（ 続紙 １ ）

京都大学	博士（　　情報　　学）	氏名	NGUYEN VIET ANH
論文題目	Studies on Efficient Methods to Learn from Structured Data (構造データを対象とした学習のための効率的手法に関する研究)		

（論文内容の要旨）

本論文は，木構造データやグラフデータなどの構造化データからの機械学習に対して，効率的な学習アルゴリズムを提案するものである．構造化データは，HTML/XMLなどのデータ交換，自然言語処理，生命情報学における化学構造データの処理など，多くの分野で利用され，蓄積されている．機械学習手法は，蓄積されたデータから有用な知識を抽出することを目的として利用が進められている．構造化データからの機械学習アルゴリズムを設計する際には，有用な結果を効率的に得るために，構造化データが備える性質を応用する．本論文では，機械学習手法が利用される状況を考慮したいくつかの設定を与えた上で，学習アルゴリズムを設計している．

第１章は序章であり，機械学習アルゴリズムの設計に必要な構造化データの基本的性質について述べ，続いて関連研究を提示して本論文の位置付けを与えている．

第２章では，根付き非順序木からなるデータベースに頻出する部分木を効率的に列挙するという機械学習の問題に対して，利用者がその意図を学習結果に反映するために制約木を与えるという条件と，利用者が学習結果を把握しやすくするために出力を閉部分木に限定するという条件を設定し，両条件を同時に満たすアルゴリズムSCCMを提案している．SCCMは，データベース中の各木における制約木の出現位置を探索し，その結果をもとにアイテム集合のデータベースを構成した上で，既存の機械学習手法を適用するというものである．さらに，SCCMが効率的に学習することを，人工データと公開されている実データを利用した計算機実験により検証している．

第３章では，データベースに次々と新たな根付き順序木が挿入されるという設定において，各時点において頻出する閉部分木の集合を，直前の時点における学習結果を利用することで効率的に求めるアルゴリズムICTreeMinerを提案している．この設定は，データベースが更新されることはその実用において必然であるという事実を反映している．ICTreeMinerは分割統治法に基づいて設計されており，直前時点で得られている各閉部分木を，データベースへの新たな木の挿入によって頻度が変化しないものと変化するものに分類した上で，後者からなる集合をもとに３種類の集合を計算し，これらの和集合を構成することにより求める閉部分木の集合を計算する．

第４章では，無向グラフを蓄積したデータベースからの機械学習に適用することを目的として，形式概念解析を用いて無向グラフの集合からなる概念束を構成するための基礎理論を与えている．形式概念解析では，対象の属性を利用して，対象の集合からなる完備束として概念束を構成する．本論文では，無向グラフの属性としてそのグラフに出現する部分グラフを採用し，さらに，完備束を計算することは非効率的であることから，属性となる部分グラフ間の包摂関係を利用することにより，不完全ながらも実用性があり，かつ効率的に計算可能な概念束の構成方法を与えている．

第５章では，前章で構成法を与えた概念束を利用して，２つの無向グラフ間の類似性を提案している．この類似性は，２つの無向グラフ間の包摂関係によって定義される類似性ではノイズの扱いが困難であり，かつ効率的な計算が困難なこともある，という問題を解決することを目的にしている．提案している類似性をグラフのクラス分類アルゴリズムに適用することで，その機械学習における有用性を検証している．

第６章では，本論文で与えた結果を概観し，結論としている．

（続紙 ２ ）

（論文審査の結果の要旨）

本論文は，木構造データやグラフデータなどの構造化データから有用な知識を得るための効率的機械学習アルゴリズムを構成することを目的としている．現在，構造化データは様々な分野で蓄積と利用が進められており，機械学習の適用も進展している．機械学習手法は，ベクトルやタプルなどの単純な構造を持つデータを対象とすることを中心に研究と開発が進められてきたため，それをそのまま構造化データに適用すると，学習が効率的でなかったり，学習結果の利用が困難となる状況がしばしば生じる．本論文はこれらの問題点を解決するため，機械学習が利用される状況を考慮した上で，効率的な学習アルゴリズムを与えている．主要な結果は以下の３つである．

１．根付き非順序木からなるデータベースを対象とした機械学習において，そこに頻出する部分木を効率的に列挙するという基本的な問題に対して，出力を，制約木とよばれる木を含み，かつ閉となる部分木に限定した学習アルゴリズムSCCMを構築している．第一の条件は利用者の意図を学習結果に反映することを，そして第二の条件は利用者による結果の把握を容易にすることを目的にしている．SCCMによる学習が効率的であることは，人工データと実データを利用した計算機実験により検証している．

２．実用においては，データベースは更新されることが前提であるという事実を考慮するため，根付き順序木を蓄積したデータベースに次々と新たな木が挿入されるという更新の状況を設定した上で，各時点において頻出する閉部分木の集合を効率的に求めるアルゴリズムICTreeMinerを設計している．ICTreeMinerは，直前の時点で得られた集合に対して，分割統治法を適用することにより，効率に学習結果を構成することを可能としている．

３．無向グラフを蓄積したデータベースを対象とした機械学習において，形式概念解析と呼ばれる代数的データ解析手法を適用するための基礎理論を与えた上で，そこで得られる概念束とよばれる束を利用して，無向グラフ間の新たな類似性を定義している．この理論においては，無向グラフの中に出現する部分グラフをもとのグラフの属性とするため，属性間にも順序関係が定義され，それを用いることによって不完全であるものの有用であり，かつ効率的に計算可能な概念束の構成を可能としている．定義された類似性の有用性は，グラフのクラス分類アルゴリズムに応用することにより検証している．

これらの研究成果はどれも全く新規な着想に基づくものである．アルゴリズムの設計や理論の構成においては，構造化データの持つ性質を効果的に利用しており，その独創性は特筆すべきである．さらに，それらの有効性を実データを含むデータを対象に計算機実験によって検証しており，機械学習分野に対する貢献度は高い．よって，本論文は博士（情報学）の学位論文として価値のあるものと認める．

また，平成２４年７月１８日に実施した論文とそれに関連する内容についての口頭試問の結果，合格と認めた．