

From High-Dimensional Classification to Clustering: High-Impact Feature Selection via ggCGA and UL-MLR

Shih-Feng Huang

Graduate Institute of Statistics, National Central University

This study develops a unified framework for high-impact feature (HIF) selection in high-dimensional classification and clustering. We first propose ggCGA, a new feature selection procedure for high-dimensional multinomial logistic regression (MLR), and establish its selection consistency under mild regularity conditions. Building on ggCGA, we further develop UL-MLR, an unsupervised learning framework that extends HIF selection to high-dimensional clustering. UL-MLR iteratively alternates between an MLR step and an unsupervised learning (UL) step. In the MLR step, ggCGA identifies HIFs using pseudo-labels generated by the UL step, while in the UL step, the pseudo-labels are updated based on the selected features. For each fixed iteration, the selection consistency of UL-MLR is established when the pseudo-labels satisfy a proposed local imbalance condition. Through this iterative procedure, UL-MLR jointly refines feature selection and cluster assignment. Extensive simulation studies and real-data applications demonstrate that ggCGA effectively identifies HIFs in high-dimensional classification, while UL-MLR achieves accurate HIF selection and reliable clustering performance in high-dimensional settings.