＜ 修 士 論 文 ＞

不均衡データに対する機械学習手法と税関不正検知への応用

（要　旨）

滋 賀 大 学 大 学 院
デ ー タ サ イ エ ン ス 研 究 科
デ ー タ サ イ エ ン ス 専 攻

修了年度 ： ２０２０年度

学籍番号 ： 6019114
氏　　名 ： 辻　　孝 辰
指導教員 ： 松井　秀俊

提出年月日 ： 2021年1月20日

近年，増加の一途を辿る貿易額や訪日外国人旅客数に対し，税関では業務のより一層の高度化・効率化が急務となっている．その取り組みの一つとして，輸出入申告等の膨大なデータを解析し，検査対象等の選定を支援することが検討されている．解析によって，不正な申告等を自動的に検知することを税関不正検知と呼ぶ．従来の税関不正検知はルールベースのものが主流であったが，機械学習の活用が世界各国の税関当局で検討されている．そして，税関不正検知に機械学習を活用する上で，取り扱うデータが不均衡であることが，解決すべき課題の一つとされている．

不均衡データとは，分類問題において，各クラスのサンプルサイズに偏りがあるデータのことをいう．不均衡データによって一般的な機械学習モデルを訓練すると，サンプルサイズの大きいクラス（多数クラス）に予測しがちなモデルが構築されてしまう．この問題に対処するため，これまで様々な手法が提案されている．既存手法においては，識別器を訓練する前に，訓練データの各クラスのサンプルを新たに生成（オーバーサンプリング）または取り除く（アンダーサンプリング）ことによって不均衡を解消するアプローチ（データに基づくアプローチ）が広くとられている．また，アンサンブル学習における弱識別器の訓練データを均衡にするために，オーバーサンプリングまたはアンダーサンプリングを利用する手法も数多く提案されている．

本研究では，まず，データに基づくアプローチ及びそのアンサンブル学習との組み合わせについて，評価を行った．評価には，訓練に要する時間，AUC-ROC，AUC-PR，正解率，再現率，適合率，F1 値，マシューズ相関係数，G-mean の 9 つの評価指標を用いて，多面的かつ公平に比較した．その結果，特定のアンサンブル学習手法では，単体の決定木を訓練するよりも良いモデルが構築されることが確認できた．さらに，この評価結果から，先行研究によく見られる，単一の指標による一面的な評価で優劣を判断することには問題があり，様々な指標による評価が重要であることも確認できた．

次に，これらの手法を税関への輸入申告の疑似データに対して適用し，その効果を検証した．この検証により，アンサンブル学習が単一の決定木より良いという傾向が，税関不正検知において特に顕著に現れることが確認できた．これは，税関不正検知で扱うデータが非常にスパースであることに起因すると考えられる．なお，オーバーサンプリングとアンサンブル学習を組み合わせた手法については，非常に訓練時間を要するため，実用的でないことも確認できた．以上の結果を踏まえ，アンダーサンプリングとアンサンブル学習を組み合わせた手法を，税関の保有する輸入申告データに適用し，その効果を検証した．

この検証では，オーバーサンプリング後の訓練データで訓練した深層学習モデルとの比較も行い，それに対して同等か，場合によっては優れた性能であることが確認できた．深層学習のモデル構築に要するコストを考慮すると，この検証結果は本研究の貢献の一つといえる．

最後に，税関不正検知において有効であることが確認できた，アンダーサンプリングとアンサンブル学習の組み合わせについて，改善策を検討した．検討にあたって，まずはより有効なアンダーサンプリング手法として，クラスタリングに基づくアンダーサンプリング（CUS）の先行研究を調査した．その中から，ClusterBal と CUSBoost について，実装し評価を行った．ClusterBal は，負例の各クラスタそのものを全ての正例とそれぞれ結合することで，弱識別器の訓練データを構築する手法である．一方，CUSBoost では，AdaBoost の各反復において，CUS を利用して訓練データを構築する手法である．この手法における CUS では，負例の各クラスタから 50%ずつ負例を抽出する．これらの手法を評価した結果，ClusterBal では負例の各クラスタを抽出し，それぞれを各弱識別器の訓練データとすることに問題があり，CUSBoost では負例の各クラスタから 50%抽出することで，依然として不均衡が解消されていないことに問題があることが判明した．

これらの問題を解決するため，負例の各クラスタから抽出する負例数を変更し，各弱識別器の訓練データを均衡とする方法と，クラスタリング前に UMAP によって次元圧縮することを提案した．提案手法を，UnderBagging，RUSBoost，EasyEnsemble のランダムアンサーサンプリングと置き換えることにより実装し，それらと性能を比較したところ，提案手法により僅かに AUC-PR，再現率が向上した．