

博士論文審査結果報告書

論 文 題 目

**Study on SVMs with Quasi-Linear Kernel for Imbalanced
Classification and Semi-Supervised Classification**

申 請 者

Bo ZHOU

情報生産システム工学専攻
ニューロコンピューティング研究

2018 年 9 月

パターン分類は機械学習において一つの重要な研究課題である。機械学習による分類の目的は、カテゴリの属性が分かっている観測を含む一連の訓練サンプルから学習することによって、特定のカテゴリに分類されるサンプルを予測することである。サポートベクターマシン (Support Vector Machine: SVM) は、マージン最大化によって優れた汎化性等の性能を持つ線形分類器として近年多くの注目を集めている。また、SVM はカーネル法によりデータを高次元特徴空間に写像することによって非線形分類問題を扱うことを可能にしている。しかし、多項式カーネルや放射基底関数 (Radial Basis Function: RBF) カーネルなど、SVM で使用できるカーネル関数は極僅かである。複雑な分類タスクでは、基本カーネル関数の線形結合や組み合わせによって、より複雑なカーネル関数を構築することができるが、基本カーネル関数の形が固定されているので、データによってカーネル関数を柔軟に合成することはできない。また、RBF カーネル関数のような基本カーネル関数を使用される場合には、観測空間から無限次元の特徴空間へのブラックボックス写像となり、ノイズに弱くなる恐れがある。このような基本カーネル関数を用いた SVM は、ノイズが多く数が少ない学習データを利用した場合には、過学習が発生しやすいという課題が残っている。

本研究では、上記課題の解決を目指し、データ依存型準線形カーネル関数を持つ準線形 SVM が提案されている (第 2 章)。準線形カーネル関数では、まず、機械学習アプローチにより複数の局所線形分類器を構築する。次に、局所線形分類器を補間し線形回帰で表わされる近似非線形分類器を構築することによって、多線形特徴空間 (Multi-Linear Feature Space: MLFS) という特徴空間を導入する。データは無限次元の特徴空間に写像される代わりに MLFS に写像される。これにより、ノイズに対してよりロバストな準線形 SVM を実現している。さらに、提案された準線形 SVM は、データ不均衡な分類問題と半教師付き分類問題に適用され、準線形 SVM に取り入れた局所線形分類器の情報を活用したクラス不均衡 SVM 分類器 (第 3 章) と半教師付き SVM 分類器 (第 4 章) の構築が行われている。

以下に、本論文の構成と各章の概略について述べ、評価を与える。

第 1 章では、非線形分類、SVM およびカーネル法、データ不均衡分類と半教師付き分類について述べ、本研究の背景・目的を明らかにしている。

第 2 章では、準線形 SVM の提案について述べている。非線形分類器の構築はクラス間の分離境界線を表現する非線形モデルを構築することであるが、従来の非線形 SVM では、この分離境界線に関する先見情報があっても利用しにくいという問題がある。そこで、本論文では、非線形分離境界線をいくつかの区分に分割し、各区分の分離境界線を局所線形モデルで近似する。また、これらの局所線形モデルを補間し非線形分離境界線を線形回帰で表現する非線形モデルを構築する。そして、SVM の定式化を適用して、マージン最大化でモデル

構造リスクを最小化するように上記の非線形モデルのパラメータを最適化する。その結果、補間関数と局所線形モデルの情報で合成される準線形カーネル関数を持つ準線形 SVM 分類器を実現している。このように得られた準線形カーネル関数は、クラス間の分離境界線に関する局所線形情報を活用することができ、データを無限次元の特徴空間に写像するのではなく、クラス間の分離境界線に関する局所線形情報で構成される MLFS に写像することで、ノイズに対してよりロバストになることが期待できる。さらに、分離境界線に沿った観測空間の分割を実現するために、改良型 k-means クラスタリング法を提案している。性能評価のため、タンパク質機能分類問題を対象とした数値実験では、RBF カーネル関数を用いた SVM では総合評価 F-score が 0.4031 であるのに対して、線形 SVM では、F-score が 0.5027 であるので、RBF カーネル関数を用いた SVM が過学習となっていることがわかる。提案された準線形 SVM では、F-score が 0.5326 であるので良い性能が得られていることが明らかにされている。また、提案した準線形カーネル関数は、データ依存型カーネル関数であり、カーネル関数学習法として展開することが可能であるので高く評価できる。

