称号及び氏名	博士（理学）寺田　幹治
学位授与の日付	平成 17 年 2 月 28 日
論　文　名	「パターン上の決定木の帰納推論」
論文審査委員	主査　佐藤 優子
	副査　石井 伸郎
	副査　馬野 元秀
	副査　寺岡 義伸
	副査　寺岡 義博
	副査　向内 康人

論文要旨

今日，科学の各分野で，得られる観測・実験のデータは膨大なものがあり，これらの分析・処理のためには計算機による解析が不可欠になっている．そこでは，膨大な観測・実験データから計算機による知識化が求められることになる．本学位論文では，観測・実験データから知識化を行なう際の枠組みとして，帰納推論として知られる学習モデルを用いる．

帰納推論とは，与えられたデータからそれらを説明する一般的な規則を推測する過程のことである．E.M. Gold (1967)は「極限における同定」と呼ばれる帰納推論の成功基準を導入し，形式言語と帰納的関数の帰納推論理論の枠組みを与えた．帰納推論の過程が，ある概念を極限同定するとは，その概念に関する例をどのように与えても，すべての可能な例がいつか現れるような与え方をする限りは，出力される推論の列が正しい目標概念に収束することをいう．本論文では，帰納推論の成功基準として「極限における同定」を採用し，極限において同定可能であることを，単に推論可能ということにする．

第 2 章では，この基準に基づいた形式言語の帰納推論とパターン言語について既知の概念と性質についてまとめる．形式言語の正例からの帰納推論では，与えられる事例は未知の言語に含まれるすべての語である．正例からの帰納推論に関して，E.M. Gold (1967)は超有限な言語族(すべての有限言語と少なくとも一つの無限言語を含む言語族)が推論可能とはならないことを示した．このことは，Chomsky の階層のもっとも下位に位置する正規言

語族さえも正例からは推論可能ではないことを意味する．それ故にしばらくの間，正例からの帰納推論は注目されることはなかったが，D. Angluin (1980)は有限証拠集合という概念を用いて，言語族が正例から推論可能であることを特徴づける定理を示した．それと同時に，有限の厚さ(finite thickness)とよばれる，推論可能であるための十分条件を与え，パターン言語とよばれる言語の族が正例から推論可能であることを示した．パターンとは，アルファベット∑に含まれる定数記号と変数記号からなる文字列であり，特に，各変数記号が一度しか現れないようなパターンを正規パターンと呼ぶ．パターン言語とは，パターンに現れるすべての変数記号に空でない文字列を代入して得られる∑上の文字列からなる集合である．

一方，K. Wright (1989)は，有限の厚さを一般化した有限の弾力性(finite elasticity)と呼ばれる概念を導入し，この性質が正例からの推論可能性であるための十分条件であること，および，この性質が言語の和で構成される言語族に関しても保持されることを示した．この結果から，予め定められた定数個以下のパターン言語の和からなる言語族が正例から推論可能であることが示される．また，T. Moriyama and M. Sato (1994)は，この性質が言語の積(共通部分)や連接などの種々の演算を施した言語族に関して保持されることを示した．

第 3 章では，これらの帰納推論の理論を背景に，パターン上の決定木の学習問題を扱う．S. Arikawa 等 (1993)や S. Miyano (1993)は，ゲノム情報のプロジェクトの中でアミノ酸残基における膜貫通領域の同定する問題を，計算学習のパラダイムの一つである「PAC 学習」の枠組みを用いて定式化を行い，その理論的研究から学習システムの開発および計算機実験まで幅広い研究を行っている．ここで，このモチーフを表現するのに採用されたモデルは，正規パターン上の決定木であり，学習システムに提示される事例は，コンピュータの切り出す膜貫通領域の正および負の事例であった．本論文では，パターン上の決定木が定義する言語族の学習問題を，正例からの「極限における同定」の枠組みで扱っている．

パターン上の決定木とは，葉以外の節のラベルがパターンで，葉のラベルが 0 または 1 の 2 分木である．∑上の文字列 w が与えられると，パターンをラベルとする節では w がそのパターンにマッチするか否か，すなわち，そのパターン言語に含まれるか否かに応じて左または右に分岐し，根から葉に至る．葉のラベルが 1 ならば文字列 w は受理され，0 ならば受理されない．パターン上の決定木 T が定義する言語は，T で受理される文字列全体の集合として定義される．本論文では，高さが 1 のパターン上の決定木 T で定義される言語族が正例から推論可能となるが，高さが 2 以上のものまで含めると，正例から推論可能とはならないことを示した．

パターン上の決定木で定義される言語は，パターン言語やそれらの補言語に高々有限回

の積演算や和演算を施して得られる言語として表現される．パターン言語族およびパターン言語の補言語からなる言語族の両方が有限の弾力性をもてば，演算の閉包性に関する結果(K. Wright (1989)および T. Moriyama and M. Sato (1994))を応用して，高さがあらかじめ定められた定数以下のパターン上の決定木で定義される言語族は，有限の弾力性をもち，正例から推論可能ということになる．しかし，パターン言語族は有限の弾力性をもつが，パターン言語の補言語からなる言語族は有限の弾力性をもたない．そこで，各パターン p に対して，コ・パターンと呼ばれる文字列 p^c を導入し，その言語をパターン p の補言語のある部分集合として定義する．その定義の下で，コ・パターン言語族が有限の弾力性を持つことを示した．さらに，パターン上の決定木で定義される言語を，パターンとコ・パターンが定義する言語に高々有限回の積演算と和演算を施して得られる言語として新たな定義を行い，その解釈のもとに，高さがあらかじめ定められた定数以下のパターン上の決定木で定義される言語族が正例から推論可能となることを示した．

