Statistica Sinica **36** (2026), 1-18
doi:https://doi.org/10.5705/ss.202023.0434

BALANCED SUBSAMPLING FOR BIG DATA WITH CATEGORICAL PREDICTORS

Lin Wang

Purdue University

Abstract: Supervised learning under measurement constraints is a common challenge in statistical and machine learning. In many applications, despite extensive design points, acquiring responses for all points is often impractical due to resource limitations. Subsampling algorithms offer a solution by selecting a subset from the design points for observing the response. Existing subsampling methods primarily assume numerical predictors, neglecting the prevalent occurrence of big data with categorical predictors across various disciplines. This paper proposes a novel balanced subsampling approach tailored for data with categorical predictors. A balanced subsample significantly reduces the cost of observing the response and possesses three desired merits. First, it is nonsingular and, therefore, allows linear regression with all dummy variables encoded from categorical predictors. Second, it offers optimal parameter estimation by minimizing the generalized variance of the estimated parameters. Third, it allows robust prediction in the sense of minimizing the worst-case prediction error. We demonstrate the superiority of balanced subsampling over existing methods through extensive simulation studies and a real-world application.

Key words and phrases: Data labeling, D-optimality, experimental design, orthogonal array, robust prediction.

1. Introduction

Supervised learning under measurement constraints is a common challenge in statistical and machine learning (Wang, Yu and Singh, 2017; Meng et al., 2021). In many applications, despite the availability of extensive predictor observations (design points), acquiring the observations of the response variable for all design points is frequently impractical due to resource limitations. For example, consider a scenario in healthcare where researchers aim to develop a predictive model for patient outcomes based on a diverse set of health-related predictors. A large set of predictor observations, such as patient demographics, medical history, and genetic information, is readily available. However, obtaining the corresponding response variable, such as a medical condition or treatment outcome, may involve invasive procedures or expensive diagnostic tests. Given the constraints of limited resources, observing the response of every individual in the dataset becomes

*Corresponding author. E-mail: linwang@purdue.edu