様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　　７　日現在

機関番号：　１０１０１
研究種目：　基盤研究(B)（一般）
研究期間：　2015～2017
課題番号：　１５Ｈ０２７１９

研究課題名（和文）eXtFS:超大規模マルチラベル分類問題における特徴選択と特徴探索

研究課題名（英文）eXtFS:Feature Selection and Exploration in extremly large multi-label classification problems

研究代表者

工藤　峰一（Kudo, Mineichi）

北海道大学・情報科学研究科・教授

研究者番号：６０２０５１０１

交付決定額（研究期間全体）：（直接経費）　　10,500,000円

研究成果の概要（和文）：本研究では、一つの対象に適切な複数のキーワードや名前をつける「マルチラベル識別」問題において、識別性能を上げることと、性能を下げずに処理速度を上げることを目的として各種の検討を行った。その成果は主に三つである。第一に、ラベル間の相関利用の重要性を示し、その利用法を各種手法により示した。第二に、実用的な処理時間を確保するために、特徴空間あるいはラベル空間におけるサンプルの近さに注目した「問題分割」が有効であることを論じ、その効果を実験的に示した。最後に、滅多に出現しないラベルや付け忘れたラベルがラベル全体に対して多くの割合を占めるため、これらに関しては特別な取扱いが必要であることを指摘した。

研究成果の概要（英文）：We have dealt with "multi-label classification" problems where an object is assigned multiple labels. This study aimed at raising the classification performance and speeding up without degradation of performance. Our achievement is three of the following. First, we have pointed out the importance of the correlation between labels and showed several ways using it. Second, to keep a realistic processing time, we showed that the problem division of samples on the basis of their features or labels in some experimental results. Last, we pointed out the necessity of a special treatment on labels that appear rarely or have been forgotten to assign.

研究分野：　パターン認識

キーワード：　マルチラベル識別　スケーラビリティ　確率構造　同時可視化　埋め込み

１．研究開始当初の背景

コンピュータやネットワークの発展に伴い、パターン認識の対象も広範囲に広がるとともに問題が多様化・大規模化し、それらに呼応するように、目的自体も変化してきた。これまでの一つの対象を一つのクラスに分けること（シングルラベル識別）から、一つの対象を複数のクラスに分けるように拡張された（マルチラベル識別）。ウェブへのキーワード付与や画像中の対象の列挙などが代表的な例である。研究当初は、シングルラベルの識別方法を拡張する試みが盛んに行われた。しかし、ラベルの組合せがラベル数の指数的な数になること、それに伴い、それぞれの組合せの学習に使えるサンプルが非常に少ないこと、そもそも、テールラベルと呼ばれるような、滅多に出てこないラベルの方がよく出てくるラベルより多い、といった特殊性が明らかになるにつれ、基礎に根差した研究が必要となった。本研究は基礎に立ち返り、問題の所在を明らかにするとともに、適切な提案を行うことを意図して開始された。

２．研究の目的

本研究では、（超）大規模マルチラベル識別問題においてこれまで以上に識別性能を向上させるとともに、訓練および検査における時間・空間計算量の削減を図ることでスケーラビリティを確保することを目的とする。

３．研究の方法

問題の分析および従来手法の分析を行い、それらの得失を明らかにした上で主に以下の検討を行う。１）識別子の性能向上のために、特徴－特徴間、特徴－ラベル間、ラベル－ラベル間の相関を調べ、どれがどれほどの識別情報を持っているか、あるいは、持っていないかを明らかにする。その上で、ラベル間に適切な確率構造を設定して学習効率の良い識別方式を提案する。２）特徴選択など、冗長性を減ずる方式を検討する他、大規模問題を複数の中・小規模問題に分けることでスケーラビリティを確保する。

４．研究成果

成果は大きく三つの視点から整理される。

（１）識別子の性能向上について

（１）－１　ラベル数が数百から数十万にものぼるこれらの問題では、ラベルの相関情報が最も重要であることは予見されたものの、その量的な評価と適切な利用法については十分考慮されてこなかった。本研究では、classifier-chain と呼ばれるベイジアンネットの枠組みを検討し、その枠組みの中で中間的な複雑さを持つ Poly-Tree と呼ばれる確率構造（次頁の図）が適切な複雑さを持つことを見いだし、それにより識別性能を上げられることを示した〔⑫〕。

