［様式－学 5］

主　論　文　要　旨

論文題名

不均衡データ環境における分類のための学習アルゴリズム

氏名　　NGUYEN MANH HIEN（ぐえん　まいん　ひえん）

主論文要旨

従来の機械学習は対象となるデータが均衡したクラス分布であることを仮定しているため、不均衡データに対しては十分満足できる結果を出すことができない。本学位論文では、静的および動的の両環境におけるクラス不均衡問題を扱っている。静的環境では、あらかじめすべての学習データが既知であるため、分類モデルの構築は一度のみとなる。静的環境においてクラス不均衡問題を扱うために、本学位論文は SVM（support vector machines）の決定境界をよりよく絞り込むことのできる新しいオーバーサンプリング手法を提案している。その手法は補間と外挿の両手法により少数クラスデータをより適切に配置する新しいオーバーサンプリング手法の考えにもとづいている。ベンチマークデータ UCI を用いた実験結果から、本提案手法は他のオーバーサンプリング手法と同様、標準 SVM よりは高い成果を達成した。

一方、動的環境では、学習データが連続してデータストリームとして与えられる。そのため、新しい学習データを用いたモデルの繰り返し更新やコンセプトドリフト（時間とともに変化するデータ分布）の扱い方といった静的環境では存在しない課題に挑戦しなければならない。本学位論文では、以下を含む新しい手法を提案している。（1）不均衡データストリームから繰り返し学習する 2 つのオンラインサンプリング手法、(2) 過去の少数クラスデータの再利用にもとづいてコンセプトドリフトとクラス不均衡の両方を用いてデータストリームを処理する 2 つの手法。実世界のデータストリームと同様のシミュレーション結果から、本提案手法は従来のどの手法よりも優れていることを確認した。さらに、本学位論文では、不均衡データストリームに関するサンプリング手法の比較研究から、学習データサイズのサンプリング手法の相対的性能への影響といった多くの新しく有用な知見が導かれた。