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 the Predictions of Gene Function and Protein Structure Using Multi-SVM and Hybrid EDA

申　請　者

CHEN, Benhui

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

2011 年 1 月

近年、多くの生物対象に実施されているゲノムプロジェクトによって大量なDNA配列・アミノ酸配列のデータが得られている。これらの配列情報における遺伝子識別は簡単ではない。そのため、既知の遺伝子配列データベースの計算機予測モデルを発展させ、新たに生成した配列情報のどこにどのような遺伝子情報があるのかを予測するバイオインフォマティクスの方法が注目されている。現在、このようなバイオインフォマティクス問題に用いられるアプローチには、統計学的モデル、人工ニューラルネットワーク、隠れマルコフモデル、サポートベクターマシン(Support Vector Machines: SVM)、進化的アルゴリズム（Evolutionary Algorithm: EA）などの機械学習手法がある。これらを利用した配列比較・配列検索、進化系統樹推定、DNA・タンパク質の機能予測、RNA・タンパク質の構造予測、相互作用・ネットワーク推定などの研究が盛んに行われている。本論文では、DNA・タンパク質の機能予測とRNA・タンパク質の構造予測という二つの課題を取り上げいる。前者は分類問題であり、SVMが、また、後者は最適化問題で、EAの一種である分布推定アルゴリズム（Estimation of Distribution Algorithm: EDA）が適用できる。

一方、バイオインフォマティクス問題は不確かさが大きく非常に複雑であるため、従来のSVMやEDAなどの機械学習手法をそのまま適用しても、予期した結果が得られず、予測精度や探索能力が非常に低くなってしまう点が指摘されている。そこで、本論文では、応用対象から何らかの構造的情報を抽出活用し、SVMやEDAなどの機械学習手法の性能を向上することを目指して、①汎化能力を調整することが可能なマルチSVMシステム、②分離超平面微調整機構を有する多重ラベル分類器、③適応的ニッチング進化戦略を導入したハイブリッド型EDAなどの提案を行い、予測精度や探索能力の向上を図っている。

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「SVM with Composite Kernel for Classification」では、汎化能力が調整可能なマルチSVMシステムの構築について述べている。遺伝子分類では、訓練データの次元が高く、サンプル数が少なく、かつ強い雑音が含まれる場合が多い。従来の非線形SVMは過学習問題を抱えており、上記の状況では学習性能が保証できない。そこで本章では、マルチ線形SVMシステムにより汎化能力が調整可能なSVMシステムを構築し、これにより過学習問題を改善している。提案方式では、従来のマルチSVMシステムにおけるローカルSVM訓練の代わりに、入力空間の分割情報に基づいて適切なカーネルを合成し、マルチ線形SVMシステムをシングルSVMと同様に実現する特長を持っている。シミュレーションでは、遺伝子機能分類の10個のベンチマーク問題に適用し評価実験を行った。従来のRBFカーネルを持つSVMでは過学習が生じ線形SVMより精度が悪くなるのに対して、提案方式では過学習が発生することがなく線形SVMより平均で評価スコア（F-score）が3.1%向上することを明らかにしている。

第３章「Multi-label Classification Based on SVM for Functional Genomics」では、遺伝子機能を予測するための改良型多重ラベル分類器の設計について述べている。多重ラベル分類問題は多クラス分類問題の拡張であり、一つのサンプルは一つのクラスだけに分類されるのではなく複数のクラスに分類可能である。多クラス問題と同様にOne-Versus-RestあるいはOne-Versus-Oneの手法で多重ラベル分類問題を複数の２クラス分類問題に変換し、複数の２クラス分類器を組み合わせることで多重ラベル分類器を構築するが、クラス間のオーバーラップや分類器の固定閾値は従来の多重ラベル分類器の精度を低下させる原因となる。そこで本章では、オーバーラッピング領域を重視した分離超平面微調整機構を有するSVM分類器を構築し、複数のSVM分類器の組み合わせにより精度のよいラベルランキングを実現している。これをベースにして、K近傍法に基づいたサンプル依存閾値機構を導入することによって多重ラベル分類器の精度を向上している。ゲノムベンチマークデータを使用した評価実験の結果、提案手法は従来のSVM多重ラベル分類器より、評価スコアを8.4%改善しており、その有効性が高く評価できる。

