早稲田大学大学院情報生産システム研究科

博士論文審査結果報告書

論　文　題　目

Study on Multi-SVM Systems and Their Applications to Pattern Recognition

申　請　者

LI, Boyang

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

2010 年 7 月

最近、サポートベクターマシン（Support Vector Machine, 以下 SVM）はパターン認識手法の一つとして注目されている。未学習データに対して高い識別性能を得るための工夫があることとカーネルトリック（Kernel Trick）に基づいて非線形の識別関数を構成できることにより、SVM は、現在知られている多くの手法の中で最も認識性能の優れた学習モデルの一つとなっている。近年 SVM を実用化するための応用研究が盛んに行われている。実世界の複雑なパターン認識問題への応用には、(1)訓練データが増加すると SVM 訓練の計算量が膨大となる大量訓練データの問題；(2) 訓練データにパターン認識に寄与しない入力変数が多い場合、SVM の訓練時間が増えて認識精度が低下する高次元の問題；(3)分類クラスが不均衡な訓練データの場合、SVM 分離超平面の位置オフセットが生じ、分類精度が大きく低下するクラスタ不均衡の問題などが研究課題として残っている。一方、SVM は、サポートベクターを用いて、2 クラスのパターン識別器を構成する手法であり、訓練データから「マージン最大化」という基準でサポートベクターのパラメータを学習する。そのため、サポートベクターになる訓練データとそうでない訓練データでは、その重要性に大きな違いがある。

本論文では、複雑なパターン認識問題に適用できるマルチ SVM システムの構築を目指して、訓練データがサポートベクターになる可能性があるかどうかの違いに注目し、サポートベクターになる可能性のある訓練データの重要性の重みを大きく設定する方式を提案している。また、上記課題を実現するために局所 SVM 訓練用アルゴリズムを提案し、分割統治戦略によるマルチ SVM システムを構築している。

本論文は 7 章から成る。以下にその概要と評価を述べる。

第 1 章「Introduction and Motivation」では、本論文の背景と目的について述べ、本論文の研究の位置付けと意義を示している。

第 2 章「Fast SVM Training Based on Separation Boundary Detection」では、大量訓練データの認識問題に対して、訓練データの選択機構を有するSVM分類器の設計について述べている。SVMの訓練では、訓練データが増えると、訓練のための計算量が膨大となり、効率的な機械学習を実現するための訓練データの取捨選択が必要になる。そこで本章では、訓練データの数を減らすため、分離超平面の検出技術を導入し、サポートベクターになる確率が高い、分離超平面付近の訓練データを検出している。また、クラスタリング手法を使用して、サポートベクターになる確率が低い訓練データを、訓練データの分布情報を維持するために少数残し多くを削除する。これにより、提案手法はサポートベクターになる確率が高い訓練データをできるだけ多く保留し、パターン認識の精度を維持しながら訓練データの数を大きく減らすことができる。シミュレーションでは、二つの人工データと三つの実データを利用し提案手法の評価を行っている。従来手法と比べ、提案手法では、訓練データの数が平均で66.7%削減で

き、訓練時間が平均で60.5%短縮できるにも関わらず、認識精度はほぼ維持できることを明らかにしている。

第３章「Feature Selection Using Correlation-based SVM Filter」では、高次元データの認識問題に対して、パターン認識に寄与する特徴(変数)を選択する手法を提案している。パターン認識が対象とするデータは、例えば、マイクロアレイ分析やテキスト分類などで用いられるデータのように、非常に多くの変数で記述される場合が多く、パターン認識に寄与しない変数も多数含まれており、認識精度を低下させる原因になる。したがって、寄与する変数の選択が必要となる。本章の提案手法では、まず、K相関性クラスタリングを用いて、変数の集合を相関性の高い部分集合（クラスタ）に区分する。次に、SVMに基づいた感度分析を行い、感度に基づき変数の重要性をランキングする。最後に、クラスタから重要性の高い変数を選択し、選択した変数(特徴)を使用すると、重要性と多様性が維持できることを明らかにしている。また、高次元実データを利用したシミュレーションでは、従来の特徴選択手法では特徴量が平均で48.7%程度しか減少しないのに対して、提案法では平均で63%減少することを示している。GA等に基づいた手法と比較して、提案手法では認識精度がほとんど低下せず、認識時間・記憶容量が節約可能であり、そのため有用性が高いと評価できる。

