Statistica Sinica **33** (2023), 1-22
doi:https://doi.org/10.5705/ss.202021.0028

AN EFFICIENT GREEDY SEARCH ALGORITHM FOR HIGH-DIMENSIONAL LINEAR DISCRIMINANT ANALYSIS

Hannan Yang, Danyu Lin and Quefeng Li

University of North Carolina at Chapel Hill

Abstract: High-dimensional classification is an important statistical problem with applications in many areas. One widely used classifier is the linear discriminant analysis (LDA). In recent years, many regularized LDA classifiers have been proposed to solve the problem of high-dimensional classification. However, these methods rely on inverting a large matrix or solving large-scale optimization problems in order to render classification rules, making them computationally prohibitive when the dimension is ultrahigh. With the emergence of big data, it has become increasingly important that we develop more efficient algorithms to solve high-dimensional LDA problems. In this paper, we propose an efficient greedy search algorithm that depends solely on closed-form formulae to learn a high-dimensional LDA rule. We establish a theoretical guarantee of its statistical properties in terms of variable selection and error rate consistency. In addition, we provide an explicit interpretation of the extra information brought by an additional feature in an LDA problem under some mild distributional assumptions. We demonstrate that the computational speed of the new algorithm is significantly better than that of other high-dimensional LDA methods, while maintaining comparable or even better classification performance.

Key words and phrases: Greedy search, high-dimensional classification, linear discriminant analysis, Mahalanobis distance, variable selection.

1. Introduction

Classification–assigning a subject to one of several classes based on certain features–is an important statistical problem. However, the recent emergence of big data poses great challenges, for it requires the efficient use of many features for classification. A simple classifier, namely, linear discriminant analysis (LDA) was widely used before the big data era (Anderson (1962)). However, as Bickel and Levina ((2004) have shown, when the number of features exceeds the sample size, a traditional LDA is no longer applicable, owing to the accumulation of errors when estimating the unknown parameters. To deal with the high-dimensional

Corresponding author: Quefeng Li, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. E-mail: quefeng@email.unc.edu.