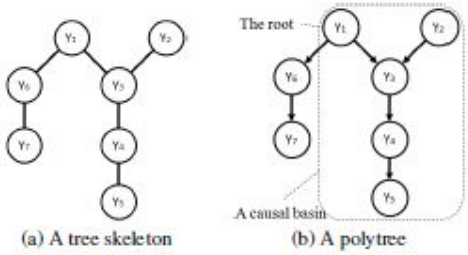
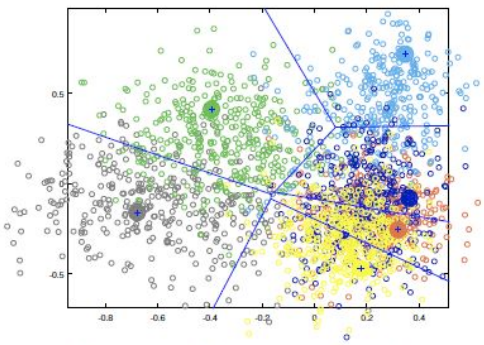


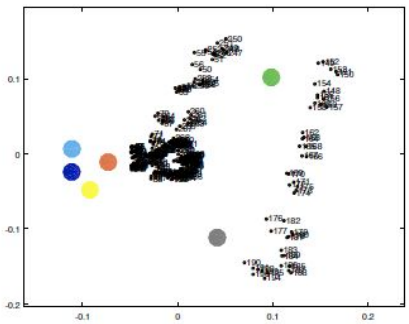
Figure 3: Example of polytree with its latent skeleton.

（１）－２　特徴選択は、規模の縮小だけでなく性能向上にも有効であることが知られている。しかし、本研究では、ラベル集合に応じて異なる特徴を選択することの重要性を指摘し、その効果を実験的に確かめた〔⑥〕。さらにこの方法を改良し、ラベルの相関と特徴選択を同時に考慮した方式を開発した〔①〕。

（１）－３　問題の理解のために、サンプル、特徴、ラベルの任意の二つの組合せを同時に可視化する手法を提案した。これにより、一つのラベル集合とその各々の要素ラベルがどのような関係にあるかを直感的に把握することができるとともに、特徴選択やラベルのクラスタリングなどの効果を直感的に理解できるようになった〔⑨〕。（下図は、サンプルとラベル、特徴とラベルの同時表示例）



(a) Single-label 6 classes by SL-A (C.R.:45.8%)



（２）問題の分割について

マルチラベル識別問題では通常ラベル数だけでなく、サンプル数や次元数も上がり、訓練時だけでなく識別時においてもその時

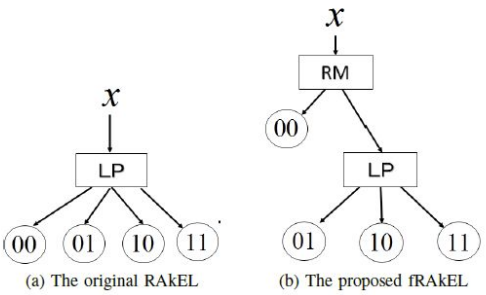
間・空間計算量を下げて実用性を得ることが重要である。特に、ラベルの組合せを予測するマルチラベル問題では、ラベル数の増加が指数的に影響する。そこで、各種の解析を行った。

（２）－１　局所性の解析について、サンプル、特徴、ラベルの局所性を定性的に整理し、特に、識別に有効な局所性を「有効局所性」と呼び、その利用法を検討した〔⑧〕。

（２）－２　サンプル集合の分割について、特徴空間の局所性に基づいた手法〔⑩〕およびラベルの局所性に基づいた手法〔⑧〕を提案し、効果を実験的に示した。

（２）－３　特徴選択および低次元への写像を検討し、二つの手法を提案した〔③、⑥〕。

（２）－４　ラベル部分集合の選択方法を検討し、これまでの方法に関して効率および性能を改良した〔⑦〕（次の図はそのアイディアの模式図）。



（３）まとめ

研究成果として提案した手法は７つに上る。本研究を通して、目標の一つであるスケーラビリティの確保は未だ十分に達成できたとは言えないものの、もう一つの目標である性能向上に対する試みはほぼ網羅したと思われる。しかし、残念ながら、結果としては性能を十分には向上させられてはいない。これは、研究途中に改めて気づかされたことではあるものの、出さえ現頻度が低いテールラベル、付け忘れて出現しなかったミッシングラベル、がラベル全体のうちの大多数を占めることがその要因である。今後はこれらのラベルに対する対処法を先に検討すべきである。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　５件）

