

Abstract

Feature screening is crucial in analysis of ultrahigh dimensional data, where the number of variables (features) is exponentially larger than the number of observations. In various ultrahigh dimensional data, variables are naturally grouped and have correlation within the group. In this article, we propose a new feature screening method for data with grouped variables. This is in clear contrast with existing literature on feature screening, which screens the variables one-by-one. Under mild technical conditions, we prove that the group screening procedure possesses a sure screening property, which is defined by Fan and Lv (2008), while controlling the false positive rate. We conduct simulation studies to demonstrate the advantages of the proposed method and apply it in data analysis of genome-wide association studies (GWAS).