HOKKAIDO UNIVERSITY

| | |
|---|---|
| Title | On Improving Multi-Label Classification via Dimension Reduction [an abstract of dissertation and a summary of dissertation review] |
| Author(s) | 孫, 露 |
| Degree Grantor | 北海道大学 |
| Degree Name | 博士(情報科学) |
| Dissertation Number | 甲第12850号 |
| Issue Date | 2017-09-25 |
| Doc URL | https://hdl.handle.net/2115/67456 |
| Rights(URL) | https://creativecommons.org/licenses/by-nc-sa/2.1/jp/ |
| Type | doctoral thesis |
| File Information | Lu_Sun_review.pdf, 審査の要旨 |

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

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　孫 露

審査担当者　主　査　教　授　工藤 峰一
副　査　教　授　小野 哲雄
副　査　教　授　杉本 雅則

学位論文題名

On Improving Multi-Label Classification via Dimension Reduction
（次元縮約によるマルチラベル識別の改善）

近年、一つの対象に一つのラベル (名前) を与える従来の「パターン認識」が拡張され、一つの対象に複数のラベルを与える「マルチラベル識別」が注目を浴びるようになった。ウェブページに適切なタグを複数つける、記事や著書を複数のジャンルに分ける、症状から複数の病名候補を挙げる、などの応用を持つ。本論文はこのマルチラベル識別において基礎的な方法論を検討したものである。

マルチラベル識別では次の二点が特徴的である。第一に、ラベルの候補が膨大 (例えば、ウェブページのラベル候補は数百万) である一方、一つの対象に関連するラベルは少なくそれらに強い相関がある。第二に、問題の規模が通常非常に大きい。サンプル数、特徴次元数、ラベル数のどれも、あるいは一部が万を超すのは当たり前である。この規模の問題は従来の方法論の直接の適用を阻む大きな要因となっている。よって、実用的なマルチラベル識別を高い精度でかつ実時間で行うには、これらを十分考慮する必要がある。本論文は、これらを順に踏まえて、新しいモデルならびに識別器の提案を行っている。

第一部では、主に、ラベルの相関を適切に扱う方法を検討している。これまでも Classifier Chain など多くの方法が提案されてきたものの、1) 識別性能が Chain の順番に強く依存する、2) 先行するクラス識別の結果を次のクラスの識別に利用するため、誤差の伝搬が起きる、3) 冗長な特徴が多く識別性能を阻害する、問題があった。本論文では、ラベルの依存関係を Polytree という特殊な確率構造で表現することで、最初の二つの問題を大きく解決できることを示した。また、最後の問題に対しては、2 段階特徴選択法を提案するとともに、特徴選択の統一的な基準を「条件付き尤度最大化」として定式化した。これらの技術を合わせて、よく使われる 12 データセットにおいてこれまでに提案されている代表的な 4 識別器と比べ平均して 11.7% ほど識別率 (ラベルの正解集合を正しく予測した割合) を向上させた。また、特徴数も元の数の平均 28.2% に削減した。

第二部では、主に規模の問題を扱い、サンプル数の削減、特徴空間の次元圧縮、ラベル空間の次元圧縮をそれぞれマルチラベル識別の枠組みで検討した。特徴空間の次元圧縮については、ラベルのないサンプルも使った学習 (半教師つき学習) の枠組みで検討を行った。目的関数を、ラベルつきサンプルにおける推定誤差項、圧縮行列のスパース項、ラベルなしサンプル間の距離保存項、ラベルの共起関係項、の 4 項の線形和で表現した。ラベル空間の次元圧縮においては、その目的関数に、ラベルの埋込み項を加える修正を行った。サンプル数の削減に関しても、これまでの大多数の方式とは異なり、特徴の部分集合においてクラスタリングすることの効果を示した。これらの次元

圧縮方法により、従来の 5 つの特徴選択手法と比べて、6 データセットにおいて平均して 3.4% ほど識別率 (今回は、Micro-F1 というラベルの正解比率に基づく指標) を向上させた (選択特徴が全体の 60% の場合)。半教師つき学習としては、ラベル付きサンプルが全体の 10% のときに、ラベルの付いてないサンプルを使うことで識別性能を 3.7% 向上させた。サンプル数の削減においては、18 データセットにおいて平均して 32.8% ほど識別率を向上させるとともに、最も少ない処理時間を記録した。

本論文の貢献は以下にまとめられる。

1. マルチクラス識別において重要な性質、ラベル間の相関、をより適切に扱うことのできる確率構造を提案し、これまでの方式と比べて 1 割程度識別性能を向上させることに成功した。
2. 大規模問題を効率的に扱うために、特徴数、ラベル数、サンプル数それぞれの縮小方法を提案した。結果として、識別性能を殆ど損なわずに特徴数を 60% 削減できることを示したほか、ラベルおよびサンプルにおける削減効果を識別性能と処理時間の両方において確認した。
3. 特徴空間とラベル空間両方の次元を同時に圧縮する統一的基準を提案した。これにより、より柔軟かつ合理的に次元圧縮することが可能となった。

これを要するに筆者は、パターン認識の急成長分野であるマルチラベル識別問題において、数理科学的方法論に基づいて種々の解析を行うとともに、これまでの識別器の性能を上回る識別器の設計および計算量の削減を行った。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。