

局所密度適応カーネルに基づく決定境界制御による極端クラス不均衡二値分類

氏 名 西村虎太郎ジェームス

研究室名 数理情報学研究室

主指導教員名 池田和司

内容梗概 (1 ページ目に収めること)

本論文は、クラス不均衡を伴う二値分類において、極端な不均衡比の下で少数クラス認識が著しく劣化するという実問題上の困難に対し、局所密度に応じて決定境界を自律的に調整する新規手法を提案し、その有効性と適用限界を体系的に検証するものである。従来の不均衡学習は、(i) SMOTE 等の過標本化による幾何構造の歪み、(ii) 近傍削減型の下標本化による情報損失、(iii) 重み付き損失・アンサンブル等における領域ごとの密度差への非適応性、といった問題を抱え、とりわけ「多数派が局所的に高密度に凝集する」型の極端不均衡では、境界が多数派側に過剰拡張して少数派領域が過小推定されやすい。

そこで本研究では、Kernel-density-Oriented Threshold Adjustment with Regional Optimization (KOTARO) を提案する。KOTARO は、局所スケール（例：各点の k 近傍半径）により点ごとに帯域幅を変えるガウス基底を割り当て、その線形結合として判別関数を構成することで、高密度な多数派領域での平滑化を抑えつつ、低密度な少数派領域では影響範囲を拡張し、固定帯域幅に起因する“にじみ出し”による境界の過剰拡張を抑制する。推定は密度適応類似度行列に対する正則化付き線形代数として実装できる一方、素朴実装では密行列演算を伴い計算量が課題となり得るため、低ランク近似等の拡張も展望として位置付ける。人工データ (Extreme Imbalance 型 / Divergent Imbalance 型) および UCI 由来の実データに対して、G-mean, F1, Precision により代表的手法 (SMOTE+SVM, RUS+SVM, 重み付き SVM, Random Forest, AdaBoost 等) と比較した結果、とくに Extreme Imbalance 型かつ高次元・強不均衡条件で少数派の見落としを抑え、従来法が境界を多数派側へ押し込みやすい状況で優位性を示した一方、特徴選択等によりノイズが低減される場合には差が縮小する傾向も観察され、提案法の適用条件と限界が明確化された。