OVERCOMING IMBALANCED AND OVERLAPPING DATA IN MULTICLASS CLASSIFICATION

DESSY ROTUA NATALINA SIAHAAN

STATISTICS AND DATA SCIENCE
FACULTY OF MATHEMATICS AND NATURAL SCIENCE
IPB UNIVERSITY
BOGOR
2024

STATEMENT OF THESIS AUTHENTICITY, INFORMATION SOURCE, AND COPYRIGHT TRANSFER AGREEMENT

I declare that the thesis entitled "Overcoming Imbalanced and Overlapping Data in Multiclass Classification" is my work under the direction of my supervisors and has not been submitted in any form to any university. Sources of information derived or quoted from published and unpublished works of other authors have been mentioned in the text and included in the Bibliography at the end of this thesis.

I assign the copyright of my writing to the IPB University.

Bogor, June 2024

Dessy Rotua Natalina Siahaan
G1501211027

SUMMARY

DESSY ROTUA NATALINA SIAHAAN. Overcoming Imbalanced and Overlapping Data in Multiclass Classification. Supervised by ANWAR FITRIANTO and KHAIRIL ANWAR NOTODIPUTRO.

Classification is a predictive model that groups data based on algorithms with categorical response variables. In binary classification, there are two class categories, while in multiclass classification, there are three or more class categories. Multiclass classification is more challenging than binary classification because of complex interaction patterns. This problem becomes even more complicated if the data contains imbalanced data issues, meaning that some classes have many observations while others have fewer. Complexity increases if the data to be modeled also overlaps because no boundaries separate the existing categories, making the classification process difficult.

This research aims to address the problem of imbalanced and overlapping data in multiclass classification. To achieve this, research was conducted on simulated and empirical data. The simulated data consisted of 10,000 observations with four predictor variables, including two numerical and categorical variables, and a response variable with three class categories. Overlap conditions are controlled using three distance scenarios (High, Medium, Low) calculated based on Euclidean distance. The distance determination is used as each class' middle value (centroid) for generating data that follows a normal distribution. The data generation process for each class controls imbalanced conditions with six scenarios based on data proportions (High, Medium, Low) and the number of minority classes. The combination of all these conditions produces 18 simulation scenarios.

Multiclass classification is encountered in everyday life, making the problem of imbalanced and overlapping data increasingly urgent to be handled appropriately. Poverty is used as a case study for overcoming imbalanced and overlapping data. Poverty is a classification problem where the data has to be integrated with the focus object (Poor) being in the minority class. The poverty status develops from binary (Poor and Not Poor) to multiclass with the addition of Extremely Poor. Overlapping variables were also found, for example, two households with heads of household with the same gender, age, and education level but different poverty statuses. The Secondary Data from the National Socio-Economic Survey (SUSENAS, *Survei Sosial Ekonomi Nasional*) in West Java in 2021 by Statistics Indonesia (BPS, *Badan Pusat Statistik*) is used, where the poverty status category is generated from the average household expenditure in a month.

The One Versus One (OVO) decomposition technique is used to simplify multiclass classification by transforming it into binary subclasses so that the binary classifier method can be applied. The final prediction results are determined by Majority Voting. The problem of imbalanced data is overcome by applying the Synthetic Minority Oversampling Technique (SMOTE) method, which adds data to the minority class by generating new synthetic data. A combination of two classification methods, a Multiple Classifier System with sequential combination, is used to overcome the problem of overlapping data. The single classifier combined is the Logistic Regression and K-Nearest Neighbour (KNN) method. The prediction

output results from the Logistic Regression method are input into the classification process using the KNN method. Combining all methods is applied to all simulation scenarios and empirical data.

The evaluation model measurements are Balanced Accuracy, Weighted F1 Score, and G-Mean. The results of the simulated data show that the proposed model is able to overcome the problem of imbalanced and overlapping data in multiclass classification. Based on all simulation scenarios, the model is able to provide satisfactory performance in both the overall model and in classifying minority classes. The model with the lowest performance is in the High-2 scenario, where the amount of data for the majority class and minority class has a huge difference and has two minority classes. The model performance works better when the imbalance proportion decreases, and there is only one minority class. Model performance is also influenced by distance scenarios that represent the degree of overlap. The High-distance scenario has the lowest performance where the data is highly overlapping. As the distance increases, the model performance improves because the data overlap decreases.

