スタッキングを利用した不均衡データに対する新たな分類手法の提案

曽我部　泰誠

不均衡問題は，分類問題において，あるクラスのサンプルサイズが他のクラスのサンプルサイズよりも多い場合に発生する。学習を行っても多数クラスと予測する結果を生じさせやすく，注目を置く少数クラスの分類が困難である問題をさす。このような不均衡データに対して本研究ではリサンプリングによるブートストラップサンプルに対して SMOTE またはアンダーサンプリングを使用し，スタッキングを行う SMOTE-stacking 及び Under-stacking を提案する。ブートストラップとは元のデータのサイズを維持したまま，ランダムに重複を許容して対象を抽出することでデータを複製する方法である。SMOTE 及びアンダーサンプリングはデータの不均衡さを修正するサンプリング手法であり，SMOTE はある少数クラスのサンプルに対して，似た性質の少数クラスのサンプルを合成することで新しく少数クラスのサンプルを作成する。少数クラスのサンプルの数を増やすことで，データの不均衡さを修正する方法である。一方アンダーサンプリングは多数クラスのサンプルをランダムにデータから削除する。多数クラスのサンプルの数を減少させることでデータの不均衡さを修正する方法である。スタッキングとは複数の弱学習器で分類を行い，その結果を特徴量とした新たなデータを作成する。そのデータをもとに分類することで最終的な分類を行う手法である。これらの手法を組み合わせたものが提案手法である。

本研究の目的はスタッキングを利用した不均衡データに対する新たな分析手法を提案し，その有用性を検証することである。機械学習の性能を評価する概念としてバイアスと分散が存在している。バイアスとはモデルが真の値からずれる程度を表す指標で，分散とはモデルの予測値がばらつく程度を表す指標である。高いバイアスは、モデルが単純すぎてデータの本質的なパターンを捉えられていないことを示し，高い分散は、モデルが訓練データに過度に適応し、新しいデータに対して一般化ができない可能性があることを示す。このバイアスと分散はトレードオフの関係にあり、どちらかが高いと適切なモデルとは言えない。モデルが複雑になるほど訓練データに適応しやすくなり、分散が増加する。一方で、モデルが単純になると訓練データに適応しにくくなり、バイアスが増加する。提案手法の利点は，使用する弱学習機に使用するデータを多様にすることで、分散を低減しつつ、弱学習器の結果をさらにメタ学習器を使用し、分類することで、バイアスを補正できることである。

シミュレーションと実データ分析の結果，①少数クラスの絶対数が多い，②データの特性上分類が困難という2つの場合において特に既存手法より提案手法が優れた分類性能を示すという結論に至り，提案手法は分類性能，汎用性に優れた手法といえる。また提案手法の SMOTE-stacking と Under-stacking の比較では，わずかに SMOTE-stacking が良い分類性をを示したが，使用可能な場面と計算時間の観点から Under-stacking の使用が推奨されるという結論に至った。(行動統計科学)