論文の内容の要旨

論文題目 Matching and Learning in Trees
木の照合と学習

氏　　名 久保山 哲二

本論文は、木構造の近似パタン照合と分類学習に関する理論的な基礎づけと新事実の発見、および、木構造の新たな分類学習アルゴリズムの開発とその情報生物学への応用に関する成果についてまとめたものである。

第1章「Introduction(序論)」では、本論文の研究の背景と動機について述べる。木構造(以後、単に木と呼ぶ)は、情報の階層構造を自然に表現し、効率的に処理するために広く用いられている抽象構造である。インターネット上に蓄積されている膨大なHTML・XML文書や、自然言語の構文木、生物学分野における系統樹、RNAの二次構造、糖鎖構造、画像処理分野における対象物のグラフモデルなどをはじめとするさまざまな対象が木として表現可能である。これらの対象の効率的な検索や、対象からの知識発見のために、木に関する多くのアルゴリズムが提案されている。なかでも、編集距離の概念に基づく木の近似照合アルゴリズムは、2つの木の間の距離や類似度を計算するだけでなく、共通パタンの発見、構造の統合等の用途に対しても適用可能な一般性をもった手法であり、適用範囲の広さからも重要性が高い。編集距離は、2つの構造を比較する一般的なフレームワークであり、一方の構造から編集操作を用いて、もう一方の構造へ変換するために要する最小の編集コストにより定義される。しかし、木の編集距離の分野では、同じアルゴリズムが気づかれることなく独立に何度も発表されたり、提案者が意図する近似照合の意味と実際に提案されているアルゴリズムとの間にギャップがあったりするなどの混乱が生じている。これは、近似照合アルゴリズムを厳密に記述・比較するための理論的なフレームワークがなかったためである。

一方、近年のサポートベクターマシン(SVM)に代表される学習器を用いたカーネル法による機械学習の爆発的な流行にともない、木の分類学習のためのさまざまな木の類似度(カーネル)が提案されている。しかし、この分野でも、提案者が意図する類似度の意味と実際に提案されている類似度の間にギャップがある場合や、厳密にカーネルとしての要件

を満たしていることが証明されていない場合が散見される。また、カーネルとして用いることが出来る計算効率のよい一般的な木の類似度計算のアルゴリズムも提案されていない。このような背景を踏まえて、本論文では木の近似照合分野における混乱を解消するための基礎理論構築と、その機械学習への応用に関する新たな研究成果を示す。

第1部の「Matching in Trees (木の照合)」は、第2章から第4章までの3つの章から成る。**第2章「Approximate Tree Matching(木の近似照合)」**では、代表的な既存研究を厳密に定式化しながら紹介する。まず木の近似照合の基礎となる文字列の近似照合について述べる。文字列の近似照合は、編集距離、アラインメント、トレースといった様々な観点から定式化可能であり、数学的に等価であることが知られている。これらの定義は、近似照合の意味を構成的手続きにより示す操作的な定義と、集合論的な概念により示す宣言的な定義に分類できる。次に、同様の観点から、木の編集距離や木のアラインメントなどの既存の多様な木の近似照合手法を概観する。

第3章「Theoretical Foundation of Approximate Tree Matching (木の近似照合の理論的基礎)」では、広い分野で独立に提案されてきた多様な木の近似照合を統一的に記述するための基礎理論を半順序代数を用いて構築する。この基礎付けにより、近似照合の操作的な定義と宣言的な定義を、数学的に厳密に橋渡しすることが可能となる。実際に、編集距離における木の編集操作に、2つの木構造間の写像としての厳密な意味を与える。

第4章「Relationship Analysis among Tree Edit Distance Measures (木の編集距離尺度間の関連性の解析)」では、第3章で構築した基礎理論を用いた解析により、10年来、知られていなかった重要な木の近似照合手法である「木のアラインメント」の宣言的な定義を示す。この結果は、2つの木を1つに結合するための必要十分条件になっており、広い応用が考えられる。さらに、従来別々のアルゴリズムであると考えられていた「木のアラインメント」と「less-constrained 編集距離」が、実は等価なアルゴリズムであることを示す。この解析の過程で、「less-constrained 編集距離」の提案論文の宣言的な定義が、実は「constrained 編集距離」の意味と等価であり、提案者の意図を正しくあらわしていないことを示す。その他にも、厳密に等価性が示されていなかったいくつかの近似照合アルゴリズムに対して、等価性を示した。また、さまざまな木の近似照合アルゴリズムをクラス分けし、これらのクラス間の階層関係を示す。

第2部の「Learning in Trees (木の学習)」は、第5章から第8章までの4つの章から成る。第5章「Kernel-based Learning for Trees (カーネルに基づく木の学習)」では、木の学習問題を、カーネル法におけるカーネル設計の問題として捉え、離散構造のカーネルの重要な設計指針として広く用いられている畳み込みカーネルについて述べる。また、畳み込みカーネルの概念に基づき提案された代表的な木のカーネルについて概観する。また、畳み込みカーネルとして提案されている既存の最も一般的な木カーネルが、実際には畳み込みカーネルのクラスではなく、カーネルとして用いるためには理論的な裏づけが必要であることを示す。

第6章「Mapping Kernel for Trees (木のマッピングカーネル)」では、既存の木カーネルを、第4章で明らかにした木の近似照合のクラス階層の観点から統一的に定式化し、既存の木カーネルの意味づけを行う。また、従来研究では理論的な裏づけなくカーネルとして用いられてきた木の類似度が、実際にカーネル法による学習問題に適用できる数学的条件を満たすことを厳密に証明する。さらに、木の近似照合の各々のクラスに対応す

る2つの木の類似度(木カーネル)を新たに提案し、最も表現力の高かった既存の木カーネルよりも、さらに高い表現力を持つことを示す。これに対して、木のアラインメントに対応する木カーネルが存在しないことを厳密に示す。

第7章「Spectrum Kernel for Trees (木のスペクトラムカーネル)」では、高速でかつ一般性の高い木カーネルとして木のスペクトラムカーネルを提案する。木のスペクトラムカーネルでは、新たな概念として、文字列のqグラムの概念を木に拡張した木のqグラム を提案し、2つの木に共通して含まれるqグラムを数え上げることにより類似性を測る高速なアルゴリズムを提案する。さらに、木のスペクトラムカーネルを拡張したグラム分布カーネルを提案する。

第8章「Application to Glycan Classification (糖鎖の分類への応用)」では、前章で開発した2つの木カーネルを、バイオインフォマティクスの分野における糖鎖構造の分類学習、および、糖鎖構造からのモチーフ抽出に実際に適用する。その結果、従来の一般的な木カーネルや、糖鎖専用のカーネルと比べて、分類能力と計算効率の両面において、高い性能を示すことを示す。また、モチーフ抽出では、糖鎖専用のカーネルよりも一般的な構造をもつモチーフを抽出できることを示す。

最後に、**第9章「Conclusion and Future Work (結論と今後の課題)」**で、本論文の成果を要約し、関連する未解決問題や今後の発展について述べる。