Statistica Sinica **35** (2025), 1-21
doi:https://doi.org/10.5705/ss.202023.0124

LEVERAGE CLASSIFIER: ANOTHER LOOK AT SUPPORT VECTOR MACHINE

Yixin Han[1], Jun Yu[2], Nan Zhang[3], Cheng Meng[4], Ping Ma*[5], Wenxuan Zhong[5] and Changliang Zou[1]

[1]*Nankai University,* [2]*Beijing Institute of Technology,*
[3]*Fudan University,* [4]*Renmin University and* [5]*University of Georgia*

Abstract: Support vector machine (SVM) is a popular classifier known for accuracy, flexibility, and robustness. However, its intensive computation has hindered its application to large-scale datasets. In this paper, we propose a new optimal leverage classifier based on linear SVM under a nonseparable setting. Our classifier aims to select an informative subset of the training sample to reduce data size, enabling efficient computation while maintaining high accuracy. We take a novel view of SVM under the general subsampling framework and rigorously investigate the statistical properties. We propose a two-step subsampling procedure consisting of a pilot estimation of the optimal subsampling probabilities and a subsampling step to construct the classifier. We develop a new Bahadur representation of the SVM coefficients and derive unconditional asymptotic distribution and optimal subsampling probabilities without giving the full sample. Numerical results demonstrate that our classifiers outperform the existing methods in terms of estimation, computation, and prediction.

Keywords and phrases: Classification, large-scale dataset, martingale, optimal subsampling, support vector machine.

1. Introduction

Consider the binary classification problem for a training sample of size N, $\mathcal{D}_N = \{(\boldsymbol{X}_j, Y_j)\}_{j=1}^N$, where $\boldsymbol{X}_j \in \mathbb{R}^p$ denotes covariates (a.k.a. features), $Y_j = \{1, -1\}$ represents class labels. The central task is to build a classifier that predicts the label based on the observed covariates. Numerous literature is available on binary classification procedures, including nearest neighbor classifiers, discriminant analysis, logistic regression, tree-based methods, support vector machine, and ensemble learning. See, for example, Hastie, Tibshirani and Friedman (2010) and Fan et al. (2020) for a comprehensive review.

Support vector machine (SVM) is a theoretically motivated classifier and has gained significant popularity in various applications (Boser, Guyon and Vapnik, 1992; Cortes and Vapnik, 1995; Vapnik, 2013). As a margin-based approach, SVM aims to find the maximum-margin hyperplane in either the original or

*Corresponding author. E-mail: pingma@uga.edu