学位論文審査の要旨

博士の専攻分野の名称　博士 (情報科学)　　氏名　BAO Jiaqi

審査担当者　主　査　教　授　工藤 峰一
副　査　教　授　今井 英幸
副　査　教　授　田中 章
副　査　教　授　中村 篤祥

学位論文題名

On the Reliability and Robustness of Linear Generalized Regression Algorithms for Classification
（識別のための線形一般化回帰アルゴリズムの信頼度とロバストネスに関する研究）

現在、機械学習は車の自動運転のみならず ChatGPT などの質疑応答システムにおいても必須の技術となっている。中でも、各種の予測を可能とする「回帰」と、対象を適切に分類する「識別」は代表的な技術である。両者には既に多くの学習方式が提案され有効性も確認されている。しかし、クリーンで正しくラベル付けされた大量データを仮定して構築された理論モデルでは実際の場面で十分な性能を発揮できないことが問題となっている。そこで、本研究は、実際に入手可能な「不完全データ」からの学習でも十分に性能を発揮することができる識別器の構成手法の開発に取り組んでいる。具体的には、三つの代表的な学習の枠組それぞれにおいて、想定される不完全要因に対処できる方式を検討している。

第一部では、ラベル付けされたデータが数において限られ、さらにデータに雑音が混入する場合や例外がある場合を検討している。この問題には、ラベル付けされてない多量のデータを利用する「半教師つき学習」のモデルに雑音耐性を付与することを考え、(RER: Robust Embedding Regression) という手法を提案している。具体的には、線形回帰モデルに正則化項として低ランク要請を加え、さらに項の評価を核ノルムなどで行うことでスパース性も併せて要請する目的関数を提案した。これにより、特徴に重畳する雑音を抑制し、例外データを除去することに成功している。

第二部では、データが不足している問題領域でも高性能な識別器を構成できるように類似している別な領域から豊富なデータを借りてくる「転移学習」に注目している。異なる二つの領域のデータ分布が特徴空間では移動によりほぼ対応つく場合を想定し、その差が同じ低次元特徴空間で吸収されるようにそれぞれの空間での射影を適切に実現している。さらに、ラベル間の分離性を強調するように特徴空間でのラベル配置を行っている。これらをやはり線形回帰モデルにおける正則化の追加として実現することで (RTL: Robust Transfer Learning) という手法を提案している。

第三部では、一つの対象に複数のラベルが与えられる「マルチラベル識別」という問題において、訓練用のデータに冗長にラベルがつけられている場合を扱っている。これはクラウドソーシングなどで複数人がつけたラベルを統合する際に見られる現象である。冗長なラベルを適切に扱うために、データの特徴空間の近接関係を表すグラフとラベル空間での近接関係を表すグラフの二つのグラフを利用している。線形回帰モデルにおいてこれらの二つのグラフが示す近接関係を維持する項を加え、特徴からラベルへの変換にスパース性を要請するノルムを採用することで (ADGD:

Adaptive Dual Graph Disambiguation) という手法を提案している。

本論文の貢献は以下にまとめられる。

1. 実際のデータには理想的なデータでは見られない各種の不完全要因が存在する。本研究はそのような要因に対して頑健な線形回帰モデルを構築し、その有効性を実際の識別問題において示した。
2. 線形回帰式において、必要な耐性をもたらす項を正則化項として追加するとともに各項の評価を適切なノルムで行うことの効果を一般的に示した。
3. 半教師つき学習、転移学習、マルチラベル識別、という三つの枠組みそれぞれにおいて RER, RTL, ADGD という線形回帰モデルをそれぞれ提案し、それらの有効性を定性的に論じた。
4. 三つの提案手法の有効性を実際のデータにより定量的に示した。RER は画像識別において、ごく少数 (4,5 個) のラベル付きサンプルからでも多量のラベルなしサンプルを利用して高性能の識別器を構成できること、さらにオクルージョンがあっても対応できることを、5 つのデータセットにおける 9 つの従来手法との比較で示した (50%〜100% の割合で 1 位)。RTL は画像とテキストの識別において 6 つの従来手法を凌ぐ転移学習性能を示した (ディープ特徴と深層学習を組み合わせた場合には同等で、それ以外の場合はほぼ 1 位)。ADGD は冗長ラベルを含む 7 つのマルチラベルデータセットにおいて約 5 割の割合で 7 つの従来手法を抑えて 1 位を得た。

これを要するに筆者は、機械学習の主題の一つである識別問題において、量的に不十分であるばかりか雑音や例外を含むといった現実のデータからでも適切に識別規則を学習する方式を開発し、三つの典型的な枠組みにおいて定性的・定量的にその有効性を示した。この成果はパターン認識の分野に貢献すること大なるものがある。よって、著者は、北海道大学博士 (情報科学) の学位を授与される資格あるものと認める。