第４章「Classification Using Fuzzy Decision-making SVM」では、訓練データに混入したノイズの影響とクラスの不均衡の問題に対し、ファジー論の視点からソフトな分離超平面の構築及び位置オフセットパラメータの計算について述べている。実データのサンプルに何らかのノイズが混入している場合、特に分離超平面付近にノイズが混入している場合は、認識結果への影響が大きい。この問題に対応するため、信頼度計算に基づくソフトなSVMファジー分離超平面を提案している。この分離超平面は、ファジーロジックに基づき、伝統的なSVMの制約を緩め、データの分類結果を調整し、分類器の分類精度の向上を図っている。また、実世界の識別問題では、クラス間のデータが不均衡である場合が多く、サポートベクターの分布も不均衡になる場合が多い。この問題に着目し、サポートベクターに対する加重調和平均法を提案している。これにより、分離超平面の位置オフセットパラメータを計算し自動校正を行っている。評価実験の結果、提案手法はノイズやクラス不均衡の影響を減少することができ、その有効性を高く評価できる。

第５章「Multiple Support Vector Classifier System」では、第2～4章で述べた単一SVM認識器を局所SVMとして用い、パターン認識の精度や訓練速度などの性能を高めるために、分割統治戦略に基づいたマルチSVM識別システムを提案している。個々の識別器を単体で用いた場合より、複数の識別器を統合し、アンサンブル識別器を構成し、アンサンブル学習法により、更に高い精度が得られることをシミュレーション実験で確認している。本章の提案手法では、従来の手法と異なり、分離超平面の検出技術を利用し分離超平面付近の領域の検出

を行い、分離超平面の周り領域に対して、区分補間融合技術に基づき、データ領域の分割を行う。また、分類器を組み合わせて分類を行う手法として、重み付き多数決による分類法を提案している。人工的に作成したデータを使用して評価を行った結果、マルチSVM識別システムは、既存手法に比べ平均で約5%の分類精度が向上するだけでなく、訓練時間が平均で50%減少することを明らかにしている。また、UCIベンチマークデータによる評価においても同じような結果が得られていることを示しており、提案手法の有用性が高いと評価される。

第6章「Multiple Support Vector Regression Network」では、時系列予測問題に対して、複数のサポートベクター回帰モデル（SVR）の統合と、性能のより高いモジュラーSVR予測ネットワークの構築と応用について述べている。従来の手法との違いは、提案手法は、次元の高い入力でなく次元の低い出力であるファイナンス時系列データの価格域を分割し、局所的な学習データを使用し、個々の回帰モデルの訓練を行うことが特徴である。異なる価格域では入出力関係が大きく違う可能性があり、そのため異なるモジュールで学習し、その結果を統合することによって、時系列の価格と変動の予測を実現している。5種の外国為替レートのデータに対し行った予測評価実験では、従来の単体SVR予測器と比較して、提案するSVR予測ネットワークは、予測誤差が減少し正解率を向上することを明らかにしている。具体的には、移動平均データの場合、予測誤差が平均で76.2%減少し、正解率が平均で約6.5%向上することを示している。

第7章「Conclusions」では、本研究により得られた成果を総括し、今後の研究課題について論じている。

以上を要約すると、本研究は、SVMを実世界の複雑なパターン認識問題へ適用するために、局所SVM訓練のための訓練データ選択法と特徴選択法および分離超平面の位置オフセット自動校正法などを提案している。これにより、新しい分割技術を導入したマルチSVM分類システムおよびマルチSVM回帰システムが可能になることを明らかにし、その有効性を示している。これらは、SVMの発展に大きく貢献し、計算知能分野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。

2010年6月18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
	早稲田大学	教授	工学博士（九州大学）	平澤宏太郎
	早稲田大学	教授	博士(工学)(九州工業大学)	鎌田清一郎