UNIVERSITI TEKNOLOGI MARA

DISTANCE-BASED UNDERSAMPLING FOR IMBALANCE DATASET: A COMPREHENSIVE SIMULATION STUDY

HEZLIN ARYANI BINTI ABD RAHMAN

Thesis submitted in fulfilment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)

College of Computing, Informatics and Mathematics

August 2023

ABSTRACT

This thesis presents a simulation study on parameter estimation for the effect of imbalance problem in binary logistic regression. As well as assessing the effect of random oversampling (ROS), random undersampling (RUS), and distance-based undersampling (E-DBUS, iE-DBUS and M-DBUS) on imbalanced data of the binary logistic regression. This study obtained the threshold for imbalance ratio and size of sample, which are affected by the presence of imbalanced. The motivation behind this study is influenced by three main factors. Firstly, imbalanced problem normally effects the accuracy in predictive models, especially in data mining and machine learning models. However, most classification focusses on other classifiers, and not much on binary logistic regression. Secondly, there is a lack of focus in studies of imbalance involving simulation studies in the area of imbalanced data, especially in visualization of the effect of imbalanced on classifiers, in this case, binary logistic regression. Thirdly, resampling strategies are a more straight-forward approach to handling imbalanced data. However, this strategy has been under-rated and various studies suggested various resampling strategies to better handle imbalance dataset. Distanced-based sampling has shown positive impact on imbalanced data. Hence, proving the positive effect on binary logistic regression is the quest for this study. Simulation studies are useful to assess and confirm the effects of parameter estimation for binary logistic regression under various conditions. The first phase of this study covers the effect of different types of covariates, imbalance ratio and sample size on parameter estimation for binary logistic regression model. Data were simulated for different sample sizes, types of covariates (continuous and categorical) and imbalance ratio. The simulation results show that the effect imbalance problem is more prominent in smaller sample sizes ($n < 2000$) and highly imbalanced data *(IR* < 10%). The effect reduces as sample size increases and data became more balanced. The effect of imbalanced were more dominant for categorical covariates compared to continuous covariates. In Phase 2, the effect of the ROS and RUS were assessed for imbalanced datasets on parameter estimation of binary logistic regression. Results shows that the ROS has better performance in curbing the effect of imbalanced compared to RUS on all different sample sizes and imbalance ratio on various types of covariates; continuous, categorical, and mixture of both, due to the doubled in the number of sample size. However, random synthetisation of observations was unfavourable, especially in statistics. Thus, in Phase 3, the simulation focused on the RUS and the distanced-based undersampling strategies in handling the effects of imbalanced datasets on parameter estimation of binary logistic regression for one continuous covariate. Comparing the results in Phase 1-3, the distance-based undersampling, either Euclidean (E-DBUS), Mahalanobis (M-DBUS) or improved-Euclidian (iE-DBUS), - based undersampling, were more reliable in curbing the effect of imbalanced problem as compared to ROS and RUS. Further, in phase 4 (evaluation), the performance of all random and the three distanced-based undersampling (E-DBUS, iE-DBUS, and M-DBUS) were investigated using 14 benchmark datasets studies, comparing the accuracy, sensitivity and specificity of the binary logistic regression model. The results showed that the M-DBUS performed the best compared to the other undersampling strategies. However, the difference in terms of performance were not far compared from E-DBUS and iE-DBUS. The significance of this study will benefit the body of knowledge of statistics and predictive data analytics, especially in the area of imbalanced data handling.

ACKNOWLEDGEMENTS

In the Name of Allah, the All-Compassionate, the All-Merciful

Alhamdulillah, all praise to Allah S.W.T for making it possible for me, in every way, to complete this journey. I would like to express my utmost gratitude towards my supervisor Prof. Dr. Yap Bee Wah whom has been my guardian angel since the beginning of my studies. I sincerely appreciate her continuous guidance, positive encouragement, and constructive criticisms. Her meticulousness to details and exemplary passion for research has inspired me to venture in this field that I once have no knowledge of. My gratitude also goes to my co-supervisor, Prof. Dr. Haibo He. He made it clear that distance is not an obstacle for research and supervision. I am also deeply indebted to Prof. Dr. Daud for his indirect guidance throughout my studies. Most indebted is to my current supervisor, Associate Professor Dr. Ahmad Zia Ul-Saufie Mohamad Japeri., for without him, I might not complete this PhD journey alone.

