Paper ID – 456

A Comprehensive Evaluation of Sampling Techniques in Addressing Class Imbalance Across Diverse Datasets

Md. Salman Mohosheu[1], MD. Abdullah al Noman[2], Asif Newaz[3] , Al-Amin[4] and Dr. Taskeed Jabid[5]
[1]Department of Electrical and Electronic Engineering, Islamic University of Technology, Gazipur, Bangladesh
[2, 3,4]Department of Electrical and Electronic Engineering, Islamic University of Technology, Gazipur, Bangladesh
[5]Department of Computer Science & Engineering, East West University, Dhaka, Bangladesh
[1]salmanmohosheu@iut-dhaka.edu , [2]alnoman6@iut-dhaka.edu, [3]eee.asifnewaz@iut-dhaka.edu, [4]al-amin@iut-dhaka.edu and [5]taskeed@ewubd.edu

Abstract: Class imbalance is a frequently occurring issue in predictive modeling. Learning from imbalanced data is a challenging task that has attracted much interest from scholars. While a substantial amount of research has been conducted to develop a plethora of approaches, only a limited number of studies have analyzed the efficacy of these techniques on a wide variety of imbalanced datasets. This study aims to fill that research gap by providing a comprehensive experimental analysis of the effectiveness of a wide range of techniques. We investigated the performance of 30 different state-of-the-art approaches used in imbalanced learning. The performance was evaluated on 84 different imbalanced datasets with varied imbalance ratios. One of the major observations of the study is that although some techniques, especially the undersampling techniques, are quite effective in handling smaller imbalances, they fail to shift the bias in the case of larger imbalances resulting in poor performance. Even the ensemble techniques, where sampling algorithms are incorporated to boost performance, do not perform well. Some of the oversampling techniques such as LEE, SMOBD, or hybrid sampling techniques such as SMOTE-ENN perform comparatively better in such scenarios. Through experimentation, we identify the best-performing techniques for different imbalanced cases. Moreover, we identify the limitations of the techniques and discuss their performance in detail in this manuscript. We believe that this study will provide great insight into these sampling techniques and will pave the way to developing new, more reliable approaches that can surmount the shortcomings of the current strategies.