さらに，正規パターン上の決定木が定義する言語族の効率的な推論問題を扱っている．正規パターンで定義される言語族は，正例から効率的に推論可能であることが知られている．まず，高さが 1 の正規パターン上の決定木で定義される言語族を正例から効率的に推論するアルゴリズムを示した．高さ 2 以上の正規パターン上の決定木を効率的に推論するアルゴリズムは得られていない．第 4 章では，効率的なアルゴリズムを構築する際に重要な鍵となる，正規パターンとコ・正規パターンで定義される言語の積集合や和集合の包含関係について，意味的包含関係と同値となる構文的な関係を示している．

最後に，第 5 章では，形式言語の Prefix-Free とよばれる生成集合の学習問題を考える．アルファベット∑上の言語 L に対して，$L \subseteq G^+$をみたす集合 G を L の生成集合といい，L ？ G で表わす．アルファベット∑は空語を含まない任意の言語 L の生成集合であり，言語 L 自身は L の生成集合である．文字列の集合 G は G の任意の文字列が G の文字列の真の prefix とならないとき，Prefix-Free であるという．一般に，与えられた言語 L の Prefix-Free 生成集合は複数個ある．Prefix-Free 集合 G は，G^+の任意の文字列が G の要素で一意的に分解できるという性質をもつので，いわゆる符号解読に用いることができる．符号理論においては，このような集合をコードとよび，符号解読ではこれらの結びつきを議論する(R.M. Capocelli (1982)).

本論文では，まず，言語 L のすべての既約な Prefix-Free 生成集合の族 PFG_L は，関係 ？のもとで完備束を構成し，最小の要素 $G_L{}^{\inf}$ と最大の要素 $G_L{}^{\sup}$ が存在することを示した．特に，$G_L{}^{\inf}$ は L におけるすべての文字列が PFG_L のなかの最も短い長さを持つという意味で興味を引く性質を持っている．次に，有限言語 L に対して，$G_L{}^{\inf}$ を求めるための効率的なアルゴリズムを提示している．有限言語とは限らない L については，L の正提示から $G_L{}^{\inf}$

を極限同定する学習問題を考え，最小の生成集合 $G_L{}^{\inf}$ が有限となるとき，$G_L{}^{\inf}$ を求める効率的な学習アルゴリズムを与えている．

審査結果の要旨

この学位論文では，「極限同定」に基づく帰納推論の枠組みを用いて，正の事例からパターン上の決定木を極限同定する問題を扱っている．E.M. Gold(1967)による極限同定の成功基準は，今日まで多数の研究が行なわれている機械学習の主要なパラダイムの一つである．

本論文では，まず，通常のパターン上の決定木で定義される言語族は，高さが 1 の場合は正例から帰納推論可能となるが，高さが 2 以上の場合まで含めると，正例から帰納推論可能とはならないという結果を得ている．そこで，各パターン p に対して，コ・パターンと呼ばれる文字列 p^c を導入し，決定木で定義される言語を，パターンとコ・パターンが定義する言語に高々有限回の積(共通部分)演算と和演算を施して得られる言語として新たな定義を行い，その解釈のもとに，正の事例からの帰納推論可能性を調べている．

本論文の主要な結果として，パターンとコ・パターンで定義される言語からなる言語族，および，それらにあらかじめ定められた定数回以下の積演算と和演算を施して得られる言語族が正例から帰納推論可能であるという重要な結果を得ている．これは，高さがあらかじめ定められた定数以下のパターン上の決定木で定義される言語族が正例から帰納推論可能であることを意味する．

さらに，効率的な帰納推論を考えるため，正規パターンに制限した決定木の帰納推論問題を扱っている．本論文の主要な結果として，高さが 1 の正規パターン上の決定木で定義される言語族が正例から効率的に帰納推論可能となることを示している．高さ 2 以上の正規パターン上の決定木を効率的に推論するアルゴリズムはまだ得られていないが，その構築の重要な鍵となる，正規パターンとコ・正規パターンで定義される言語の積集合や和集合の包含関係について，意味的包含関係と同値となる構文的な関係を示している．

また，本論文では，形式言語の Prefix-Free 生成集合の概念を導入し，ある言語の正例が与えられたとき，その Prefix-Free 生成集合を極限同定するという問題についても扱っている．まず，既約な Prefix-Free 生成集合の族が完備束となることを示し，有限言語 L に対して，その最小元 $G_L{}^{\inf}$ を効率的に求めるアルゴリズムを与えている．このアルゴリズムを用いて，$G_L{}^{\inf}$ が有限となる場合，有限とは限らない言語 L の正例から $G_L{}^{\inf}$ を効率的に極限同定可能であることを示している．

以上のとおり，パターン上の決定木が正例から学習可能であること，特に，高さ 1 の正規パターン上の決定木が効率的に学習可能となるなどの肯定的な結果を得ており，ゲノム情報科学などへの応用も期待できる．学位論文審査委員会は，学位論文の審査ならびに最終試験の結果から，申請者に対して博士(理学)の学位を授与することを適当と認める．