I would also like to thank Dr. Anne Porter, Dr. Pam Davey and Dr. Carolle Birrell from the University of Wollongong, Australia for giving a different perspective to my simulation studies. My gratitude also goes to the current dean, Prof. Dr. Haryani Haron, the mantan dean of FSKM, Prof. Azlinah Mohamed, and also the Head of Center of Statistical and Decision Science Studies, PM Dr. Sayang Mohd Deni, for their continuous encouragement and support. I would like to acknowledge the Ministry of Higher Education and Universiti Teknologi MARA for the research grant provided. I would also like to acknowledge the Head of IPSIS FSKM UiTM, Dr. Fuziah Ishak and her team. They are very cooperative in every way.

Last but not least my gratitude to my parents PM Haji Abd Rahman Jantan and for providing me the best education since the beginning and remain my inspiration throughout my entire life. Not to forget, my Ibu, that has indirectly supported my journey since day one. Special thanks to my beloved husband Mohd Ali Yusni bin Abu Samah for his never-ending support and abundance of love that kept me progressing each day. To my children, Muhammad Alif Hazimi, Nur Aimy Hamizah, Muhammad Ariq Harraz and Nur Adreena Hasya, although things certainly got tough, your love have made me the iron lady I became. To my DivaZumbaholic sisters, which has been my shoulder to cry on since we met, my gratitude knows no boundaries. Not to forget, my supervisory siblings, Dr Hamzah Abdul Hamid and Dr. Ainur Amira Kamaruddin, knowing both of you has made my sort-of never-ending PhD journey so much colourful. Also, to my sibings from different parents, Nur Maizura Lin, Zuraida Khairudin, Sharina Salmi Azmi, Norbaizura Kamarudin, Norulhidayah Md Isa, and Mohd Razif Shamsudin, thank you for supporting me since day one. Not least, my gratitude goes to everyone that has not been mentioned but has contributed in any way to any part of this thesis.

I thank you all with doa, Jazaka-allahu Khayran (May Allah bless you with good).

Al-Fatihah for those losing the battle in study and life.

Al-Fatihah for my Mama, Almarhumah Norhayati binti Mohd Taib.

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1.1 Background of Study

Recent developments in science and technology, especially in the technology involving database and data warehousing, has led to the evolution of data storage and the explosion of the availability of voluminous data or so-called the Big Data Challenge. This has created a gold mine opportunity for data scientist, and researchers' involvement in applying the data mining and statistical methodologies to a wide range of applications, from social science studies in finance, business, and education to more critical fields such as epidemiology, pharmaceutical, disaster prevention and plague detection.

One of the biggest challenges of Big Data Analytics is the issue of imbalanced data sets (IDS). It was first highlighted as a major problem-causing issue in the First Conference of the Association of Advancement in Artificial Intelligence in the year 2000. Since then, the *imbalanced learning* problem has gained significant interest among many data scientists and academia, resulting in various algorithm development and enhancement techniques for practical applications. The main goal is to find the solution to improve classification involving imbalanced datasets. Profoundly, the key issue when dealing with IDS is that the ability of most standard machine learning and data mining algorithms to significantly predict an outcome is compromised due to the characteristics of the IDS itself. This is due to the fact that the standard machine learning and data mining algorithms were developed without the issue of IDS in mind.

The study on IDS, not only present a new challenge to the data scientist community, but also raised many critically challenging queries towards the implementation of data mining techniques and machine learning algorithms applied in critical fields such as mentioned above. The interest of many data scientists towards the issue of IDS can be immensely reflected in the number of uprising publications in recent years, especially review articles (Weiss & Provost, 2003; Batuwita & Palade, 2013; Bekkar & Alitouche, 2013; Dianah et al., 2022; Fernández et al., 2017; Galar et al., 2011; Ganganwar, 2012; He & Garcia, 2009; Kaur et al., 2019; Kotsiantis, 2006; Longadge et al., 2013; Ramyachitra & Manikandan, 2014; Visa & Ralescu, 2005).