Imbalanced data in classification

Abstract : Imbalanced data in classification tasks presents a significant challenge, as traditional machine learning algorithms often struggle to detect minority classes effectively. This work presents an asymptotic study of ensemble methods focusing on infinite random forests and bagged nearest neighbors trained on subsamples without replacement within binary classification. We establish a Central Limit Theorem (CLT) and propose a resampling strategy for imbalanced data. The resulting estimator requires debiasing via odds ratio to achieve consistency. Our analysis shows that debiased estimators achieve variance reduction compared to standard methods trained on original data. A numerical illustration of these results is carried out on fraud datasets.