Statistica Sinica **36** (2026), 1-20
doi:https://doi.org/10.5705/ss.202024.0012

OPTIMAL MODEL AVERAGING FOR IMBALANCED CLASSIFICATION

Ze Chen[1], Jun Liao[2], Wangli Xu[2] and Yuhong Yang*[3]

[1]*Shandong University,* [2]*Renmin University of China and* [3]*Tsinghua University*

Abstract: Imbalanced data with a high-dimensional input has been widely encountered in many areas of applications. In this situation, it usually becomes essential to reduce redundant variables via model selection to improve the classification performance. However, with a large number of variables, model selection uncertainty is typically very high. To deal with this problem, we present a feasible model averaging procedure based on a cost-sensitive support vector machine (CSSVM) coupled with a cost-sensitive data-driven weight choice criterion for imbalanced classification. Theoretical justifications are provided in two distinct scenarios. When the data exhibits a weak imbalance, we derive a relatively fast uniform convergence rate of the CSSVM solution. In contrast, when the data possesses a strong imbalance, the convergence rate becomes much slower. In both scenarios, an asymptotic optimality of the proposed model averaging approach in the sense of minimizing the out-of-sample hinge loss is established. Moreover, to reduce the computational burden imposed by a large number of candidate models for model averaging, we develop the CSSVM with an L_1-norm penalty to prepare candidate models. Numerical analysis shows the superiority of the proposed model averaging procedure over existing imbalanced classification methods.

Key words and phrases: Asymptotic optimality, imbalanced data, model averaging, uniform convergence rate.

1. Introduction

1.1. Imbalanced classification problems

In binary classification, imbalanced data, sometimes called rare event data, occurs when the number of instances of a certain class (the minority class) is significantly smaller than the number of instances of the opposite class (the majority class). The imbalanced classification problems (ICPs) were identified by Yang and Wu (2006) as one of ten challenging problems in data mining research. ICPs have shown up in many real-world applications, such as medical science (Rahman and Davis, 2013), telecommunications (Babu and Ananthanarayanan, 2018), and bioinformatics (Bugnon et al., 2019).

*Corresponding author. E-mail: yyangsc@tsinghua.edu.cn