2025 - 15/03/2025)
  - Learn Model (7 days)
    - Our main method (Conditional Generative Model)
    - Other method (Comparison)
  - Write the model's code (18 days)
    - Conditional Generative Model
    - Other method
  - Write - Method (5 days)

- Experiments and Results Analysis (16/03/2025 - 19/04/2025)
  - Experiments (14 days)
    - Conditional Generative Model
    - Other Method
  - Results Analysis (14 days)
    - Dataset: Impact of different methods on the dataset.
    - Samples: The quality of the samples generated by the evaluation.
    - Classifier Performance: NN, SVM
  - Write - Experiment and Analysis (7 days)
- Write Master Thesis (20/04/2025 - 03/05/2025)
  - Look back at the thesis and reorganize (10 days)
  - Write - Conclusion (4 days)
- Prepare Presentation (04/05/2025 - 17/05/2025)
  - Prepare slides for presentation (7 days)
  - Practice presentation (7 days)

References

[1] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," *Artificial Intelligence Review*, vol. 57, no. 10, p. 273, Sep. 2024. [Online]. Available: https://doi.org/10.1007/s10462-024-10884-2

[2] F. Aslam, A. I. Hunjra, Z. Ftiti, W. Louhichi, and T. Shams, "Insurance fraud detection: Evidence from artificial intelligence and machine learning," *Research in International Business and Finance*, vol. 62, p. 101744, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0275531922001325

[3] P. Gnip, L. Vokorokos, and P. Drotár, "Selective oversampling approach for strongly imbalanced data," *PeerJ Computer Science*, vol. 7, p. e604, 2021. [Online]. Available: https://doi.org/10.7717/peerj-cs.604

[4] M. Bach, A. Werner, and M. Palt, "The proposal of undersampling method for learning from imbalanced datasets," *Procedia Computer Science*, vol. 159, pp. 125–134, 2019, knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 23rd International Conference KES2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050919313456

[5] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015.

[6] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014. [Online]. Available: https://arxiv.org/abs/1411.1784

[7] M. Chen, S. Mei, J. Fan, and M. Wang, "An overview of diffusion models: Applications, guided generation, statistical rates and optimization," 2024. [Online]. Available: https://arxiv.org/abs/2404.07771

[8] A. A. Khan, O. Chaudhari, and R. Chandra, "A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation," *Expert Systems with Applications*, vol. 244, p. 122778, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423032803

Signature page

This document has been electronically signed using eduSign.

eduSign

Electronically signed by
Andrey Shternshis
Date and time of signature
2024-12-13 12:25 UTC
Authenticated by
Uppsala University

Electronically signed by
Lovisa Eriksson
Date and time of signature
2024-12-16 09:08 UTC
Authenticated by
Uppsala universitet

Electronically signed by
Chenglong Li
Date and time of signature
2024-12-16 09:12 UTC
Authenticated by
Uppsala universitet