

(続紙 1)

京都大学	博士 (情報学)	氏名	彭 思棋 (PENG SIQI)
論文題目	Studies on Several Methods for Integrating Formal Concept Analysis and Statistical Machine Learning (形式概念分析と統計的機械学習を統合するためのいくつかの手法に関する研究)		
(論文内容の要旨)			
<p>形式概念分析(以下ではFCAと表す)は、2項関係を表すデータから形式概念とよばれる閉包を列挙し、形式概念間の半順序関係に基づく概念束を構築することにより知識を獲得する手法である。本論文は、FCAを統計的機械学習と統合することにより、2項関係データからの予測問題を解くことを目的とした一連の手法の開発について研究結果をまとめたものであり、全6章から構成されている。</p> <p>第1章は序論であり、本論文の目的とその内容を概観している。FCAと、その拡張である3項関係解析(TCA)についての基本的な定義を与えている。FCAでは、対象の有限集合と属性の有限集合との間の2項関係を表すデータから形式概念を列挙することにより分析を行う。TCAは対象、属性に加えて条件の有限集合が与えられたときに、それら3項間の関係を表すデータから3項概念とよばれる閉包を列挙することにより分析する。</p> <p>第2章では、TCAを効率的に行う手法として、ゼロサプレス型二分決定グラフ(ZDD)を利用したZ-TCAを提案している。FCAに対しては、形式概念間の半順序関係の性質に基づいた自然で効率的な形式概念の列挙アルゴリズムが設計可能である一方、TCAにおける3項概念間の半順序関係は同様の性質を持たないため、3項概念の効率的な列挙アルゴリズムは提案されていなかった。本論文では、形式概念の列挙アルゴリズムを2層の入れ子構造にし、集合の表現にZDDを用いれば、実用上十分高速に3項概念を列挙するアルゴリズムZ-TCAを構成できることを示している。また、映画に関する実データを用いてZ-TCAの有効性を確認し、人工データを用いてアルゴリズムが有効に機能する条件を導いている。</p> <p>第3章では、FCAによって得られる概念束を簡約する手法を提案している。FCAを適用する原データが大規模になると、得られる概念束も大規模かつ複雑になり、概念束から有用な知識を獲得することが困難になるため、概念束を簡約する必要性が生じる。本論文では、整数計画法を利用することにより、原データの改変をできるだけ小さくし、かつ新たな概念の追加や、得られている概念の削除をできるだけ少なくしながら、概念束の簡約が可能な手法を提案している。この手法について、生命科学分野のデータを用いて有効性を確認している。</p> <p>第4章では、二部グラフのリンク予測問題に対して、FCAと行列分解を組合せた機械学習手法を提案している。FCAが対象とする2項関係データは、対象と概念を節点とする二部グラフを表していると解釈することができる。本論文では、この二部グラフにおいて観測されていないリンクを予測する、という問題に対して、原データにFCAを適用して得られた形式概念を用いて統計的機械学習に必要な負例を生成し、行列分解を用いた協調フィルタリングを利用したアルゴリズムを提案している。さらに提案手法を生命科学分野の実データに適用することにより、行列分解を利用するがFCAを利用しなかった既存手法に比べて高い予測性能を持つことを確認している。</p>			

第5章では、FCAで得られた概念束と深層学習のアーキテクチャの一つであるBERTを組み合わせることにより、二部グラフのリンク予測問題を解く手法を提案している。BERTを用いれば自然言語で記述された文における単語間の参照関係を予測できることに着目し、対象あるいは属性を単語とみなし、概念を文、概念束における最大概念から最小概念への経路を文の系列とみなした上で、二部グラフのリンクを単語の参照関係とみなしてリンク予測問題にBERTを適用する手法を提案している。この手法についても生命科学分野の実データに適用することにより、既存手法よりも優れた予測性能を持つことを確認している。

第6章は結論であり、本論文で得られた成果をFCA4SMLという呼称でまとめた上で、将来への展望を述べている。

(続紙 2)

(論文審査の結果の要旨)

本論文は、2項関係を表す関係データからの予測問題を解くことを目標として、形式概念分析(FCA)を拡張する方法と、FCAと統計的機械学習を統合する手法を開発することを目的としている。FCAにおいては、対象と属性の間の2項関係を表すデータから、形式概念とよばれる閉包を列挙し、得られた形式概念がなす概念束と呼ばれる束構造から知識を得る。主な結果は以下の通りである。

1. FCAを拡張し、対象と属性と条件の三者間の関係を表す3項関係データを分析し、3項概念と呼ばれる閉包を列挙して知識を獲得する手法であるTCAに対して、ゼロサプレス型二分決定グラフ(ZDD)を用いた高速なアルゴリズムZ-TCAを提案している。TCAに対して3項概念を効率的に列挙するアルゴリズムは提案されていなかったが、形式概念の列挙アルゴリズムを2層の入れ子構造にした上で、ZDDを用いて集合演算を高速化したZ-TCAが実用上十分高速に3項概念を列挙することを示している。
2. FCAで得られる概念束が原データの大規模化にしたがって大規模かつ複雑化し、そこから知識を獲得することが困難になるという問題を解決するために、整数計画法を利用して概念束を簡約する手法を提案している。原データの改変をできるだけ小さくし、新たな概念を追加や、得られている概念の削除をできるだけ少なくしながら、概念束を簡約することを可能にしている。
3. 二部グラフにおいて観測されていないリンクを予測するリンク予測問題に対して、FCAが対象とする2項関係データが二部グラフを表していると解釈し、FCAと行列分解を組合せたアルゴリズムを提案している。二部グラフにFCAを適用して得られた形式概念を用いて機械学習に必要な負例を生成し、行列分解と協調フィルタリングを利用することにより実現している。
4. FCAと自然言語処理を対応させ、深層学習のアーキテクチャの一つであるBERTを用いて二部グラフのリンク予測問題を解く手法を提案している。FCAにおける対象あるいは属性を自然言語で記述された文の単語とみなし、FCAで得られた概念を文、概念束における経路を文の系列とみなした上でBERTを適用する、という手法である。

これらの研究成果はどれも新規な着想に基づくものである。アルゴリズムや手法の設計、理論の構成においては、FCA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と人工データを対象に計算機実験によって検証しており、知識発見と機械学習の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あるものと認める。また、令和6年8月21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。さらに、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。