機械学習を用いたデータ分析と知識習得を同時に支援するソフトウェアの研究

鴨志田 亮太

電気通信大学大学院 情報理工学研究科

博士（工学）の学位申請論文

2022 年 3 月

機械学習を用いたデータ分析と知識習得を同時に支援するソフトウェアの研究

博士論文審査委員会

主査　大須賀 昭彦　教授
委員　石川 冬樹　客員准教授
委員　田中 健次　教授
委員　柏原 昭博　教授
委員　西野 哲朗　教授

Study about a tool that simultaneously helps entry-level engineers conduct data analysis using machine learning and acquire knowledge about that

Ryota Kamoshida

Abstract

Machine learning (ML) is one of the most important technology trends, because the volumes and varieties of available data are rapidly growing. Due to the rapid growth of ML demand, entry-level ML engineers who do not know enough about ML and data analysis sometimes engage in the ML data analysis projects. Beginners can easily analyze data using various ML open source software (OSS) and searching for the instructions for using the software on the web. However, if they do not know about proper data analysis procedures, they may not only obtain inadequate results but also draw wrong conclusions from the results.

To address these problems, we have developed an open source Python library, MALSS, which simultaneously helps beginners conduct the ML data analysis and acquire knowledge about the ML data analysis. MALSS helps beginners conduct the ML data analysis by automating the data analysis pipeline. The usual drawback of the automation is that users cannot acquire knowledge about the ML data analysis because the automation treats the analysis process as a black box. MALSS overcomes this drawback and helps beginners acquire knowledge about the ML data analysis by giving them information about the ML data analysis during or after the ML data analysis.

This paper is organized as follows. Chapter 1 describes the background of the study and clarifying our approach and scope of the study as an introduction. We discusses the related technologies of this study in chapter 2. Chapter 3 reveals the challenges of the previous studies and proposes a solution to these challenges that simultaneously helping to conduct the ML data analysis and to acquire knowledge about the ML

data analysis.

In chapter 4, MALSS automates entire process of the ML data analysis and makes an analysis report after the analysis to help users acquire knowledge about the ML data analysis. The proposed method focuses on the supervised machine learning tasks and automates the process of feature engineering and prototyping. Afterwards, MALSS shows the analysis report to users to help them acquire knowledge about supervised machine learning data analysis. The analysis report includes not only the results of the ML data analysis but also the process of the ML data analysis and the common pitfalls in the ML data analysis, which helps users to acquire the appropriate knowledge about that.

We validated the effectiveness of our approach by using test datasets and by running a learning effect test. The results shows that users can conduct the ML data analysis appropriately by using MALSS and acquire knowledge about the ML data analysis through the analysis.

We extend MALSS so that it can help the ML data analysis in accordance with the training data and the result of the subprocesses of the ML data analysis through a graphical user interface (GUI) in chapter 5. Automating whole process of the ML data analysis makes it difficult to help ML data analysis in accordance with the input training data and the results of the subprocesses of the ML data analysis. To handle the problem, we extended the MALSS. The extended MALSS has GUI. The GUI enables users to conduct the ML data analysis for each subprocess and to change the settings of the subprocess in accordance with the inputs of the users. The proposed method helps users to acquire knowledge about the ML data analysis by presenting the information about the anaysis through GUI not using the analysis report.

We validated the effectiveness of our approach by using test datasets and by running a learning effect test in the same way as the previous chapter. We confirmed that users using the proposed extended MALSS outperformed those using the conventional MALSS and acquired basic knowledge about supervised ML data analysis the same as those using the conventional MALSS.

In chapter 6, we discuss how to help to conduct unsupervised machine learning

data analysis. Most AutoML OSS, which automates the process of the ML data analysis, focuses on supervised learning tasks, and little attention has been given to unsupervised learning tasks. Our proposed method help users to conduct clustering data analysis, which is an unsupervised learning task. The MALSS automates the process of clustering data analysis and help users to estimate the appropriate number of the clusters, which is one of the main purposes of clustering data analysis. Furthermore, the MALSS helps users to acquire knowledge about clustering data analysis by generating an analysis report after automated clustering data analysis is conducted.

We validated the effectiveness of our approach by using open datasets and by running an experiment on a crowdsourcing platform.

In chapter 7, we summarize to what extent MALSS can help to conduct machine learning data analysis and acquire the knowledge about that based on the proposed methods.

We confirmed that MALSS can help beginner data scientists to conduct machine learning data analysis and acquire the knowledge about that. Furthermore, we also showed that MALSS can help the beginners recognize when the results of the data analysis are inadequate.

We summarize the results of our study and discuss the future works in chapter 8. In this study, we proposed the system that simultaneously helps beginners conduct the ML data analysis and acquire knowledge about the ML data analysis. The results of our studies indicate the validity of the proposed system.

機械学習を用いたデータ分析と知識習得を同時に支援するソフトウェアの研究

鴨志田 亮太

概要

技術の進歩により扱えるデータの量と種類が急速に増加したことで，収集したデータを分析するために機械学習技術が注目を集めている．機械学習技術のニーズが急増したため，特にビジネスの分野では，経験や知識の不足した担当者が分析に従事せざるを得ないことがある．近年はオープンソースソフトウェアのライブラリを利用することで，高度な機械学習アルゴリズムを容易に利用することが可能になっている．しかし，データ分析に関する適切な知識が不足している場合，誤った分析手順により間違った分析結果を得たり，間違った分析結果から誤った意思決定を行う恐れがある．

このような課題に対し，本研究ではデータ分析の知識を習得しながら実際に分析を行うことのできるツール MALSS（Machine Learning Support System）を開発した．MALSS は機械学習を用いたデータ分析のプロセスを自動化することで，知識や経験が不足した分析者でも適切な手順で分析を遂行することを可能とする．しかし，分析の自動化は分析プロセスの中身をブラックボックス化してしまうため，分析者が分析に関する知識を習得することができない．機械学習を用いたデータ分析に関する適切な知識が身につかなければ，ツールの支援範囲を超えた発展的な分析を行えるようにならない．そこで MALSS では分析中または分析後に，機械学習を用いたデータ分析に関する情報を提供し，分析者の知識習得を支援する．

本論文は以下のように構成される．第 1 章は序論として，研究の背景を述べ，本研究の目的と意義を明らかにする．第 2 章では従来研究について論述し，第 3 章で従来研究の課題について述べた後，課題の解決策として，分析の遂行と知識習得の同時支援という本研究の提案を行ったうえで，本研究のスコープを明確にする．

第 4 章では，分析全体を一気通貫に自動化し，自動化した分析完了後に分析レポートを用いて分析者の知識習得を支援する手法を提案する．提案手法は教師あり学習をターゲットとし，データの前処理，モデルの学習，評価のプロセスを自動化することで分析遂行を

支援する．さらに，分析終了後に分析レポートを生成し分析者に提示することで，分析者の教師あり学習に関する知識習得を支援する．分析レポートには分析結果だけでなく，データ分析のプロセスや，各プロセスにおける実行時の注意点についても記載することで，分析者がこれらのプロセスに関する適切な知識を習得することを支援する．

提案手法の有効性を検証するために，模擬データ分析実験と知識確認テストを行い，MALSS を利用することで適切なデータ分析を実施可能であることと，MALSS を利用したデータ分析を通して機械学習を用いたデータ分析に関する知識を習得できることを確認した．

第 5 章では，グラフィカルユーザーインタフェース（GUI）を用いて，データ分析プロセスの途中結果に応じて分析実施内容を変更する手法を提案する．分析自動化と分析レポートによる分析支援では，分析プロセス全体を自動化するため，分析プロセスの途中結果に応じて分析内容を変更することが困難であるという課題があった．そこで提案手法では，分析のサブプロセスごとの実行を可能とし，GUI を介して，サブプロセスの実行結果に応じて分析者の入力を受け付け，プロセスの実施内容を変更可能なように MALSS を拡張した．提案手法では分析完了後の分析レポートではなく，分析の途中に GUI 上に必要な情報を提示することで，分析者の知識習得を支援する．

本章においても，提案手法の有効性を検証するために，第 4 章と同様の模擬データ分析実験と知識確認テストを行った．検証の結果，GUI を介して分析プロセスの途中結果に応じて分析内容を変更することで，より適切な分析支援が可能なことと，GUI を用いた分析支援により，分析レポートと同等の知識習得が可能であることを確認した．

第 6 章では，教師なし学習のうち，クラスタリング分析の支援方法について論じる．機械学習を用いたデータ分析プロセスを自動化する AutoML は様々な OSS ツールが提案されているが，その多くが教師あり学習にフォーカスしている．提案手法では，教師なし学習の代表的なタスクの 1 つである，クラスタリング分析を支援する．クラスタリング分析の主要な目的の一つである，最適なクラスター数推定を，分析プロセスの自動化により支援し，分析レポートによりクラスタリング分析に関する知識習得を支援する．

提案手法の有効性を検証するために模擬データ分析実験と，クラウドソーシングを利用した知識確認テストを行った結果，提案手法は既存手法と比較しても高精度にクラスター数推定が可能であることを確認し，分析レポートは，特に知識の不足した分析者の知識習得に効果的であることを確認した．

第 7 章では，第 4 章から第 6 章までの提案を踏まえ，第 3 章で述べた MALSS のスコープにおいて，何をどの程度まで支援することができるのかを整理する．

本研究の提案手法により，MALSS は支援対象範囲において，データサイエンティスト協会が定める見習いデータサイエンティストに求められる，分析遂行および分析に関する知識習得を支援できることを確認した．また，MALSS は適切な分析結果が得られないケースにおいても，分析結果が不十分であることに気がつくことができるよう支援することによって，分析者が与えられた責務を果たすことを支援することができることを示した．

第 8 章では本研究の結果をまとめ，得られた研究成果について述べるとともに，今後の課題について整理した．本研究では，機械学習を用いたデータ分析の知識や経験の不足した分析者が分析の実務に従事する際に，適切に分析を遂行することを支援すると同時に，分析者が機械学習を用いたデータ分析に関する知識を習得することを支援するツールを開発し，本ツールが有用であることを示した．

目次

図目次

表目次

第 1 章

序論

1.1 研究の背景

コンピュータの性能向上や多種多様なセンサの普及，通信ネットワークの高速化などにより，収集・蓄積が可能なデータの種類と量が急激に増大している [3]．そのため，収集・蓄積したビッグデータを分析し，ビジネスに有効な知見を発見（データマイニング）し意思決定に活かしていくデータサイエンティストという職種に注目が集まっている [4]．有名な企業レビューサイト Glassdoor が公表している Best Jobs in America [*1] において，データサイエンティストは 2016 年から 2019 年まで 1 位に，2020 年も 3 位にランキングされている．マッキンゼーは，米国では 2018 年までに深い分析スキルをもつ人材が 14 万から 19 万人，ビッグデータを分析し意思決定するマネージャーやアナリストが 150 万人不足すると報告している [3]．人材不足の課題は日本も例外ではない．同マッキンゼーの報告によると，高度なデータ分析の訓練を受けた大学卒業生の数は，米国や中国，インドなどの人口の多い国のみならず，ポーランドや英国，フランスなど，日本より人口の少ない国と比較しても少ないことが指摘されている．さらに国内において，このようなデータ分析スキルを有する人材の数は，2005 年から 2008 年において減少傾向にある．

データサイエンティストに求められるスキルの一つに，収集・蓄積されたデータから有用なルールを抽出し，新たなデータに対する予測などを行う機械学習が挙げられる [5]．これまで，機械学習技術の利用には専門的なスキルが求められたが，近年，scikit-learn [6]，TensorFlow [7] など，オープンソースの機械学習ツールが充実してきたことで，専門家でなくとも容易に機械学習技術を利用することが可能となった．

[*1] https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0 20.htm (2021/11/13 access)

特に，AutoML（Automated Machine Learning）とよばれる分析自動化ツール [8, 9, 10, 11] は，特徴量エンジニアリングやプロトタイピングといった機械学習を用いたデータ分析のプロセスを自動化することで，機械学習を用いたデータ分析に関する知識を一切もたなくとも，数行のプログラミングコードを書くだけで分析を遂行することを可能とする．AutoML 以外にも，GUI を用いてプログラミングすることなしに機械学習を用いたデータ分析のプロセス構築を支援する，いわゆるノーコード，ローコード開発ツールとよばれる OSS も存在する [12, 13].

このようなツールの充実と，人材不足という背景から，機械学習に関する知識・経験が浅い者がデータ分析業務に従事した場合，機械学習技術自体はツールを利用できたとしても，分析手順の誤りなどにより十分な分析結果を出すことができないことがある [14]．分析手順の誤りにより誤った分析結果を得た場合，その結果に基づき誤った意思決定をする恐れもある．本来であれば，実務に従事する前に On-the-Job Training などにより，十分な時間をかけて知識を習得し，その後実務に従事するべきであるが，前述のように人材不足という背景から，そのような時間を確保することができずに実務にあたらざるを得ない分析者も多く存在するのが，実業分野における実情である．

1.2 本研究の目的と意義

前述の背景を受けて本研究では，分析者の知識・経験不足に基づくデータ分析の質低下の問題を解決することを目的とする．この目的を達成するために，筆者は，分析の遂行と知識の習得を同時に支援することが可能なシステムを提案する．提案するシステムは，MALSS（MAchine Learning Support System）と名づけ，Python のオープンソースソフトウェアライブラリとして開発している [*2].

MALSS は分析者の機械学習を用いたデータ分析遂行を支援するために，分析を自動化する．分析を自動化することで，知識・経験不足の分析者が分析を行った場合でも，適切な分析手順で一定以上の質の分析を行うことが可能となる．

しかし，分析を自動化するだけでは，分析プロセスがブラックボックス化してしまい，知識習得に結びつかず，分析者が発展的な分析を行うことができない．そこで MALSS は分析中あるいは分析後に，機械学習を用いたデータ分析に関する情報の提示を行う．分析者は提示された情報を参照することで，分析を行いながら必要な知識を習得することができ，分析結果を正しく理解したうえで，次の分析施策を立案することが可能となる．

[*2] http://pypi.python.org/pypi/malss/

本研究では，初めに教師あり学習の回帰タスクと分類（識別）タスクをターゲットとし，機械学習を用いたデータ分析における特徴量エンジニアリングとプロトタイピングのプロセスを一気通貫に支援する方法を提案する．本手法では，分析の完了後に分析レポートという形でデータ分析に関する情報を分析者に提示し，分析者の知識習得を支援する．分析遂行支援および知識習得支援の有効性は，模擬データ分析実験と知識確認テストにより評価した．

次に，分析プロセスをサブプロセスごとに遂行し，サブプロセスの実行結果に応じて分析者の判断を受け付け，判断に応じ，以降の分析内容を変更する手法を提案する．分析プロセス全体を一気通貫に支援する方式は簡便である一方，分析結果に応じた分析内容の変更が困難であるという課題を有する．そこで本提案手法では，グラフィカルユーザーインタフェース（GUI）を備え，サブプロセスごとに分析者の入力を受け付けることで，分析結果に応じた分析内容の変更を可能とする．知識習得支援はこの GUI を介して分析に関する情報を提示することで行う．本提案手法の有効性も，模擬データ分析実験と知識確認テストにより評価した．

最後に，提案手法を教師なし学習に適用する．教師なし学習は教師あり学習と異なり，データに正解ラベルとして利用できる情報を含まず，分析結果を分析者自身が解釈する必要があるため，分析の自動化が困難である．そのため多くの AutoML の OSS は教師あり学習をターゲットとしている．本研究では，教師なし学習の中でクラスタリングタスクにフォーカスし，適切なクラスター数の推定に分析目的を限定することで，教師なし学習の分析遂行支援と知識習得支援を可能とする．本提案手法の有効性は，模擬データ分析実験と，クラウドソーシングを利用した知識確認テストにより評価した．

以上より，本研究の意義は，機械学習を用いたデータ分析における，教師あり学習と教師なし学習のクラスタリングタスクをターゲットとし，知識が経験が不足した分析者の分析遂行と，分析に関する知識習得を同時に支援するシステムの提案と，実際に開発した OSS ツールを用いた有効性の検証，である．

1.3 本論文の構成

本論文の以降の構成は以下のとおりである．第 2 章において，既存の分析遂行支援手法と知識習得支援手法について述べる．第 3 章では既存手法の課題について述べ，課題解決のアプローチとして，分析の遂行と分析に関する知識習得を同時に支援するシステムを提案する．また，提案手法のスコープを明確にする．第 4 章では，提案システムについて，教師あり学習をターゲットとし，分析の自動化と分析後の分析レポートを用いた支援方法

について説明し，第 5 章で，分析支援範囲を広げるための，グラフィカルユーザーインタフェースを用いた分析中のインタラクティブな分析遂行・知識習得支援方法を提案する．第 6 章では提案手法の教師なし学習への拡張について述べる．第 7 章で本研究全体を通して MALSS が支援できる範囲を明らかにし，最後に第 8 章でまとめと今後の課題について述べる．

第 2 章

従来研究

本章では，分析の遂行支援と，分析に関する知識習得支援の従来技術について説明する．2.1 節で分析遂行支援について，2.2 節で分析に関する知識習得支援について述べる．

2.1　分析遂行支援

分析遂行支援の方法としては，分析パイプラインや機械学習モデルの作成などを支援する方法と，分析プロセスを自動化する方法（AutoML）が挙げられる．

KNIME [12] や Orange [13] といった OSS は，GUI を用いてプログラミングすることなしにデータ分析のワークフローを構築することを支援する．図 2.1 は KNIME を用いて分析ワークフローを構築する例である．この図では決定木アルゴリズムを用いて分類（識別）タスクを遂行している．図のように，GUI 上で分析プロセスに相当するノードを配置し，ノード同士をコネクタで接続することで分析プロセスを構築することができる．このようにワークフローの構築を簡便にすることで，分析者の分析遂行を支援することができる，しかしこの支援方法は，分析者が適切なデータ分析のワークフローを構築する知識をもっていることを前提としている．一方 MALSS は，このような知識をもたない分析者でも，適切な分析ワークフローで分析を行えるよう支援することを目的としている．

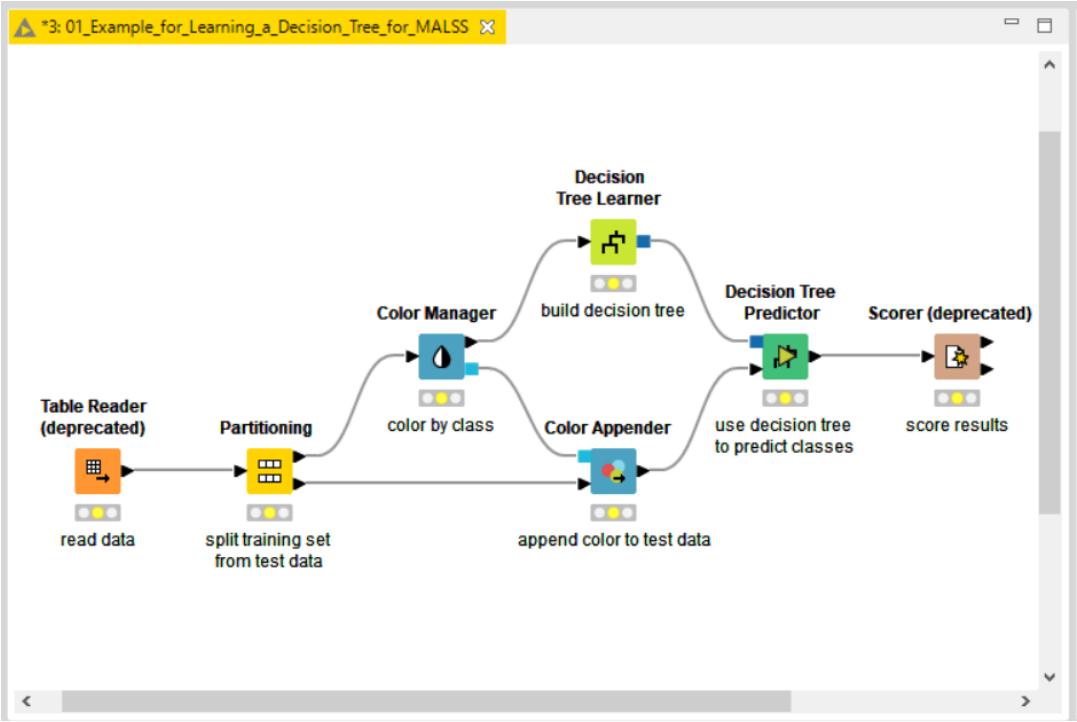


図 2.1　KNIME による分析ワークフロー構築の例.

AutoML の OSS としては，auto-sklearn [8]，TPOT [9]，H2O [10]，ATM [11] が知られている．これらのソフトウェアは，機械学習の中でもラベルや数値を予測する教師あり学習をターゲットとし，特徴量の生成や選択，複数の予測モデルの学習，ハイパーパラメータの調整やモデルのアンサンブル（複数のモデルを組み合わせより高性能なモデルを作成する）など，機械学習の一連のワークフローを自動化する．auto-sklearn は Python の機械学習ライブラリとしてデファクトスタンダードになっている scikit-learn [6] と同じような API で分析の自動化を行うことが可能であり，ベイズ最適化により効率的にハイパーパラメータの調整を行うことができるという特徴をもつ．TPOT はモデルの性能とシンプルさのトレードオフを考慮したパレート最適なモデルを作成することができる．H2O は本来はモデル作成から展開（deploy）までを支援するプラットフォームであるが，分析自動化機能も提供している．ATM は複数ユーザーの利用を可能とする分散型の AutoML システムであり，教師あり学習の一つである分類タスクのみをサポートしている．

図 2.2 は auto-sklearn を用いて分析を行う Python スクリプトの例である．この例では，データの読み込みなどのコードは省略している．AutoML に関係するコードはこの 4 行のみであり，この 4 行で分析者はアルゴリズムや分析プロセスの詳細を意識することな

く，アルゴリズムの選択やハイパーパラメータの調整といった一連の分析プロセスを自動で遂行することができる．

```
1 import autosklearn.classification
2
3 cls = autosklearn.classification.AutoSklearnClassifier()
4 cls.fit(X_train, y_train)
5 predictions = cls.predict(X_test)
```

図 2.2　auto-sklearn ライブラリを用いて分類（識別）タスクを実行する Python コード例

MALSS は機械学習を用いたデータ分析に関する知識や経験が不足した分析者が，適切な手順で分析を遂行することを支援するために，分析の自動化，つまり AutoML による分析遂行支援を行う．通常の AutoML では，分析の遂行を自動化するため，分析の中身がブラックボックスとなり，知識や経験が不足した分析者は分析に関する知識を身につけることができない．MALSS は，分析後あるいは分析中に機械学習を用いたデータ分析に関する情報を提示することで，分析の遂行を支援するのと同時に，分析者が分析に関する知識を身につけることも支援する．

2.2　知識習得支援

MALSS が支援対象とするのは機械学習を用いたデータ分析の中で，特徴量エンジニアリングとプロトタイピングのプロセスである．具体的には，それぞれのプロセスにおいてどのような処理（特徴量エンジニアリングのプロセスにおけるカテゴリ変数の処理や，プロトタイピングのプロセスにおける交差検証など）を行う必要があり，その処理を行う際にどのような点に気を付ける必要があるか（モデルが訓練データに過剰に適合した過学習状態にならないよう気を付けるなど）といった知識である．

こうした知識習得支援の方法としては，書籍や技術情報共有 Web コンテンツなどが一般的である．機械学習技術が注目を集めるにつれ，関連書籍は急激に増加している．図 2.3 は Google Ngram Viewer [*1]で，書籍中の machine learning という単語の出現頻度を可視化したものである．図をみると，機械学習技術が世の中の注目を集め始めた 2012 年以降，急速に関連書籍が増えていることが見て取れる（2012 年はトロント大のチームが開発した深層学習モデル AlexNet [15] が，物体認識のコンペ ILSVRC で大差で優勝し注目を集めた年である）．

[*1] https://books.google.com/ngrams/ (2021/9/20 access)

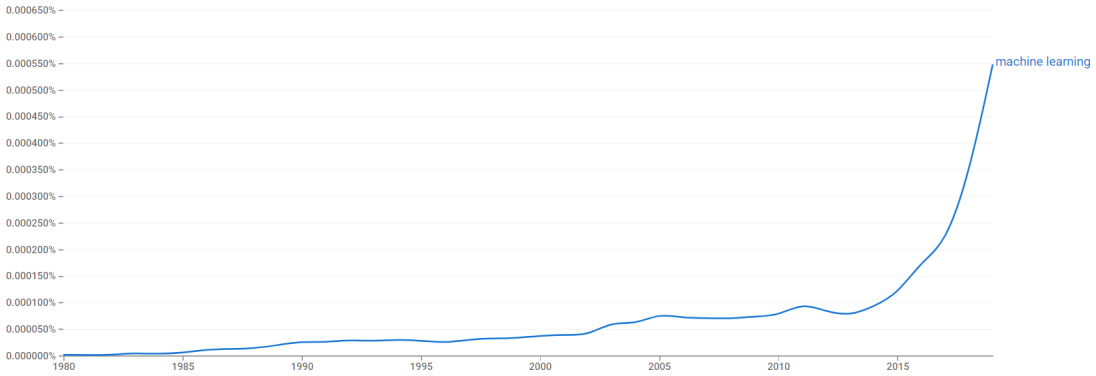


図 2.3　Google Ngram Viewer による machine learning の検索結果（2021.09.20 時点）.

書籍や Web コンテンツの他に，学習支援を対象としたシステムが古くから提案されている．「プログラム学習（Programmed Instruction）」は，細かく分割した問題を順番に提示し，生徒の回答に対して正誤の応答を即座に与える過程で学習が行われるように工夫された学習方法であり，現在の e-learning などの基礎となる考え方である．これを実現するシステムとして，古くはティーチングマシン（Teaching Machines）が提案されている [16]．ティーチングマシンは，従来の講師による講義と比較して，生徒が自分のペースで学習を進められるという利点をもつ．一方，学習プロセスが直線的であり，全ての学習者が同じプログラムを学習しなければならないという欠点も指摘されていた．

このようなティーチングマシンの課題に対応するため，適応型学習システム（Adaptive Learning Systems）が提案された．適応型学習システムでは，生徒個々の予備知識や学習スタイルの違いなどを識別し，その違いに応じた学習環境を提供することを志向している [17]．近年では，より知的な適応を可能とするために，あるいは生徒の学習体験を向上させるために，機械学習技術やロボット技術を適用する取り組みも盛んである [18]．これらは特に知的学習支援システム（Intelligent Tutoring Systems）とよばれている．

プログラム学習に関連する技術の進展や，ネットワーク環境の向上により動画コンテンツを容易に利用できるようになってきている社会的な状況を背景に, 近年は Massive Open Online Course（MOOC）とよばれるオンライン講義が注目を集めている．Coursera [*2], edX [*3]，Udacity [*4]などが有名であり，受講者は機械学習やデータ分析に関する多くの講義を無料で受講することができる（有料の講義も存在する）．これらの講義では，受講者はプログラミングタスクなどを通じて実践的なスキルを身につけることも可能である．

[*2] https://www.coursera.org/ (2022/2/10 access)
[*3] https://www.edx.org/ (2022/2/10 access)
[*4] https://www.udacity.com/ (2022/2/10 access)

以下のリストは，機械学習のオンラインコンテンツで最も有名なものの一つである，Coursera で Andrew Ng 氏が講義する Machine Learning 講座 [*5]のシラバスである．括弧内の数字は修了に要する履修時間であり，合計すると修了に 26 時間を要する．

1. Introduction (2h)
2. Linear Regression with One Variable (2h)
3. Linear Algebra Review (2h)
4. Linear Regression with Multiple Variables (3h)
5. Octave/Matlab Tutorial (5h)
6. Logistic Regression (2h)
7. Regularization (5h)
8. Neural Network Representation (5h)

企業では，新人研修 [*6]や，On-the-Job Training の研修 [19] として，機械学習に関する教育機会が用意されていることが多い．

このように様々な知識習得支援方法が存在するが，上に述べたように，一定の知識を身に着けるために多くの時間を要することが一つの課題である．本研究では，機械学習を用いたデータ分析に関する知識や経験が不足した分析者が，分析者の不足という背景から十分な準備時間を与えられずに分析の実務に従事するケースを想定しており，このようなケースにおいて，書籍や MOOC などのオンライン講義で十分に時間をかけ知識を習得してから実務に従事するということは現実的ではない．そこで MALSS は，機械学習を用いたデータ分析に関する基礎的な知識に絞り，業務の遂行と同時に知識の習得を行うことを可能とする．

[*5] https://www.coursera.org/learn/machine-learning (2021/11/13 access)

[*6] https://rand.pepabo.com/article/2020/07/22/machine-learning-introduction/ (2021/11/13 access)

第 3 章

機械学習を用いたデータ分析の遂行とデータ分析に関する知識習得の同時支援

3.1　従来技術の課題

第 2 章で述べたように，分析遂行支援，知識習得支援のための種々の方法が提案されているが，知識や経験が不足した分析者が実務に従事する際の支援手段としては課題が残っている．課題の 1 つは，多くの知識習得支援方法は実際の業務とは独立しており，実務に従事する前に十分な時間をかけて知識を習得する必要があることである．従来の知識習得支援方法では，データ分析人材の不足を背景に，今まさに目の前の実務に従事する必要のある，知識や経験が不足した分析者の支援を行うことは難しい．このような課題に対し，AutoML 技術は，機械学習を用いたデータ分析の一連のプロセスを自動化することで，知識や経験が不足した分析者であっても適切な手順で分析を行うことを可能とする．しかし，AutoML は分析の中身をブラックボックス化するため，分析者が機械学習やデータ分析に関する正しい知識を習得することができず，誤った分析結果に基づき誤った意思決定をする恐れや，AutoML ツールの支援範囲を超えた分析を自力で行うことができるようにならないという課題がある．

また，従来の AutoML OSS はほとんどが教師あり学習にフォーカスしており，教師なし学習をサポートしていない．正解データを含まないデータから何らかの規則性を発見することを目的とする教師なし学習は，教師あり学習と並んで機械学習によるデータ分析における主要なタスクの一つである．教師なし学習はデータに正解を含まないため，デー

タ分析の結果の妥当性を主観的に評価する必要があり，そのためデータ分析のプロセスを自動化することが困難である．上述の AutoML OSS の中では，H2O [10] のみが教師なし学習の一つであるクラスタリング分析のアルゴリズムである k-means アルゴリズムをサポートしている．H2O はクラスター数推定機能をもち，データ前処理の一つであるデータ正規化機能を含むため，分析の一部を自動化していると言うことができる．しかし H2O は分析に関する知識習得を支援することはできない．

3.2 課題に対する提案

3.1 節で述べたような，分析者の知識・経験不足に基づくデータ分析の質低下の問題に対する従来技術の課題を解決するために，筆者は，分析の遂行と知識の習得を同時に支援する方法を提案する．分析遂行の支援は機械学習を用いたデータ分析プロセスの自動化（AutoML）により実現する．分析を自動化することで，知識・経験不足の分析者が分析を行った場合でも，適切な分析手順で一定以上の質の分析を行うことが可能となる．それに加え，自動化による分析遂行支援の際に，機械学習を用いたデータ分析に関する知識を提示することで，分析遂行中の分析者の知識習得を支援する．分析遂行と同時に分析者の知識習得を支援することで，分析プロセスがブラックボックス化してしまうことを防ぎ，分析者は分析結果を正しく理解したうえで，次の分析施策を立案することが可能となる．

提案手法のシステムは，汎用プログラミング言語 Python のオープンソースソフトウェア（OSS）ライブラリとして開発している [*1]．

3.3 支援のスコープ

本節では，本研究で提案する，知識や経験が不足した分析者の分析遂行と分析に関する知識習得を支援するシステム MALSS の支援範囲を明確にする．

3.3.1 データ分析のタスク

機械学習のタスクは，データの内容やモデリングの基準によって表 3.1 に示すように，教師あり学習と教師なし学習に大別される．教師あり学習は，分析するデータに予測したい正解の値が含まれており，教師なし学習は正解の値が含まれていない．教師あり学習の代表的なタスクとしては，正解のデータがラベルである分類（識別）と，正解のデータが

[*1] http://pypi.python.org/pypi/malss/

数値である回帰がある．教師なし学習の代表的なタスクとしては，類似のデータをカテゴライズするクラスタリングや，よく出現するパターンを見出す頻出パタンマイニング，他のデータから大きく外れたデータを見つけ出す外れ値検出などがある．

教師あり学習はデータに正解の値を含むという性質から，教師なし学習に比べると，モデルの性能評価基準が明確であり，モデリング，および性能評価のプロセスを定型化しやすい．そのため，3.1 節でも述べたように，多くの AutoML OSS が教師あり学習にフォーカスしている．

MALSS は，教師あり学習の回帰タスクと分類（識別）タスクと教師なし学習のクラスタリングタスクを支援する．それに加えて，クラスタリングタスクにおいて，支援対象を最適なクラスター数の推定にフォーカスすることで，教師なし学習の支援も対象とする．

表 3.1　機械学習のタスク分類

教師あり学習	教師なし学習
回帰	クラスタリング
分類（識別）	頻出パタンマイニング
	外れ値検出
	等

3.3.2　データ形式

本研究ではテーブルデータ（Tabular data）とよばれる表形式で表されるデータ形式を入力とするデータ分析を支援対象とする．機械学習を用いたデータ分析が対象とするデータ形式は構造化データと非構造化データに大別され，構造化データの代表的なデータがテーブルデータである（表 3.2 参照）．非構造化データはデータ種別に応じたデータハンドリングやアルゴリズムが必要になることが多い一方，テーブルデータはフォーマットが構造化されているため多くの機械学習ライブラリが対応しており，共通の前処理技術やアルゴリズムを利用できるという利点がある．

表 3.2　分析に用いられるデータ形式

構造化データ	非構造化データ
テーブルデータ	画像データ
	音声データ
	文書データ
	センサデータ
	等

3.3.3　データ分析のプロセス

CRISP-DM

機械学習を用いたデータ分析のプロセスは，古くは CRISP-DM（Cross-Industry Standard Process for Data Mining）[20] とよばれるデータマイニングの標準プロセスで説明されることが多い．CRISP-DM は，SPSS，Teradata，Daimler AG，NCR，OHRA がメンバーとなっているコンソーシアムから提案されたデータマイニング標準プロセスの一つであり，図 3.1 のようなプロセスで表される．CRISP-DM のデータ分析プロセスは，Business Understanding，Data Understanding，Data Preparation，Modeling，Evaluation，Deployment の 6 つのフェーズから構成される．図中の矢印は重要かつ頻繁に発生するプロセスの流れを表しており，各フェーズの順序は固定されているわけではない．

上記データ分析プロセスの中で，Data Preparation および Modeling は個別プロジェクトへの依存度が比較的小さく，定型化の余地が大きいため，自動化による支援の効果が大きい．そこで MALSS では Data Preparation，および Modeling プロセスを支援することを目的とする．

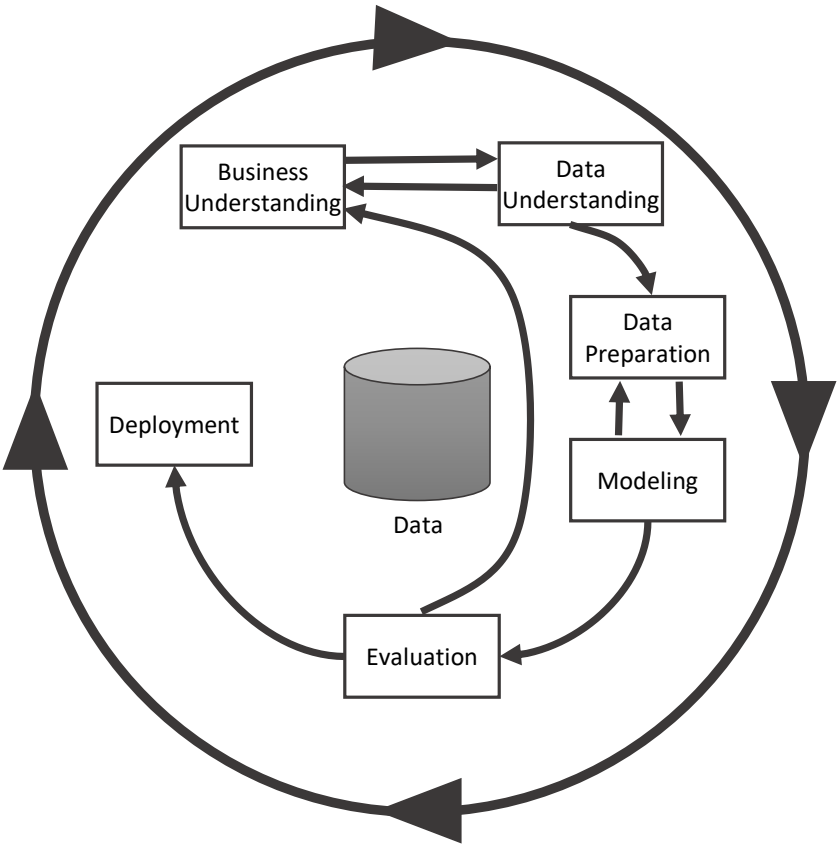


図 3.1　CRISP-DM のデータ分析プロセス

MLOps

近年では，データ分析のプロセスを，機械学習を利用したシステムの開発から運用（MLOps）までの一貫したプロセスとして整理することも増えてきている [1, 21].

図 3.2 は Google 社が整理する MLOps のライフサイクルである [1]．MALSS は，MLOps のライフサイクル全体プロセスの中で，「機械学習モデル開発」の主要プロセスである「実験とプロトタイピング」内の「特徴量エンジニアリング」と「プロトタイピング」のプロセスを支援する．特徴量エンジニアリングのプロセスは CRISP-DM における Data Preparation のプロセスに，プロトタイピングのプロセスは Modeling のプロセスにそれぞれ対応している．

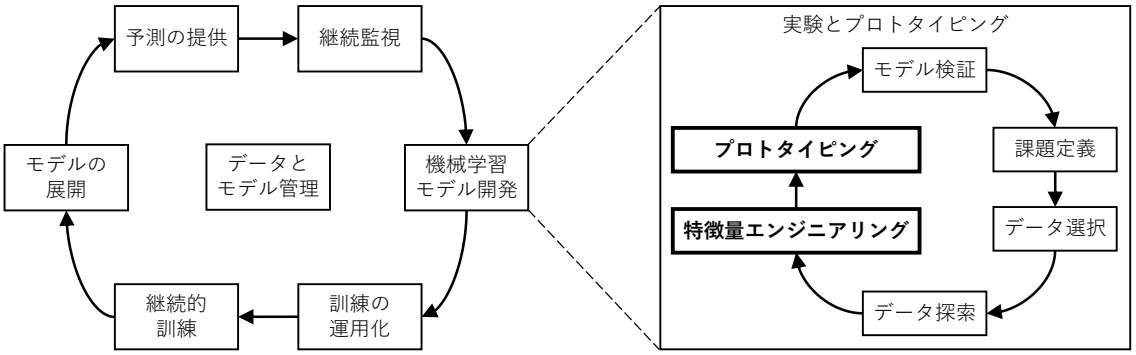


図 3.2　MLOps のライフサイクル．先行研究 [1] の図 2 と図 5 から著者が再構成．太字部分が MALSS の支援範囲.

本研究では，上記の特徴量エンジニアリングとプロトタイピングのプロセスを，2 種類の方法で支援する．1 つ目は，上記のプロセス全体を一気通貫に自動化し，自動化した分析の完了後に，分析レポートという形で分析に関する知識を提示することで，分析者の知識習得を支援する方法である．本手法については，教師あり学習を対象とした手法について第 4 章で，教師なし学習を対象とした手法について第 6 章にて詳細を述べる．

2 つ目の支援方法では，分析をサブプロセスごとに遂行し，サブプロセスの分析結果に応じて分析者の判断を受け付け，分析内容を変更する．分析者の判断の受け付け，および分析者の知識習得を支援するための分析に関する情報の提示は，グラフィカルユーザーインタフェース（GUI）を介して行う．本手法については教師あり学習を対象とした手法について第 5 章で詳細に説明を行う．

3.3.4　分析者の分類

Harris らによる分類

データ分析者の分類としては Harris らの分類が有名である．Harris らは，分析者を，彼らへのアンケートをもとに，Data Businessperson，Data Creative，Data Developer，Data Researcher の 4 タイプに分類した [5]．表 3.3 は参考文献中の分類図を表形式にまとめたものである．表の列が分析者の分類を，行がスキルを表しており，タイプごとの各スキルの高さを低・中・高の 3 段階で表している．

Data Businessperson タイプはビジネススキルが高く，プログラミングスキルは必ずしも高くない．このタイプにはデータ分析のコンサルタントやデータ分析企業のマネージャなどが該当する．本人が分析を行うというよりは，分析された結果に基づき意思決定などを行うことが多い．

Data Researcher タイプは統計スキルが特に高く，大学や研究機関の研究者などが該当する．データ分析により新たな知見を見出すことが主目的であり，システム開発を行う必要はないため，分析には BI ツールなどを利用することが可能であり，必ずしも高いプログラミングスキルは求められない．

Data Creative，および Data Developer はプログラミングスキルが高く，データ分析の結果をもとにシステム開発を行う人々である．

本研究の目的は，Data Creative や Data Developer のようなプログラミングを伴うデータ分析作業者の中で，特に機械学習の知識・経験が十分でない分析者を対象として，データ分析のプロセスにおいて定型化可能な部分を自動化することで支援するとともに，機械学習によるデータ分析に関する情報提供を行うことで知識習得を支援することである．

表 3.3　データ分析従事者の分類

	Data Businessperson	Data Creative	Data Developer	Data Researcher
ビジネス	高	低	低	低
機械学習／ビッグデータ	中	高	高	低
数学／オペレーションズリサーチ	低	低	中	中
プログラミング	低	高	高	低
統計	中	高	低	高

Kim らによる分類

Kim らは，Microsoft のデータサイエンティスト 793 名を対象に調査を行い，データサイエンティストを表 3.4 に示す 9 つのクラスターに分類している [2]．MALSS は，データ分析を行うデータサイエンティストである「データ形成者」と「データ分析者」を支援対象とする．データ形成者とデータ分析者は，Harris らの分類では Deta Creative と Data Developer に該当すると考えられる．Kim らの調査では，データサイエンティストの役割ではないが，職務の一部としてデータ分析を行う「副業者」というクラスターが定義されている．しかし，「副業者」クラスター以外のクラスターに属するデータサイエンティストであっても，ソフトウェアエンジニアやプログラムマネージャーなど，データサイエンティスト以外の役職でありながらデータサイエンス業務を行っている者が多いことが報告されている．さらに調査の中では，他のエンジニアリングの役割からデータサイエンティストの役割に移行する人たちに対する機械学習などのトレーニングの必要性と，業

務中に新しい知識を習得することの困難さが指摘されている．MALSS はこのような背景において，知識が不十分なまま分析業務に従事せざるを得ないデータサイエンティストの業務遂行と知識習得を支援することを目的とする．