第４章「Hierarchical Multi-label Classification for Gene Function Prediction」では、遺伝子機能予測のための階層的多重ラベル分類器の設計について述べている。階層的多重ラベル分類問題では、クラスのラベルがお互いに独立ではなく根付き木構造の関係を持つため、True Path Rule（TPR）という拘束条件を満たさなければならない。クラスに対応している木のノード毎にOne-Versus-Rest方式でSVM分類器を適用し、根付き木構造の関係を持つ複数のSVMを組み合わせることで階層的多重ラベル分類器を構築している。本章では、クラスのラベル間に拘束条件のあるTPRを分類器の精度の向上に活用すると同時に、２クラス分類器の分類誤差の影響が低減できる重み付改良型TPR分類器を提案している。一方、遺伝子機能予測における階層的多重ラベル分類問題では、通常数百のクラスがあるため、One-Versus-Rest方式で、正例（Positive）が負例（Negative）より遥かに少ないという高度なクラス不均衡になる場合が多くある。この不均衡の影響を低減するために、本章では階層型SMOTE（Synthetic Minority Over-sampling Technique）法を提案し正例のサンプルを生成して不均衡を解消している。酵母（Yeast）の４つのデータセットによるベンチマーク評価実験の結果、従来の階層的分類器や階層的TPR分類器と比べて、平均で評価スコアがそれぞれ16%と8.6%改善できることを明らかにしている。

第５章「Protein Structure Prediction on HP Model Using Hybrid EDA」では、HPモデルに基づいたハイブリッドEDAによるタンパク質構造の予測について述べている。HPモデルでは、配列上のすべてのタイプのアミノ酸が疎水性（Hydrophobic: H）または親水性（Polar: P）の２種類だけに簡単化されて、また、２次元あるいは３次元正方格子モデルでは、エネルギーが最も低くなるような自己回避折りたたみ構造を求めている。これらはNP完全問題であるため、最適な構造を見つけることは非常に困難であるが、最適構造ではHアミノ酸が

中心でPアミノ酸が周辺に存在すると言われている。そこで本章では、できるだけ短い時間で良い構造を見つけるために、この知見を重視し、低いエネルギーだけでなく大きいH核を形成しやすくなるような合成型適応度関数を導入し、さらに自己回避規則を違反した無効な個体の修復のための高効率なBR（Backtracking Repairing）法を導入したハイブリッド改良型EDAを構築している。11本のHP配列に適用したベンチマークによる実験の結果、従来のEDAでは最適構造が見つけられなかった一部の長いHP配列に対して、提案ハイブリッド改良型EDAは最適構造を発見できるだけでなく、長さ50以上のHP配列において、平均で44.8%のCPU計算時間が短縮できることを示している。

第６章「Protein HP Model Folding Based on Adaptive Niching EDA」では、適応的ニッチングEDAを導入したHPモデルによるタンパク質構造の予測について述べている。EDAは他のEAと同様に複雑な問題に適用する場合に早熟収束という問題点がある。そこで本章では、適応的ニッチングEDAを提案している。適応的ニッチングEDAでは、まず、AP (Affinity Propagation)クラスタリング法でEDAの集団を分割し適応的ニッチを構成する。次に、個体の適応度や探索空間の探索履歴情報などを取り入れ、確率選択によるボルツマン(Boltzmann)機構で適応的に個体の選択を行って多様探索と集中探索のバランスをとり早熟収束問題に対処している。8本のHP配列に適用したベンチマーク評価実験の結果、従来のEDAより、提案方式では最適構造の発見確率が平均で27.8%向上することを明らかにしている。

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やEDAなどの機械学習手法をバイオインフォマティクス問題へ適用するために、SVMの汎化能力調整法、分類器の分離超平面微調整法、高効率なBR法、重み付TPR階層的多重ラベル分類器および適応的ニッチングEDAなどの提案を行い、シミュレーション実験によりその有効性を示している。これらは機械学習、バイオインフォマティクス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 12 月 9 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（九州大学）　平澤 宏太郎

早稲田大学　教授　博士(工学)(九州大学)　岩井原 瑞穂

北九州市立大学 准教授　博士(工学)（慶應義塾大学）　孫 連明