① L Sun and M Kudo, “Optimization of Classifier Chains via Conditional Likelihood Maximization.” Pattern Recognition. 74(2018), 503-517.**(査読有)**

② Koji Tabata, Atsuyoshi Nakamura and Mineichi Kudo: An Efficient Approximate Algorithm for the 1-Median Problem on a Graph. IEICE Transactions 100-D(5): 994-1002 (2017). **(査読有)**

③ L Sun, M Kudo and K Kimura, “READER: Robust Semi-Supervised Multi-Label Dimension Reduction”, IEICE, E100-D-10(2017), 2597-2604. **(査読有)**

④ Ryo Watanabe, Junpei Komiyama, Atsuyoshi Nakamura and Mineichi Kudo, KL-UCB-Based Policy for Budgeted Multi-armed Bandits with Stochastic Action Costs. IEICE Transaction, E100-A(11)(2017), 2470-2486. **(査読有)**

⑤ Atsuyoshi Nakamura, Ichigaku Takigawa, Hisashi Tosaka, Mineichi Kudo and Hiroshi Mamitsuka, “Mining approximate patterns with frequent locally optimal occurrences”, Discrete Applied Mathematics, 200(2016), 123-152. **(査読有)**

〔学会発表〕（計　７件）

⑥ Lu Sun, Mineichi Kudo and Keigo Kimura, “Multi-Label Classification with Meta-Label-Specific Features.” in Proceedings of the 23rd International Conference on Pattern Recognition (ICPR 2016), Cancun, Mexico.

⑦ Keigo Kimura, Mineichi Kudo, Lu Sun and Sadamori Koujaku, “Fast Random k-labelsets for Large-Scale Multi-Label Classification.” in ICPR 2016, Cancun, Mexico.

⑧ Batzaya Norov-Erdene, Mineichi Kudo, Lu Sun and Keigo Kimura, “Locality in Multi-Label Classification Problems.” in ICPR 2016, Cancun, Mexico.

⑨ Mineichi Kudo, Keigo Kimura, Michael Haindl, Hiroshi Tenmoto, ” Simultaneous Visualization of Samples, Features and Multi-Labels.” in ICPR 2016, Cancun, Mexico.

⑩ Lu Sun, Mineichi Kudo and Keigo Kimura, “A Scalable Clustering-Based Local Multi-Label Classification Method.” in ECAI 2016, 261-268, 2016, The Hague, Netherlands.

⑪ Keigo Kimura, Mineichi Kudo, Lu Sun, “Simultaneous Nonlinear Label-Instance Embedding for Multi-label Classification.” in S+SSPR 2016, Merida, Mexico.

⑫ Lu Sun and Mineichi Kudo, “Polytree-Augmented Classifier Chains for Multi-Label Classification”,

In IJCAI2015, 3834-3840, Buenos Aires, Argentina.

〔図書〕（計　0 件）

〔産業財産権〕

○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕
ホームページ等

マルチラベル識別問題ツールキット：
https://github.com/KKimura360/MLC_toolbox
（これまでに提案されたマルチラベル識別手法を網羅するとともに、任意の組合せを容易に行えるよう工夫した。）

６．研究組織

(1)研究代表者
工藤　峰一（MINEICHI KUDO）
北海道大学・情報科学研究科・教授
研究者番号：60205101

(2)研究分担者
今井　英幸（HIDEYUKI IMAI）
北海道大学・情報科学研究科・教授
研究者番号：10213216

中村　篤祥（ATSUYOSHI NAKAMURA）
北海道大学・情報科学研究科・教授
研究者番号：50344487