氏　　名	永島　寛子
学位の種類	博士（理学）
学位記番号	甲第理 8 号
学位授与年月日	2021（令和 3）年 3 月 17 日
学位授与の要件	東京女子大学学位規程第 3 条第 3 項第 1 号
学位論文題目	Study on Mathematical Framework for Data Analysis Based on Machine Learning 機械学習に基づくデータ分析のための数理基盤と方法論に関する研究
論文審査委員	主査　教　授　加藤　由花 副査　教　授　大阿久　俊則 副査　教　授　荻田　武史 副査　千葉商科大学商経学部教授 橋本　隆子

内容の要旨および審査の結果の要旨

Ⅰ．論文内容の要旨

近年、センサーやウェラブル端末から収集されるデータなど、分析に利用されるデータの量と種類が急増し、顧客動向分析、工場の生産管理、ロボットの自律移動など様々な分野でデータの利活用が活発に行われている。これらデータ分析の精度を高めるためには、外れ値や欠損値への対応、測定機器ごとに異なる単位の統一など、いわゆるデータの『前処理』が不可欠であり、実際、このような前処理には、データ分析のリソースの 80%以上が費やされているという報告がある。そのため、これまでも、ソフトウェア・ツールを用いた手動による方法や、機械学習アルゴリズムを用いた自動化された方法など、様々なデータ前処理手法が提案されてきた。しかし、これら既存の手法には、以下の二つの課題が存在する。一つは、単一の手法では全ての種類の欠損データに対応できないという点、もう一つは、機械学習を用いた手法は専門知識が必要であるため、環境に応じたカスタマイズが困難であるという点である。

本論文では、これらの課題の解決を目的に、前処理の半自動化を実現するためのフレームワーク APREP-DM（automated pre-processing for data mining）、およびフ

レームワークでの利用を前提としたデータ補完手法 APREP-S（automated pre-processing for sensor data）を提案している。APREP-DM は、分析の目的や基準を定義するビジネス理解（business understanding）のステップをフレームワークに内包することに特徴があり、これにより前処理の半自動化を実現する。APREP-S は、事前に定義しておいた複数のデータ補完手法（統計値、時系列解析、機械学習アルゴリズムなど）から、補完箇所の特徴に合った最適な手法を選択することにより、様々な種類の欠損データに対応した補完を実現する。適切な手法を選択するために、APREP-S では、論文内で新たに定義した確率モデルを用いて、候補となる補完手法をランク付けする。確率モデルのパラメータは、与えられたデータから最尤推定により学習するが、本論文ではこの学習過程を、ベイズ推論と PBE（programming by example、例示によるプログラムの自動生成）の概念に基づいて設計している。そのため、学習とランク付けのプロセスは逐次的・対話的に行われることになり、この繰り返しの過程で対象環境に対する人間の知識を確率モデルに組み込んでいる。

本論文では、APREP-DM の有効性を、シナリオベースの評価および定性評価により検証している。その結果、ビジネス理解のステップをフレームワークに組み込むことで前処理の半自動化が可能になること、APREP-DM が他の手法と比較しセンサーデータ解析においてバランスの取れたフレームワークであることが示されている。APREP-S については、補完対象データの周期、学習データの種類、更新処理、モデルで使用する特徴の 4 つの観点から、データセットを用いた実験を行っている。その結果、データの種類によらず、対象となる特徴量に応じて適切な補完手法が選択されること、学習・ランク付けの処理を繰り返すごとに推論の精度が向上することが示されている。

以上により、提案したデータマイニングフレームワークとデータ補完手法は、精度よく、分析プロセスに必要なリソースを削減できる手法であると結論づけている。

Ⅱ．審査の結果の要旨

1．論文の構成

本論文は、6 つの章で構成されている。

第 1 章では、研究の背景を説明した上で、研究の目的とそれを達成する際の課題、解決策が述べられている。

第 2 章では、トピックごとに関連研究をまとめることで、本研究の位置付けを明確にしている。

第 3 章では、本論文で用いる主な確率モデルとして、多クラスベイズロジスティック回帰、隠れマルコフモデル、k-Shape について説明がなされている。さらに、提案手法のベースとなる理論として、PBE（Programing by Example）の概念が紹介されている。

第 4 章では、前処理の半自動化を実現するためのフレームワーク APREP-DM の詳細が説明されている。また、シナリオベースの機能検証および既存手法との定性的な特

徴比較により、フレームワークの有効性を検証した結果が示されている。

第 5 章では、フレームワークでの利用を前提としたデータ補完手法 APREP-S の提案を行っている。さらに、気象データおよびウェラブルセンサーデータという 2 種類のデータセットを用いた評価実験が行われており、その結果が示されている。

第 6 章では、本論文の結論と今後の課題が述べられている。

2. 論文の特徴

本論文は、データ分析における前処理の半自動化を研究対象としており、既に査読付き専門論文誌に掲載された 2 本の論文、および査読付き国際会議論文誌に掲載された 1 本の論文の研究成果をもとに、内容を再構成してまとめたものである。

現在、深層学習に代表される様々な機械学習手法がコモディティ化してきており、様々なツールを用いることにより、比較的簡単に様々なデータ分析が行えるようになってきている。データ分析のかなりの部分が自動化されていると言えるが、機械学習アルゴリズムで利用するための入力データを、ドメインごとの知識を取り入れながらどのように構築していくかという部分については、ほとんど検討がなされていないのが実情である。例えば、工場内でセンサーデータを収集して、リアルタイムで異常検知を行うシステムを考えた場合、異常検知ツールにデータを入力するためには、フロア構成等、工場ごとの環境を考慮してツールに入力するデータを前処理により作成する必要がある。本論文は、この『データ前処理』に着目し、機械学習ツールに入力するためのデータセットを半自動で生成するための枠組みを提案したものである。

提案手法は、環境や条件ごとに人間がいくつかの前処理の例を手動でシステムに教示することにより、確率モデルとして定義された前処理プログラムのパラメータが学習され、利用環境に適したプログラムへと半自動的にチューニングが行われる点に特徴がある。つまり、本論文では、「機械学習ツールに入力するデータセットを生成するための」機械学習モデルが提案されている。この学習を実現するための数理基盤と、人間が教示を行うための仕組みを合わせてフレームワークを構築するという点が論文の骨子となっている。

3. 論文の評価

データからの仮説発見を目的とするデータ科学は、現在、飛躍的に進展している重要な研究分野である。特に、その一分野である機械学習・深層学習の発展は目覚ましく、近年、機械学習を利用したシステムは急速に社会に浸透しつつある。一方、機械学習を組み込んだシステムを実現するためには、ドメイン知識（業務や環境に依存したノウハウ等）を取り入れたシステムの改変・調整が必須であるが、この方法論は未だ確立されておらず、システム構築時に試行錯誤が繰り返されることが問題になっている。本論文で提案しているフレームワークは、データ前処理を半自動化することにより、この試行錯誤を大幅に減少することが期待でき、インパクトのあるすばらしい成果である。

データ前処理については、多くの研究が機械学習による処理の完全自動化を目指しているのに対し、本論文では手動と自動のハイブリッド方式（半自動化方式）を提案している。これは、手動方式を組み合わせることにより、ドメイン知識を前処理プログラムに組み込むことを指向しているためである。手法の実現方法としては、自動プログラミングの分野での研究成果である例示プログラミング（PBE）の概念を採用し、人間がいくつかの前処理例を教示することにより、プログラムのパラメータが自動的に更新されていく仕組みを、フレームワークとともに提案している。これは、機械学習システムの構築に、人間参加型 AI とも呼ばれる Human-in-the-loop の概念を取り入れた新たな試みであり、学術的に評価に値する。

前処理プログラムは、具体的な確率モデルとして定義されており、事前に定義しておくデータ補完手法を入れ替えることで、幅広い環境に適用可能である。また、人間が教示を行うための仕組みを合わせて提供することにより、学習と教示のプロセスを逐次的・対話的に行うことを可能にしている。その結果、十分な学習データがなくても逐次的に前処理の精度を上げていくことに成功している点は大きな成果である。

以上のように、本論文で提案しているデータ前処理のためのフレームワークは、データ分析の分野に新たな概念を取り入れた学術的に評価すべき研究であるとともに、幅広い分野への応用・適用が期待でき、応用数理学分野において高い価値を持つ良い成果である。

４．最終試験の概要

最終試験として、オンライン形式で公開の論文発表会を行った。研究の背景から始めて、提案手法のベースとなる基本的事項を示しながら、その数理的な側面を解説した。その後、提案するフレームワーク、データ補完手法について、有効性を検証するための評価実験の結果を交えながら、詳細に説明した。発表内容は論理的に整理され、十分な準備をされたものであった。トピックは多岐に渡っていたが、それらが適切に構造化され、聴衆の理解度を意識しながら説明が行われており、良い発表であったといえる。

外国語の試験においても、論文の内容が適切にまとめられ、丁寧に解説がなされた。

以上のように、論文の内容、最終試験の結果ともに、本学博士（理学）に十分値すると判断できる。