（ 続紙 １ ）

京都大学	博士（情報学）	氏名	劉　鵬宇　（LIU Pengyu）
論文題目	Extracting Rules from Trained Machine Learning Models with Applications in Bioinformatics （機械学習モデルからの知識抽出と生命情報学への応用）		

（論文内容の要旨）

本論文は、データから機械学習による構築した数理モデルから知識を抽出するための手法とその生命情報学への応用について述べられており、５章から構成されている。

第１章では、研究の背景と動機、成果の概要、論文の構成について述べている。

第２章では、本論文で利用する機械学習モデル、具体的には、階層型ニューラルネットワークおよび勾配ブースティング（Gradient Boosting）法について説明している。

第３章では、線形閾値関数を活性化関数とする階層型ニューラルネットワークを対象に、学習ずみのニューラルネットワークモデルから知識を抽出するための二種類の手法を提案し、その計算量を解析している。一つ目の手法は、各ニューロンに割り当てられた線形閾値関数からブール関数という表現形式で知識を抽出し、それらを組み合わせるというものである。具体的には、線形閾値関数がNested Canalyzing Functionというブール関数の部分クラス、および、多数決関数という部分クラスに対応する場合に、線形閾値関数からそれぞれの表現形式で正確にブール関数を出力するアルゴリズムを提案している。さらに、これらと再帰型の手続きを組み合わせ、任意の線形閾値関数からブール関数を抽出するアルゴリズムを提案している。そして、これらのアルゴリズムの時間計算量を解析するとともに、各ニューロンにこのアルゴリズムを適用しその結果を統合することにより、ニューラルネットワーク全体に対するブール関数を抽出する手法を提案している。二つ目の手法は、入力変数に対する0,1ベクトルが一様分布に従うことを仮定したときに、一部の入力変数値に対する出力変数値の条件付き確率として知識を抽出するというものである。この問題に対して動的計画法に基づくアルゴリズムを開発し、その時間計算量を解析している。さらに、一つ目の手法については生物情報ネットワークデータおよび公開データを用いた計算機実験、二つ目の手法についてはシミュレーションデータおよび公開データを用いた計算機実験を行い、その有効性を評価している。

第４章では、RNA切断酵素によるマイクロRNAの切断部位をマイクロRNAの配列データから予測するための新規手法を提案している。これまでに配列情報および（既存の予測プログラムにより得られた）RNA二次構造情報に基づく特徴ベクトルとサポートベクターマシン（SVM）を組み合わせた予測手法が提案されていたが、本研究においては、RNA二次構造情報および相補配列情報を利用した特徴量と配列間のクラスタリング結果を反映した特徴量に基づく特徴ベクトルを新たに開発し、それを勾配ブースティング法に入力として与えることにより、学習データからモデルを学習し、新たなデータに対して予測を行う手法を提案している。さらに、勾配ブースティング法の学習結果から配列上の各位置の重要性を抽出することにより、配列データから知識を抽出する手法を提案している。そして、マイクロRNAの切断部位に関するベンチマークデータを用いて、SVMを用いた二種類の既存手法と比較し、その有効性を評価するとともに、統計解析により得られた知識の評価を行っている。

第５章は結論であり、本研究をまとめるとともに、今後の研究の方向性や課題について述べている。

（論文審査の結果の要旨）

本論文は、階層型ニューラルネットワークからブール関数および条件付き確率の形式で知識を抽出する手法、および、勾配ブースティング法を用いたRNA切断酵素によるマイクロRNAの切断部位予測のための学習・予測手法と知識抽出手法について述べたものであり、得られた成果は以下のとおりである。

（１）ニューラルネットワークは様々な予測問題に応用されているが、学習結果の解釈が困難である場合が多い。そこで、この問題の克服に向けて、線形閾値関数に基づく階層型ニューラルネットワークからブール関数および条件付き確率の形式で知識抽出を行う新規手法を開発した。前者においては、閾値関数がNested Canalyzing Functionおよび多数決関数というブール関数の部分クラスに対応する場合に、閾値関数からそれぞれの形式でブール関数を抽出するアルゴリズムを開発し、それらが多項式時間で動作することを示した。さらに、それらのアルゴリズムと再帰的な手法を組み合わせた指数時間アルゴリズムを開発し、個々の頂点からの抽出結果を組み合わせてネットワーク全体に対するブール関数を抽出する手法も開発した。後者においては、一部の入力変数に対する出力変数の条件付き確率という形式で知識を抽出するという問題を定義し、２個の頂点からなる隠れ層が１層しかない場合でもNP困難となることを示した。一方、定数個の頂点からなる隠れ層が１層だけの場合に対する動的計画法に基づくアルゴリズムを開発し、それが擬多項式時間で動作することを示し、さらにその結果を定数個の隠れ層を持つ場合に拡張した。そして、前者においては、生物情報ネットワークデータを用いた計算機実験により生物学的に有用な知識を抽出できることを示すとともに、公開データを用いた計算機実験により既存のブール関数抽出法と同程度以上の精度を持つブール関数を抽出できることを示した。後者においては、シミュレーションデータを用いた計算機実験によりランダムサンプリングに基づく単純な方法と比較し、より正確に確率を推定できることを示すとともに、公開データに適用し有用と考えられる規則を抽出できることを示した。

（２）RNA切断酵素によるマイクロRNAの切断部位予測はマイクロRNAの機能推定に役立つ可能性があり、これまでサポートベクターマシン（SVM）を用いた手法などが提案されてきた。本論文では、この問題に対する予測精度を向上させるために、配列データや予測された二次構造データに基づく特徴量に加え、配列間の編集距離をもとにAffinity Propagation法というクラスタリング法を適用することにより得る新規の特徴量を提案した。そして、その特徴量を勾配ブースティング（Gradient Boosting）法という機械学習手法と組み合わせた予測手法を開発した。ベンチマークデータを用いた計算機実験によりSVMを用いる２種類の既存手法と比較し、多くの場合においてより高い予測精度が得られることを示した。さらに、勾配ブースティング法により得られる特徴量の重要度指標をもとに各配列位置の重要度を評価し、マイクロRNAの前駆体配列の中央に近い位置が切断部位予測において重要であるとの仮説を得た。

以上、本論文では機械学習モデルからの知識抽出という情報学における重要な研究課題に取り組み、ニューラルネットワークからの知識抽出のための新規手法、および、マイクロRNA切断部位予測およびその配列位置の重要度抽出のための新規手法を提案し、それぞれの手法をバイオインフォマティクス関連データなどを用いた計

算機実験により評価した。提案手法のいずれもが新規性、有用性が高く、当該分野の発展のために十分な寄与をしている。よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、令和３年４月１５日、論文内容とそれに関連した事項について試問を行った結果、合格と認めた。なお、本論文のインターネットでの全文公表についても支障がないことを確認した。

要旨公開可能日：　　年　　月　　日以降