| | |
|---|---|
| 氏名 | 鹿島久嗣（かしまひさし） |
| 学位(専攻分野) | 博士（情報学） |
| 学位記番号 | 情博第240号 |
| 学位授与の日付 | 平成19年3月23日 |
| 学位授与の要件 | 学位規則第4条第1項該当 |
| 研究科・専攻 | 情報学研究科知能情報学専攻 |
| 学位論文題目 | Machine Learning Approaches for Structured Data<br>（構造データ解析のための機械学習法） |
| 論文調査委員 | （主査）教授 阿久津達也　教授 山本章博　教授 田中利幸 |


## 論文内容の要旨

本論文は，木構造およびグラフ構造の解析や予測のための機械学習手法について述べており6章から構成されている。

第1章は本論文の背景となる既存研究を概観し，動機，研究の独自性について述べている。特に，構造データを内部構造と外部構造の2種類に分類し，本論文の前半と後半において，それぞれに対するアプローチを述べている。

本論文の前半の2～5章では，DNA配列やXML，あるいは化合物などの内部構造をもったデータを扱うためのカーネル関数の設計について論じている。

第2章では，まずカーネル法の基本について導入的な解説を行った後，構造データに対するカーネル関数設計の一般的な枠組みとして知られている，カーネル関数を部分構造に基づき設計する畳み込みカーネルについて解説している。

第3章では，木構造をもったデータに対する，畳み込みカーネルの設計について論じている。まず，順序木に対するカーネル関数を，順序木に含まれる部分木を用いて定義し，これを効率的に計算するために，動的計画法に基づくアルゴリズムを提案している。さらに，部分木を数える際のあいまい性を許す拡張を行っている。また，順序木よりも一般的な木構造データに対しては，同様に定義されたカーネル関数を効率的に計算することができないという計算困難性の証明を与えている。

第4章では，グラフ構造をもったデータに対するカーネル関数の設計について論じている。最も一般的な部分構造として部分グラフを用いた場合のカーネル関数は効率的に計算できないことが知られているため，畳み込みカーネルのひとつである周辺化カーネルの枠組みにおいて，ランダムウォークによって生成されるパスを部分構造として用いることで，カーネルの計算を連立一次方程式に帰着し，効率的にカーネルの計算を行うアルゴリズムを提案している。化合物を用いた計算機実験では，他の代表的手法であるパターンマイニングに基づく手法と比較して遜色ない予測精度が得られることを示している。

第5章では，構造データのラベル付け問題に取り組んでいる。ラベル付け問題は，自然言語処理やバイオインフォマティクスで見られる問題で，分類問題の一般化として定式化されるが，第3章と第4章で提案した技法に基づいて，ラベル付け問題においてもカーネル関数を効率的に計算する方法を提案している。提案手法の性能の検証は，自然言語処理における固有表現抽出問題を用いて行われ，短い文脈のみに基づく手法と比較して，精度の改善が見られることを示している。

本論文の後半部分となる第6章では，外部構造，すなわちWWWや社会ネットワーク，生体ネットワークなどのネットワーク構造を扱う機械学習問題として，リンク予測問題を取り上げ，この問題に対する新しいアプローチを提案している。リンク予測問題は，ネットワーク構造の既知の部分を手がかりに，未知の構造を予測する問題であり，社会ネットワークや生体ネットワークの構造予測などの応用がある。ネットワーク上でのリンクの存在が「コピー＆ペースト」の機構によって，時間とともに移り変わっていくようなモデルを考え，このモデルの定常分布を用いて，ネットワーク構造の既知の部分にあてはめることによって，ネットワーク構造の未知の部分を推定する枠組みを提案している。また，最終的に得られる最適化問題において，EMに類似したアルゴリズムに基づく，効率的な解法を提案している。提案手法の予測性能は，実際の生体ネットワークデータにおいて検証され，既存の構造情報に基づく単純な予測手法と比較して，大きく性能が向上することが

示されている。

第6章は結論であり，本研究のまとめと今後の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，構造データ解析のための機械学習法について述べたもので，得られた成果は以下のとおりである。

⑴ RNA二次構造，糖鎖構造などの生物情報データや，HTMLデータなどは木構造を用いて表現できるため，木構造データに対する分類や予測は重要な問題であり，近年，カーネル法に基づく研究が数多く行われている。本論文では，畳み込みカーネルの枠組みに基づき，順序木データに対するカーネル関数を設計し，さらに，実際のHTMLデータに適用することにより有効性を示した。一方，無順序木に対しては同様に定義されたカーネル関数の計算が困難（＃P完全）であることを示した。

⑵ グラフ構造に対するカーネル関数の設計も薬剤設計などの応用があるため重要である。本論文では，ランダムウォークによって生成されるパスに基づいてグラフ構造に対するカーネル関数を定義し，その計算を連立一次方程式に帰着することにより効率的に計算する手法を提案した。さらに，実際の化学構造データに適用し有効性を示した。

⑶ 構造データのラベル付け問題は，自然言語処理やバイオインフォマティクスに応用を持つ重要な問題である。本論文では，上記で提案した技法に基づき，ラベル付け問題に対するカーネル関数を定義し，それを効率的に計算する手法を提案した。さらに，提案手法を自然言語処理の問題に適用し有効性を示した。

⑷ ネットワーク構造の解析は情報学，社会学，バイオインフォマティクスなど幅広い分野に応用を持つ重要な問題である。本論文では，ネットワーク構造の既知の部分を手がかりに，未知の部分を予測するというリンク構造予測問題に対し，ネットワーク上でのリンクが確率的な「コピー＆ペースト」により変化していくモデルを考え，このモデルの定常状態に既知の構造をあてはめることによりネットワーク構造を推定する新たな方法論を提案した。そして，実際の代謝ネットワークおよびタンパク質相互作用ネットワークのデータを用いて，提案手法との比較を行い，提案手法がこれらの既存手法の予測精度を大きく上回ることを示した。

以上，本論文は機械学習およびバイオインフォマティクスの両者において重要な研究テーマである構造データの解析に関して，新規で拡張性の高いカーネル関数を提案するとともに，ネットワークの部分構造予測に対する新たな方法論を提案している。さらに，いずれの手法についても実データを用いた計算機実験を通じてその有効性を示しており，当該分野の発展のために十分な寄与している。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，平成19年2月22日実施した論文内容とそれに関連した試問の結果合格と認めた。