Model comparison is done to ensure that the performance of the proposed model is better than the others. The models used for comparison are the Logistic Regression Model with SMOTE, the KNN Model with SMOTE, the MCS Model without SMOTE, the Logistic Regression Model without SMOTE, and the KNN Model without SMOTE. Model comparison is carried out by comparing existing evaluation measures. As a result, the proposed model gave the best and superior performance compared to the other five models; some models were even unable to classify the minority class well, so it was found that the F1 Score and G-Mean evaluation measures produced 0.

The results of poverty data classification also show good performance. The Balanced Accuracy obtains 80%, and the F1 Score is 0.8. Meanwhile, the G-Mean value needs to be optimized; it only gets 0.3 because the poverty data is similar to the High-2 Scenario with a High distance. This simulation scenario has the lowest performance among the other scenarios because of the high level of imbalance and overlap. However, other models are entirely unable to classify minority classes.

Keywords: Imbalanced Data, Multiclass, Multiple Classifier System, Overlapping Data, Simulation

OVERCOMING IMBALANCED AND OVERLAPPING DATA IN MULTICLASS CLASSIFICATION

DESSY ROTUA NATALINA SIAHAAN

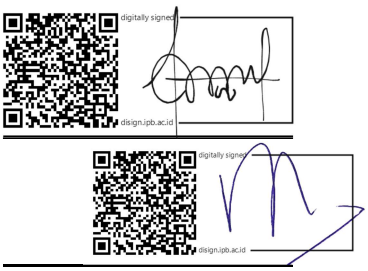
Thesis
as a partial fulfillment for the degree of
Master of Science
in
Statistics and Data Science Program

STATISTICS AND DATA SCIENCE
FACULTY OF MATHEMATICS AND NATURAL SCIENCE
IPB UNIVERSITY
BOGOR
2024

Examiner Committee:
Dr. Azka Ubaidillah M.Si.

Thesis Title : Overcoming Imbalanced and Overlapping Data in Multiclass Classification
Name : Dessy Rotua Natalina Siahaan
Student ID : G1501211027

Approved by:



Main Supervisor:
Dr. Anwar Fitrianto, M.Sc.

Co-Supervisor:
Prof. Dr. Khairil Anwar Notodiputro, M.S.

Acknowledged by:

Head of Statistics and Data Science Study Program:
Dr. Agus Mohamad Soleh, S.Si., MT.
NIP. 19750315 199903 1 004

Dean of Faculty of Mathematics and Natural Sciences:
Dr. Berry Juliandi, S.Si., M.Si.
NIP. 19780723 200701 1 001

Examination date: 26 June 2024 Graduation date:

ACKNOWLEDGMENT

I'd like to express my gratitude to the Lord Jesus Christ for His unwavering love, which has enabled me to successfully complete my thesis research. The research, conducted from July 2022 to November 2023, focuses on the challenges encountered in the classification process, with the title 'Overcoming Imbalanced and Overlapping Data in Multiclass Classification.

I am thankful to my supervisors, Dr. Anwar Fitrianto, M.Sc, and Prof. Dr. Khairil Anwar Notodiputro, M.S, for their guidance and valuable suggestions. My sincere appreciation to Dr. Azka Ubaidillah, M.Si for serving as the examiner and providing a comprehensive evaluation of this research. I would also like to extend my appreciation to The Indonesia Endowment Funds for Education (LPDP, *Lembaga Pengelola Dana Pendidikan*) for their support from the beginning of my studies until the completion of this paper and my master's degree.

Special thanks to my beloved mother, Dr. Ida Mariati Hutabarat, M.Si, for her motivation, as well as my younger siblings, Fajar Josua Siahaan and Ray Joy Hasian Siahaan, who inspire me to be a role model for them. I am also grateful to my lovely man, Manuel Septiyanto Marbun, S.Si, M.B.A, who always provided support, encouragement, and even daily assistance so that I was able to pass this.

I would like to acknowledge my extended family, especially Uda Lambok Siahaan, Inanguda Mercyrene Simanjuntak, Kezia Siahaan, and Feodora Siahaan for their continuous support and understanding. A heartfelt thanks to my best friend, Apnesia Feronika Nainggolan, S.Si, for her accompanied and assistance at every opportunity. I am thankful to my friends in the Master of Statistics and Data Science class of 2021 for the insightful discussions we have had. I hope we can all fulfill our responsibilities until the end. Also, a big thank you to the Association of Postgraduate Statistics Professionals (HIMPRO, *Himpunan Profesi*) in 2021/2022 and 2022/2023 for adding colour to my academic journey.

I acknowledge that there may be mistakes and shortcomings in this research, and I apologize for them. I hope this paper will be beneficial to those who need it and contribute to the advancement of science. May the Lord be with us all.

Bogor, June 2024

Dessy Rotua Natalina Siahaan

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDIX