Effectiveness of Semi-Supervised Learning and Multi-Source Data in Detailed Urban Landuse Mapping with a Few Labeled Samples

少数のラベル付きサンプルによる詳細な都市土地利用マッピングにおける半教師付き学習とマルチソースデータの有効性

概要

詳細な都市土地利用情報は、スマートシティ管理において基本的な役割を果たす。十分なサンプルサイズは、都市の土地利用分類のための機械学習アルゴリズムにおいて、非常に重要な事前要求として認識されている。しかし、リモートセンシング画像だけから土地利用のカテゴリーを認識し、ラベル付けすることはしばしば困難である。また、現地調査には多くの時間と人的資源、金銭的コストがかかる。そのため、都市の土地利用分類に関する先行研究では、空間分布が非常に不均一で、ラベル付けされたサンプルのサイズが小さいことに依存していることが多い。本研究は、少数のラベル付きサンプルで詳細な都市土地利用分類を行うために、マルチソースデータを用いた半教師付き分類フレームワークの有効性を探ることを目的とする。不一致ベースの半教師あり学習アプローチであるCo-Forestを採用し、従来の教師あり手法（ランダムフォレストやXGBoostなど）と比較した。土地利用区分の物理的・社会経済的特徴を示す光学・夜間光リモートセンシングや地理空間ビッグデータなど、複数の地理空間データを利用した。深圳市の都市土地利用分類をケースとした結果、半教師付き手法の分類精度は概ね従来の教師付き手法と同等であり、異なる訓練セットの比率の下で同等の結果を達成するためには、より少ないラベル付きサンプルが必要であることが示された。サンプルサイズが小さい場合、精度は合計5%以下の訓練サンプルで安定する傾向がある。また、マルチソースデータを使用した場合の分類精度は、単一のデータソースを適用した場合よりも有意に高いことが示された。これらのデータの中では、地図POIと高解像度光学リモートセンシングデータが分類に大きく寄与し、モバイルデータと夜間光リモートセンシングデータがそれに続く。

キーワード：都市土地利用、小サンプル学習、半教師付き分類、サンプリング戦略、マルチソース地理空間データ

1. はじめに

スマートシティ構築の発展に伴い、正確で詳細な都市土地利用情報は、都市計画、資源配分、行政にとって基本的な役割を果たしている。最新の都市土地利用地図は、スマート社会の管理において高い需要がある。リモートセンシング技術は、広範囲の観測と変化への迅速な対応が可能であるため、都市の土地利用と土地被覆の分類に関する多くの研究に広く利用されている[1,2,3,4]。従来の都市土地利用分類技術は、マルチスペクトル・リモートセンシング画像に基づいている。リモートセンシング画像の空間解像度が向上するにつれて、より正確な分類を得るために、スペクトル特徴に加えて、幾何学的特徴やテクスチャ特徴が採用されている[5,6,7]。しかし、都市の土地における社会的属性は複雑であるため、画像データから詳細な都市土地利用情報を導き出すことはしばしば困難である。ビッグデータ時代の到来に伴い、タクシー軌跡[8]、地図上のPOI（Point of Interest）データ[9,10]、ジオタグ付き写真[11]、ソーシャルメディアデータ[12,13]、携帯電話データ[14,15]などのマルチソースの地理空間データから得られる人間行動や社会経済的特徴が、都市の機能ゾーニングや土地利用分類に活用されている。

数多くの機械学習アルゴリズムが開発され、都市の土地利用分類に応用されている。機械学習アルゴリズムにおいて、学習サンプルの数は非常に重要であり、モデルの学習性能に直接関係する。例えば、WielandとPittore [16]は、マルチスケールリモートセンシングデータに基づく都市土地利用パターンの認識に、ナイーブベイズ（NB）、k-最近傍（kNN）、ランダムツリー（RT）、サポートベクターマシン（SVM）など、いくつかの機械学習アルゴリズムを採用した。これらのアルゴリズムは、学習サンプルサイズが十分大きければ、より良いパフォーマンスを示すことが示された。Sunら[17]は、マルチソースデータの統合に基づく都市既成市街地の抽出において、ロジスティック回帰（LR）、決定木（DT）、ランダムフォレスト（RF）、勾配ブースティング決定木（GBDT）、AdaBoostを含む別の5つの分類器を比較した。彼らは、都市既成市街地の公式マップを使用してラベル付きサンプルを取得し、モデル学習に大量のサンプルを採用した。Caoら[18]は、マルチソースの地理空間データに基づく都市の土地利用分類において、eXtreme gradient boosting（XGBoost）アルゴリズムと古典的な機械学習アルゴリズムを比較した。彼らはモデル学習に80％のラベル付きサンプルを選択した。Zhangら[19]は、RFアルゴリズムをリモートセンシングとソーシャルセンシングデータの組み合わせに基づく都市土地利用分類に適用した。彼らは、ラベル付きサンプルの50％をトレーニングサンプルとして使用した。これまでの研究から、モデルの訓練と学習には十分なサンプルサイズが必要である。しかし、リモートセンシング画像のみから土地利用区分を認識し、ラベル付けすることは、手作業による解釈の補助があったとしても、困難な場合が多い。あるいは、現地調査は労力と時間がかかり、人的資源と金銭的コストが高くつく[20]。特に詳細な都市土地利用分類の場合、「既成市街地」と大雑把にラベリングするのではなく、住宅地と商業地の区別など、より明確で詳細な土地利用ラベルを取得することが課題となる。限られたサンプルでは、都市土地利用の信頼できる分類結果を得ることはほとんど不可能である。その上、サンプリングコストと分類精度のバランスが取れる理想的な学習サイズを示した研究はほとんどない。モデル学習のためのサンプルの選択は比較的恣意的であり、その結果、汎化性能が低下する可能性がある。サンプルのサイズが小さいことを前提とした都市土地利用分類におけるサンプリング戦略の理論的な指針はまだ得られていない。以上の理由により、ほとんどの詳細な都市土地利用分類は、小さなサンプルの学習という問題に直面している。

近年、小サンプル学習の問題を解決するために、半教師付き学習法がますます開発されている[21,22]。しかし、詳細な都市土地利用分類に半教師付き分類を適用した場合の有効性は報告されていない。さらに、多ソースデータは、より良い分類精度のために、より多くの次元の特徴を提供することができると広く信じられているが、冗長性や矛盾した情報さえも導入する可能性がある[18,23]。したがって、本研究では2つの研究課題に取り組む。(1)半教師付き分類フレームワークは、サンプル数が少ない詳細な都市土地利用分類に有効か？ (2)マルチソース地理空間ビッグデータの利用は、詳細な都市土地利用分類の精度を効果的に向上させるか？

多数のラベル付きサンプルを得ることが困難であることを考慮し、我々は、マルチソースの地理空間データから物理的・社会経済的特徴を取り入れることで、詳細な都市土地利用分類に半教師付き分類法を採用することを目指した。また、少ないサンプル数のもとで、信頼性の高い分類結果を得るための最適な訓練セットの比率を見つけるために、訓練サンプルの比率を変えて分類の安定性をテストすることを試みた。

2. 関連作業

リモートセンシングや社会属性データを含むマルチソースの地理空間ビッグデータに基づいて都市の土地利用分類を実施するためには、いくつかの重要な課題がある。(1)ディープラーニングを含む従来の機械学習アルゴリズムでは、信頼性の高い分類器を構築するために多数の学習サンプルが必要である。しかし、現地調査による都市土地利用タイプの信頼性の高いラベリング作業は時間がかかり、人的・金銭的リソースも必要となるため、多くのラベル付きサンプルを得ることは通常困難である。(2)サンプリングのコストとモデル学習効果のバランスを考慮すると、都市土地利用分類における様々なアルゴリズムに最適な学習サンプルのサイズは、まだ十分に検討されていない。本節では、既存の都市土地利用分類手法と、限られた訓練サンプルにおける分類の安定性の問題に関する関連する議論をレビューする。

2.1.

都市の土地利用分類手法過去数十年間、都市の土地利用や土地被覆の分類、都市機能ゾーニングのために、さまざまな機械学習アルゴリズムが開発されてきた。リモートセンシング画像とは別に、マルチソースのソーシャルセンシングデータを採用するケースも増えている。マルチソースデータの活用は、都市土地利用分類分野の重要な方向性となっている[24]。最も一般的に使用されているアルゴリズムには、SVM、DT、RFがあり、これらは比較分析におけるベンチマークともみなされている。Mountrakisら[25]は、リモートセンシングアプリケーションにおけるSVMアルゴリズムをレビューし、SVMアルゴリズムがマルチクラス画像分類タスクに適していることを指摘した。アルゴリズムが単純であることを考慮すると、DTアルゴリズムは、異なるスケールで複雑な特徴を融合する利点があり、リモートセンシングの都市土地利用分類に広く使用されている[26]。ツリーベースのDTアルゴリズムをさらに進化させたものとして、RFアルゴリズ ムもよく使われる手法である。Talukdarら[27]は、マルチスペクトル・リモートセンシング画像に基づく土地利用と土地被覆の分類において、RF、SVM、人工ニューラルネットワーク（ANN）を含むいくつかの機械学習アルゴリズムを要約した。その結果、RFアルゴリズムが最も優れた性能を示した。Zhangら[28]は、地図POIデータに基づく土地利用分類において、機械学習アルゴリズムを比較した。彼らは、RFアルゴリズムのようなツリーベースの手法が、ベイジアンネットワーク、ルールベース学習、SVMよりも良いパフォーマンスを示したという結論を導き出した。ディープラーニングの発展に伴い、高レベルの特徴を持つディープニューラルネットワークアルゴリズムが都市機能ゾーンの認識に使用されている[29]。Jozdaniら[30]は、オブジェクトベースの土地利用分類において、深層学習アルゴリズムと従来の機械学習アルゴリズムを比較した。彼らは、XGBoostアルゴリズムのような伝統的な機械学習アルゴリズムは、深層学習アルゴリズムと同等の結果を示した。そのほか、ガウス混合モデル、k-meansアルゴリズム、カーネル密度分類アルゴリズム、階層クラスタリングアルゴリズムなどの教師なしアルゴリズムが、都市の土地利用分類に適用されている[31,32]。分類結果は一般に、教師あり手法ほど良くない。

2.2. サンプル・サイズが小さい場合の分類の安定性

実用的なアプリケーションでは、都市土地利用分類において限られたサンプルを使用することは非常に一般的である。なぜなら、ラベル付きサンプルは常に稀であり、多数のラベルなしサンプルにラベルを付けるにはコストがかかりすぎるからである。サンプルサイズが小さいとは、ラベル付けされたサンプルまたは学習サンプルの割合が、分類されるサンプルに対して小さい状況を指す。Zhaoら[33]は、ラベル付きサンプルのサイズが小さい機械学習の現在の研究をレビューした。彼らは、この分野では進歩が見られるものの、まだ多くの課題があると結論付けている。Liら[34]は、マルチスペクトル・リモートセンシング・データに基づき、教師あり分類器と教師なし分類器を用いて、都市土地利用分類における学習サンプル・サイズの影響をテストした。その結果、樹木ベースの分類器は訓練サンプルサイズに対してより敏感であることが示された。Suら[35]は、マルチソースの地理空間データに基づき、RFアルゴリズムを用いて、詳細な都市土地利用分類における訓練サンプルサイズの影響を分析した。彼らは、訓練サンプルの割合が7％以下であれば安定した結果が得られると指摘している。ラベル付きサンプルの増加という観点から、Gongら[36]は、分類モデルの汎化性能を向上させるために、中国の30以上の都市でラベル付き都市土地利用サンプルを大量に入手するクラウドソーシング手法を採用した。しかし、30都市の分類対象となるサンプルの総数を考慮すると、ラベル付けされたサンプルの数は、特定の都市土地利用カテゴリーにとってはまだ少ないと考えられる。さらに、都市の土地区画に関する土地利用情報のラベリングは、専門家（都市計画家など）の経験に大きく依存しており、通常、クラウドソーシング・アプローチで入手することは困難である。

3. 材料と方法

本研究では、道路セグメンテーションに基づく区画ベースの土地利用分類を実施した。モデル学習と分類のために、高解像度の光学リモートセンシング画像からのマルチスペクトル特徴とテクスチャー特徴、夜間照明（NTL）リモートセンシング画像からの明るさ特徴、地図POIと携帯電話データからの人間の活動と行動特徴を含む、多ソース地理空間ビッグデータからの都市土地利用特徴セットを選択した。多次元特徴に基づき、詳細な都市土地利用分類のために半教師付きCo-Forest分類フレームワークを採用し、RFやXGBoostなどの最も一般的な教師付きツリーベース分類器と比較した。サンプルの少なさがモデルの性能に与える影響を分析するため、学習サンプルの割合を変えた場合の分類モデルの安定性もテストした。

3.1. 調査地域

我々は深圳市を調査地域として選んだ（図1a）。深センは中国南部の沿岸都市で、香港との国境にある。面積は1997km2で、2020年末には人口が1,300万人を超えると予想されている。中国の改革開放政策の先駆的な都市のひとつである。過去数十年の急速な都市開発により、都市景観の変化など劇的な変化を経験している。深センは、中国の包括的改革の国家パイロット都市に指定され、広東-香港-マカオ大湾区の建設をリードしている。新しい住宅地、工業団地、交通網、観光インフラが計画されている。今後、都市景観のさらなる変化が予想される。中国で最も急速に成長している都市の1つである深センは、よりオープンで包括的な地理空間ビッグデータを提供することができ、都市研究の自然なテンプレートとみなされている。

図1. 図1 調査地域の位置（a）と分類する土地区画の道路区分（b）。

3.2. 土地区画の道路セグメンテーション

オープンソースのデータセットであるOpenStreetMap（OSM）（https://www.openstreetmap.org、2021年12月6日アクセス）の道路ネットワーク情報は、調査地域全体を土地区画に分割するために利用された。OSMデータから、幹線道路と二次道路という2つのレベルの重要な道路を利用した。道路網のジオメトリは、OSM データに一本の線（すなわち、道路中心線）で示されている。バッファゾーンは、道路レベルによって決定された道路の幅に基づいて、これらの線に適用された。土地被覆ではなく、詳細な都市土地利用を分類することを目的としたため、森林、草地、湿地などの非建築地域をマスクするために、GAIA\_2018（http://data.ess.tsinghua.edu.cn/gaia.html、2021年12月6日アクセス）の深圳の不浸透面データを利用した。土地区画のサイズは不均一であった。セグメンテーションの後、面積が1000平方メートル未満の非常に小さな土地区画が除去された。最終的に、都市部の土地区画の数は6800以上となった。図1bは、深圳におけるこれらの土地区画の分布を示している。

3.3. データとデータの前処理

都市の土地利用の特徴を多次元的に示すために、複数のソースの地理空間データを利用した（図2）。これらは、Sentinel-2高解像度リモートセンシングデータ（出典：https://earthengine.google.com/、2021年12月6日アクセス）、Luojia-1 NTLリモートセンシングデータ（出典：http://59.175.109.173:8888/app/login.html、2021年12月6日アクセス）、Gaode Map POIデータ（出典：https://lbs.amap.com、2021年12月6日アクセス）、および中国の大手サードパーティービッグデータ企業が提供するGPS位置情報ベースのモバイルビッグデータである。モバイルデータは、グリッド（解像度約140m）内のアクティブなモバイルデバイスの累積数を月別に記録した。日中（午前9時～午後5時）と夜間（午後9時～午前5時）の統計が採用された。データの時間的整合性を保つため、過去のPOIデータを公的な方法で入手することが困難であったため、地図POIデータを除き、全てのデータセットは2019年の同時期から収集された。したがって，時間的整合性をできるだけ保つために，2020 年の POI データを収集した．表 1 にデータの基本的な特徴をまとめる．

図2. (a)高解像度光学リモートセンシング画像、(b)NTLリモートセンシング画像、(c)携帯電話データ、(d)地図POIデータ。

表1. 詳細な都市土地利用分類のためのデータの特徴。

リモートセンシングデータ（Sentinel-2とLuojia-1）に対する雲や降水の影響を軽減するために、秋と冬に雲のない合成画像を利用した。深センが移民都市であることを考慮すると、人間活動は休日経済や人口移動の影響を受けやすい。モバイルデータは、2018年10月（国慶節祝日）、2018年11月（非祝日）、2019年2月（旧正月）に収集された。データの前処理として、データクレンジングと座標変換を行った。地図POIデータに関しては、前処理には都市土地利用分類システムに合わせるためのカテゴリー再分類も含まれた。

3.4. 特徴選択と次元削減

マルチスケールデータは、データ特徴（土地区画に対する属性）の形で土地区画に基づいて空間的に統一された。これらの属性（例えば、土地区画の平均NTL輝度値）を生成するために統計的特性が利用された。事前知識に基づいて、物理的特徴および社会経済的特徴を含む初期特徴を手作業で選択した。前者は主に、帯域スペクトル特性、植生指数、夜間の明るさなどのリモートセンシング画像データからである。後者は主に地図POIとモバイルビッグデータからで、昼間と夜間の人口密度、異なる月の人口、POI密度、タイプなどである。

ほとんどの機械学習アルゴリズムと同様に、分類モデルの複雑さを単純化するために特徴空間の次元を小さくする必要がある。都市の土地利用タイプを最もよく記述し、提示できるいくつかの特徴が最終的に決定された。一般に、次元削減には、特徴抽出と特徴選択という2つのアプローチがある[37]。特徴抽出は、特徴の変換または特徴の組み合わせによって新しい特徴空間を構築することを目的とする。これは、元の特徴情報を失う可能性のある新しい特徴の生成を伴う。新しく生成された属性は、通常、物理的に説明することが困難である。一方、特徴選択は、特徴の物理的解釈可能性を保持できる、元の属性特徴のサブセットを得ることを目的とする。特徴量の物理的な解釈可能性を維持するために、特徴量の次元削減には特徴選択が採用された。相関係数(r)は、特徴次元を最小化するために0.95の閾値で冗長性をフィルタリングするために利用された。rが0.95より大きい場合、対になった特徴のうち1つだけが残る。特徴量の削減後、特徴量のサブセットをモデル学習に利用した。

3.5. 半教師付き複数特徴分類フレームワーク

半教師付き学習の主な目的は、ラベル付けされたサンプルとラベル付けされていないサンプルの両方を用いて分類器を訓練することである。分類モデルはまずラベル付きサンプルから学習され、次にラベルなしサンプルによって改良される[38]。我々の研究では、半教師付きCo-Forestアルゴリズムを詳細な都市土地利用分類に適用した。Co-Forestアルゴリズムは、不一致ベースの半教師付き分類手法であり、RFアルゴリズムに基づく共学習（共学習スタイルのRFアルゴリズム）の拡張とみなされる[39]。Co-Forestでは、ランダム決定木のようなN個（N > 3）の分類器が、最初にラベル付きサンプル集合に基づいて個別に学習される。もし分類器がラベル付けされていないサンプルをある信頼度でラベル付けすることに合意すれば、元のサンプルと新しくラベル付けされたサンプルに基づいて新しい学習セットが生成され、分類器が再学習される。ラベル付けされていないサンプルに対する自動ラベリング戦略は、より高いコストパフォーマンスを特徴とする[40]。図3は、マルチソースの地理空間データとCo-Forestアルゴリズムに基づく、詳細な都市土地利用分類の研究フレームワークを示している。

3.6. モデルの調整と改良

都市の土地利用分類は典型的な多項式分類タスクである。訓練セットには不均衡の問題があるという共通の問題がある。例えば、様々な都市土地利用タイプは都市に均等に分布しておらず、サンプリングバイアスを引き起こしやすい。その上、サンプルが明確な定義や属性を持つ従来のターゲット認識や抽出タスクとは異なり、都市の土地区画は、特に多くの発達したメガシティでは、土地利用タイプが混在している可能性がある（例えば、商業・住宅混在）。これは、半教師付きモデルにおける学習結果の偏差の原因となる。上記の問題を最小化するために、Co-Forestアルゴリズムの3つの改善スキームを以下のように適用した。

改善スキーム1：サンプルの不均衡問題に対処するため、初期化と分類器構築の過程でサンプルの重みを追加する；一定の信頼レベルに達しない限り、ラベル付けされていないサンプルをラベル付けされたサンプル集合に追加すべきではないという判断をモデルに設定する（混合土地利用サンプルを除外する）。

改善スキーム2：改善スキーム1に基づいて、反復プロセスにおける誤差率の制限を追加する（すなわち、オーバーフィッティングを避けるために、誤差を0.2未満に制限して反復を終了する）。

改善スキーム3：改善スキーム2に基づいて、ラベルなしサンプルをラベル付きサンプルセットに追加した後、ノイズカットステップを実行する。

3.7. モデルの評価と精度評価

本研究では、モデル評価の信頼性を確保するため、k-fold (k = 5) 交差検証をモデル性能評価に採用した。分類精度を定量化するために、混同行列を使用し、総合精度（OA）とカッパ係数[41]を評価指標とした。

3.8. サンプルサイズが小さいことの影響分析

小さなサンプルサイズが分類結果に与える影響を調査するため、訓練サンプルの割合を変えて分類の安定性をテストしました。すべての訓練サンプルに基づいて、訓練サンプル数が毎回1%ずつ減少するにつれて、分類モデルがテストされました。サンプルの分布の割合を一定に保つため、各サンプリングプロセスには層化ランダムサンプリング法を採用した。学習サンプル数の最適なコストパフォーマンスは、精度の変化率によって決定された。変化率(m)は以下の式で計算される： m = (A - a\_k)/A, ここで、Aは全ての学習サンプルを使用した場合の最良の精度を表し、a\_kはk%の学習サンプルを使用した場合の精度を表す。

モデル学習の信頼性を高めるため、訓練サンプルは訓練セットからランダムに選択され、訓練セットの比率ごとに5回モデル学習が行われた。5回の精度の平均が、その訓練セットの比率での分類精度として採用された。

4. 実験

4.1. 特徴のサブセット

研究者の予備知識に基づき、まず都市の土地利用タイプの主な特徴として、マルチソースの地理空間データから59の特徴が選択された。冗長な特徴を排除するために、これらの特徴のうちの任意の2つの特徴間の相関係数が計算された。最終的に、34の特徴量がモデル学習に採用された。表2に、データの種類別に選択された特徴を示す。

表2. 次元削減後の特徴量

4.2. 都市土地利用分類システム

先行研究[34,35]と土地利用分類の国家標準（GB/T 21010-2017）を参考に、住宅（R）、商業（C）、工業（I）、交通（T）、公共管理・サービス（P）の5つのレベル1の都市土地利用区分からなる都市土地利用分類システムが採用された。表3は、詳細な土地利用分類を説明や例とともに示したものである。

表3. 深セン市の都市土地利用分類システム

4.3. ラベリングとTrain/Test分割

図4は、現地調査と超高解像度（VHR）画像の手動判読による2つのソースを含む、ラベル付きサンプルの分布を示している。現地調査データには、深圳市計画自然資源局都市計画土地資源研究センターから入手した162区画のラベル付き土地区画が含まれている。その他のラベル付きサンプルは、人間の視覚によるVHR画像の解釈から得られたものである。視覚的解釈プロセスは、Google EarthのVHR画像に基づいており、Gaode mapsやMicrosoft Bing mapsなどの地図アプリで支援されている。そのほか、現地調査データも手作業による画像解釈を補助するために使用される。サンプルの品質管理を通じて、全土地区画の約15％を占める合計1021のラベル付きサンプルが、モデルの訓練とテストのために最終的に収集された。

すべてのラベル付けされたサンプルは、トレーニングセットとテストセットの2つのグループに分けられた。層別サンプリング法は、異なる土地利用区分に対して同じサンプル分布を確保するために採用された。モデルの学習結果を比較できるようにするため、全土地区画の3％に相当する一定数のラベル付きサンプル（すなわち204）がテストセットとして選択された。残りのラベル付きサンプル（つまり817）は、全土地区画の12％を占め、トレーニングセットとして採用された。最適なトレーニングセットの比率を見つけるために、モデルのトレーニングに使用するラベル付きサンプルの比率を毎回1%ずつ減少させた。各比率について、訓練サンプルはランダムに5回選択された。モデルの学習結果は、固定テストセットによって評価された。各トレーニングセット比率における5回の計算結果の平均を、その比率におけるモデルの精度とした。

4.4. 実験環境とパラメータ

実験はMac OSプラットフォーム（4コアCPU、16Gメモリ）で実施した。Co-Forestアルゴリズムと修正版の実装は、Java言語、JDKバージョン8.1、Waikato Environment for Knowledge Analysis (Weka)フレームワーク、バージョン3.8.4に基づいている。Co-Forestアルゴリズムの共学習分類子の数は3から20に設定された。

5. 結果と分析

5.1. Co-Forestのアルゴリズム改良のパフォーマンス

図5は、オリジナルのCo-Forestアルゴリズムと改良版の分類精度を、それぞれ総合精度とカッパ係数で示す。結果は、3つの改善スキームすべてが、Co-Forestアルゴリズムのオリジナル・バージョンよりも優れていることを示した。その中でも、改良スキーム2は、他の改良スキームと比較して、より多くの回数No.1のランキングを獲得しており、より優れた性能を有していた。したがって、さらなる比較と分析では、改善スキーム2が好ましい方法として採用された。

図5. アルゴリズム改良の性能評価、(a)総合精度、(b)カッパ係数。

5.2. 従来の教師ありアルゴリズムとの比較

モデルの改良に基づき、半教師付きCo-Forestアルゴリズムの分類精度を、RFやXGBoostを含む伝統的な教師ありアルゴリズムと、訓練サンプルサイズの異なるレベルで比較した。図6から、半教師ありアルゴリズムの性能は、7%以上の訓練セット比率を使用する場合、他の2つの教師ありアルゴリズムよりも優れていた。言い換えると、提案するCo-Forest分類法は、より少ないラベル付き訓練サンプルを使用することで、同等の、さらにはより良い分類結果を達成することができる。表4は、異なる精度レベルにおける3つのアルゴリズムの最小サンプルサイズの必要性を示している。類似のツリーベース分類器と比較すると、半教師付き学習フレームワーク（Co-Forest）は、同じ精度レベルを達成するために、ラベル付きサンプルを17~20%削減できる。

図6. 従来の教師ありアルゴリズムとの精度の比較。

表4. 異なる精度レベルを達成するために必要な最小サンプルサイズ。

5.3. 訓練サンプルサイズの影響

訓練サンプル・サイズが分類性能に与える影響を分析するために、図7はサンプル・サイズが小さい場合の分類精度の変化率を示しています。訓練セットの比率が12%のときの最高精度を基準とすると、訓練サンプルサイズが小さくなるにつれてモデル性能は低下している。訓練サンプルサイズが5%より小さくなると、分類精度は急速に低下した。従って、学習サンプルサイズが5％以下であれば、高いコストパフォーマンス（サンプリングにかかる人件費と分類精度の比）を得ることができる。

図7. 全トレーニングサンプル（合計12%）の結果を基準として、トレーニングサイズの比率を変えた場合の分類精度の変化率。

5.4. マルチソース地理空間データの重要性

図8は、マルチソースデータのさまざまな組み合わせが分類精度に及ぼす影響を示している。すべてのマルチフィーチャーを使用することで、分類の精度が向上する。一般に、より多くの特徴を追加した方が、より高い精度を得ることができる。ソースを別々に考えた場合、地図POIと高解像度光学リモートセンシング（Sentinel-2）データは、他のデータセットよりも良い結果を示した。

図8. マルチソースデータの様々な組み合わせによる詳細な都市土地利用分類の精度。

5.5. 少ないサンプルによる詳細な都市土地利用マッピング

修正Co-Forestアルゴリズムとマルチソースデータに基づき、図9は深圳における5%の学習サンプルでの詳細な都市土地利用分類結果を示している。詳細な都市土地利用分類の空間分布は、公式の都市計画スキームとある程度一致していた。住宅地と商業地は主に繁華街に分布し、工業地は主に郊外に分布していた。

図9. 5%の学習サンプルを用いた半教師付きCo-Forestアルゴリズムに基づく深セン市の詳細な都市土地利用マッピング（2019年）。

分類結果を定量化するために、図10に混同行列と精度評価を示す。分類の全体的な精度は0.79であった。特定の都市土地利用カテゴリの精度については、生産者精度と利用者精度が、それぞれ省略エラーと委託エラーを測定する。その結果、「住宅」、「工業」、「公共管理・サービス」タイプが高い精度を達成した。分類の誤りは、主に「商業」と「住宅」の誤分類、「公共管理・サービス」と「住宅」の誤分類に集中した。

図10. 混同行列と精度評価、土地利用分類のコード1から5は、それぞれ住宅（1）、商業（2）、工業（3）、交通（4）、公共管理・サービス（5）を表す。

6. ディスカッション

6.1. 都市土地利用分類における小サンプル学習

学習サンプルサイズの選択は常に経験的プロセスである。通常のガイダンスは「可能な限り使用する」であるが、モデルの訓練とテストのためにラベル付きサンプルを収集するコストと時間を考慮すると、それはトレードオフである[42]。限られたラベル付きサンプルの都市土地利用分類の場合、ラベル付きサンプルの利用を改善するアイデアは、主にデータレベルまたはモデルレベルにある。前者には高次元データの利用が含まれ、後者には能動学習や半教師付き学習によるラベルなしサンプルの自動学習が含まれる。本研究では、この2つの考え方を併用する。

建物」や「既成市街地」といった土地利用の物理的特性は、リモートセンシング画像データから直接得ることができる。しかし、リモートセンシングデータだけでは、詳細な都市土地利用の高度な意味情報を得るには不十分である。高次元の土地利用特徴を提供することで、社会経済ビッグデータは詳細な都市土地利用の分類へのより良いアプローチを提供する。例えば、地図上のPOIデータは通常、都市の土地利用分類の特定と非常に密接な関係があると考えられている[43]。単一のデータソースではまだ不十分であることを指摘しておく。我々の実験によると、地図POIデータだけ、あるいはPOIデータと光学リモートセンシングデータの組み合わせだけでは、満足のいく分類結果を得ることは難しい。これはマルチフィーチャの重要性を裏付けるものであり、マルチソース、マルチモーダルデータを利用することで、より高い分類精度を達成することができる。

学習サンプルよりも多くのパラメータが存在する可能性があるため、学習サンプルサイズが小さい場合、ディープニューラルネットワークが失敗する原因と一般的に考えられている。RFやXGBoostのような木ベースの分類器は、より優れた分類効果を示すことが証明されており、都市の土地利用分類アプリケーションで広く利用されている[44]。そこで本研究では、比較のために半教師付き木ベースの分類フレームワークを採用した。その結果、半教師付き分類法は、教師付き分類法よりも優れた性能を示し、分類精度を低下させることなく、モデル学習のためのラベル付きサンプルの需要を効果的に削減することが証明された。

6.2. 少ない訓練サンプル数での分類の安定性

サンプルサイズが小さい場合の分類の安定性を検証するために、以前の研究では、マルチソースの地理空間データに基づくRFアルゴリズムでは、学習サンプルの数は全土地区画の7％以下であるべきであると指摘した[35]。我々の研究もこの結論に同意する。半教師付きCo-Forestアルゴリズムでは、5％のサンプルサイズでも、許容範囲内の分類精度のばらつきを確保できる。我々の研究はまた、訓練サンプルサイズが7%より大きい場合、半教師付きアルゴリズムが教師付きアルゴリズムより優れた性能を達成することを示した。

6.3. 限界と不確実性

ほとんどの先行研究では、詳細な都市土地利用区分ではなく、大まかな土地利用／土地被覆区分の分類に焦点が当てられている。セマンティック・セグメンテーションが難しいため、詳細な都市土地利用分類の精度の高い結果を得ることは難しい。深圳のようなメガシティにおける詳細な都市土地利用分類の全体的な精度は、0.76より低いと報告する学者もいる[35,36]。本研究では、分類タスクの絶対精度よりも、小さなサンプルサイズに基づく土地利用分類を改善するために、マルチソースの地理空間データと半教師付き分類器を適用することの有効性に重点を置いた。

本研究では、主に純度の高い土地利用サンプル（例えば、支配的なタイプが土地区画内の面積の90％以上を占める）を学習サンプルとした。しかし、急速に発展した都市である深圳には、「商業と住宅が混在する」土地など、さまざまな土地利用の混合モデルがある。混合利用モデルには、水平空間と垂直空間における混合が含まれる。このため、分類にある程度の不確実性が生じる可能性がある。他の都市への一般化（例えば、ある都市から別の都市へオリジナルのラベルを使用する）を考える場合、大きな課題は都市の土地利用構造の違いから生じる。同じ都市土地利用であっても、都市が異なれば物理的な特徴が異なる場合がある。したがって、モデルの汎化性能を検証するには、より多くの都市事例と土地利用ラベルが必要である。

7. 結論

本研究では、半教師付きCo-Forestアルゴリズムとマルチソースの地理空間データを用いて、サンプル数の少ない詳細な都市土地利用分類の有効性を検討した。都市土地利用分類の実践において、大量のラベル付きサンプルを収集することは非常に困難であり、高いコストがかかることから、安定した分類結果を維持するための最適なトレーニングセットの比率も検証した。深圳市をケースとした場合、半教師付きCo-Forest法は、訓練セット比率のレベルを下げれば（17-20%削減）、RFやXGBoostといった従来の教師付き分類器と同等の結果を示した。モデルの性能は、訓練サンプルサイズが合計で5%未満になると急激に低下した。したがって、分類精度の低下を許容範囲内に抑えるためには、5％以上の訓練サンプルが必要である。本研究はまた、分類精度を大幅に向上させたマルチソース、マルチモーダルデータの重要性を確認した。その中でも、POIデータと高解像度リモートセンシングデータの貢献度が高い。

今後、提案手法を他の変化の激しい都市に拡張し、汎化性能を評価する予定である。ラベル付きサンプルをより効率的に利用するために、教師なし強化アルゴリズムなどのデータ強化手法を導入し、既存のラベル付きサンプルとラベルなしサンプルに基づいて新しいラベル付きサンプルを生成する。また、より多くのラベルを作成することで、混合土地利用の特徴を分析する。