第 3 章では、準線形 SVM を利用した不均衡データを扱えるクラス不均衡 SVM 分類器の構築について述べている。クラス不均衡分類の目標は、不均衡データセットにおける少数派クラスを正しく特定することである。SMOTE (Synthetic Minority Oversampling Technique) はこの目標を達成するための有効なオーバーサンプリング法として知られ、少数派クラスのサンプルを新しく線形合成で生成し、バランスの取れたデータセットを提供することで不均衡問題を解決する。しかし、SMOTE では、クラス間の分離境界線が非線形である場合に、多数派クラスと重複した誤った少数派クラスのサンプルを生成して分類が難しくなる可能性がある。そこで、本論文では、新しい線形区分化 SMOTE を提案している。まず、ボトムアップ最小スパニングツリーベースのクラスタリング方法を導入し、観測空間を分離境界線に沿っていくつかの線形分離可能な局所パーティションに分割する。次に、SMOTE を適用し各局所線形パーティション内に限定することによって正しい少数派クラスのサンプルを生成する。最後に、バランスの取れたデータセットに対して、局所線形パーティションの情報を活用して準線形カーネル関数を合成し準線形 SVM 分類器を実現している。性能評価の数値例では、提案された線形区分化 SMOTE と従来の SMOTE を Yeast 等の実データからなる 14 データセットに適用して比較し、再現率 Recall では、7 データセットは提案された線形区分化 SMOTE が従来の SMOTE より優れており、1 データセットは同値、6 データセットは従来の SMOTE が優れていた（以後これを win/tie/lose で 7/1/6 と記す）が、総合評価 F-score と G-mean ではそれぞれ 14/0/0 と 10/0/4 となり、提案された線形区分化 SMOTE は従来の SMOTE より不均衡なデータ分類に有効であることを明らかにしている。

第 4 章では、準線形 SVM を利用した半教師付き SVM 分類器の構築について述べている。半教師付き分類ではラベル付きデータに加えてラベル無しデータも利用して分類器の構築を行う。従来、ラベル無しデータを利用するためのモデル仮定として、クラスタ仮定やグラフ仮定などがある。前者は、ラベルありラベル無しに関係なく、データ密度に基づいたクラスタを形成し、分離境界線が低密度領域を通ることを仮定する。一方、後者は、ノードがデータでエッジがデータ間の類似度でグラフを構築し、構築されたグラフ上でラベルありデータからラベル無しデータへラベルをラベルの変化が滑らかになるように伝搬することを仮定する。本論文では、クラスタ仮定とグラフ仮定を組み合わせた新しい半教師付き SVM 分類器の構築を行っている。まず、繰り返し DBSCAN (Density-based Spatial Clustering of Applications with Noise) クラスタリングアルゴリズムを導入し、密度クラスタを検出して局所線形分類器を構築する。次に局所線形分類器を補間し低密度領域を横切る近似非線形分離境界を形成することによって準線形カーネル関数を合成する。そして、グラフ構成に基づく Laplacian SVM 分類器を構築する。そこでは、準線形カーネル関数がデータ間の類似度計算ならびに Laplacian SVM 分類器のカーネル関数として活用され、半教師付き Laplacian SVM 分類器を構築している。クラスタ仮定とグラフ仮定を同時に利用することによって、高性能な半教師付き準線形 Laplacian SVM 分類器を実現している。性能評価の数値例では、提案された Laplacian SVM 分類器と従来の RBF カーネルを用いた Laplacian SVM 分類器を前出の 7 データセットに適用し比較した結果、総合評価 F-score と Accuracy では、それぞれ 7/0/0 と 6/0/1 となり、提案された半教師付き Laplacian SVM 分類器の有効性が明らかにされている。

第 5 章では結論として、本研究の成果について総括し、今後の研究課題を論じている。

以上を要約すると、本論文では、クラス間の分離境界線に関する局所線形情報を活用した準線形カーネル関数の合成を行い、準線形 SVM の提案を行っている。また、準線形 SVM をクラス不均衡分類問題および半教師付き分類問題に適用しその有効性を明らかにしている。これらの成果は機械学習、ニューロコンピューティング分野の発展に寄与するところ大である。よって、本論文は、博士（工学）の学位論文として価値あるものと認める。

2018 年 9 月 4 日

審査員

主査	早稲田大学	教授	博士(情報工学)(九州工業大学)	古月 敬之
副査	早稲田大学	教授	工学博士(早稲田大学)	吉江 修
	早稲田大学	教授	博士(工学)(早稲田大学)	藤村 茂