表 3.4　Microsoft におけるデータサイエンティストのクラスター [2]．表は著者が作成．

クラスター	役割	スキル
博識者	全ての活動に従事	多様なスキル
データエバンジェリスト	データ駆動型意思決定の推進	ビジネス・製品開発スキル
データ準備者	データの照会・準備	構造化データ処理
データ形成者	データの準備・分析	アルゴリズム，機械学習
データ分析者	データ分析	統計，数学
プラットフォーム構築者	プラットフォーム構築	ビッグデータ，分散システム
50% 副業者	—	—
20% 副業者	—	—
インサイトアクター	得られた知見に基づき活動	—（小数のため考察対象外）

3.3.5　分析および習得する知識のレベル

MALSS の目的は，知識が不十分な分析者が分析に関する基本的な知識を身に着け，分析自動化で対応できる範囲を超えたより高度な分析を自身で行っていくことができるようになることである．大城らは，AI・データ分析プロジェクトの実務経験があるが，自身の判断で一通りのプロジェクトに関するタスクを進めることはできないデータサイエンティストを，ジュニアデータサイエンティストと定義している [22]．本研究では，AI・データ分析プロジェクトの実務経験が無い，あるいは不十分であるにも関わらず，ジュニアデータサイエンティストの職務を求められている分析者が，ジュニアデータサイエンティストとして基本的な水準の分析を遂行するのと同時に，ジュニアデータサイエンティストとして知っておくべき基本的な知識の習得を支援することを目的とする．3.3.3 節で述べたように，MALSS は CRISP-DM のプロセスの中で，Data Preparation，Modeling のプロセスを支援対象としており，前段の Business Understanding，Data Understanding プロセスは，シニアデータサイエンティストが，あるいはシニアデータサイエンティストと一緒に検討することを想定している．MALSS は，知識や経験が不足しているジュニアデータサイエンティストが，実務の中で繰り返し利用していくことで，ジュニアデータサイエンティストに求められる基本的な知識を習得し，MALSS の支援範囲を超えたより発

展的な分析を行えるようになるまでをサポートすることを目指している．

上記の目的のため，多くの AutoML ソフトウェアが高性能化のために複雑なモデルを学習するのに対し，MALSS はシンプルで基本的なモデリングを行うことを特徴とする．具体的には，MALSS はより良い分析結果を得るために，複数の機械学習アルゴリズムの分析結果を統合するアンサンブル学習をサポートしない．アンサンブル学習は複数のアルゴリズムを統合するという性質上，モデルの学習・評価手順が複雑になるだけでなく，分析結果の解釈も困難になりやすいためである．同様の理由から，既存の特徴量を組合せ新たな特徴量を追加する特徴量生成技術も利用しない．加えて，MALSS は深層学習技術 [23] もサポートしない．深層学習技術はニューラルネットワークアルゴリズムを大規模かつ複雑にしたもので，近年のコンピュータ性能の向上によりその有効性が認識され大きな注目を集めている技術である．しかし，深層学習技術は未だ発展途上の技術であり典型的な活用パターンが確立しているとは言い難いことから，知識や経験が不足した分析者が正しく扱うのは困難であると判断した．

知識習得支援において習得を支援する知識レベルについては，上述の通り，基本的な分析を行うにあたり必要となる知識レベルに限定する．分析者に提示する内容については，主に，Python の機械学習ライブラリとしてデファクトスタンダードになっている scikit-learn のチュートリアルページ [*2] と，MOOC の機械学習講座として最もポピュラーなものの一つである，Andrew Ng 氏が講義する Machine Learning 講座 [*3] の内容と，その他多数の初学者向け教科書を参考に著者が選定した．具体的な提示内容については，以降の各章で詳細を示す．

[*2] http://scikit-learn.org/stable/tutorial/ (2021/11/13 access)

[*3] https://www.coursera.org/learn/machine-learning (2021/11/13 access)

第 4 章

分析自動化と分析レポートを用いた教師あり学習によるデータ分析の支援

本章では，機械学習を用いたデータ分析プロセスの中で，特徴量エンジニアリングとプロトタイピングのプロセスを一気通貫に自動化し，自動化した分析の完了後に分析レポートという形で分析に関する知識を提示することで，分析者の上記分析プロセスに関する知識習得を支援する方法を提案する．

4.1　動機づけの例

初めに，動機づけの例として，不適切な分析により分析結果が不十分となるケースを考える．図 4.1 は Python の機械学習ライブラリ scikit-learn [6] を用いて回帰分析を行う例を示したものである．$\cos(1.5\pi x)$（真の値）に平均 0，分散 0.16 の正規分布に従うノイズを付加した観測値について，目的変数 y を説明変数 x に回帰する単回帰分析であり，アルゴリズムは回帰木 [24] を利用している．外部ライブラリを利用することで利用者はアルゴリズムの中身を意識することなく，数行のコードを書くだけで分析を実行することができる．この例において，学習したモデルは与えられたデータ（訓練データ）に完全にフィッティングしており訓練誤差は 0 である．

しかし図 4.2 に示すように，モデルは真の値を再現しているとはいい難く，未知のデータに対する予測結果の誤差（汎化誤差）も 0 とはならない．これは，モデルが訓練データに過剰に適応している過学習という状態である．過学習を防ぐための手法としては特徴量

選択や次元削減などが，過学習の度合いを評価する方法としては交差検証などが用いられる [25]．しかし，これらの手法は一般的に機械学習アルゴリズムとは独立しており，分析者がその必要性を正しく認識していなければ，適切な手法を選択し利用することができない．

```
from numpy import random, cos, pi, sort, newaxis
from sklearn.metrics import mean_squared_error as mse
import matplotlib.pyplot as plt
from sklearn.tree import DecisionTreeRegressor as dtr

random.seed(0)
true_fun = lambda X: cos(1.5 * pi * X)
X_train = sort(random.rand(30))
y_train = true_fun(X_train) + random.randn(30) * 0.4

clf = dtr().fit(X_train[:, newaxis], y_train)
print u'訓練誤差:', # 0.0
print mse(y_train, clf.predict(X_train[:, newaxis]))

X_test = sort(random.rand(30))
y_test = true_fun(X_test) + random.randn(30) * 0.4
print u'汎化誤差:', # 0.41980081569
print mse(y_test, clf.predict(X_test[:, newaxis]))
```

図 4.1　scikit-learn ライブラリを用いて回帰分析する Python コード例

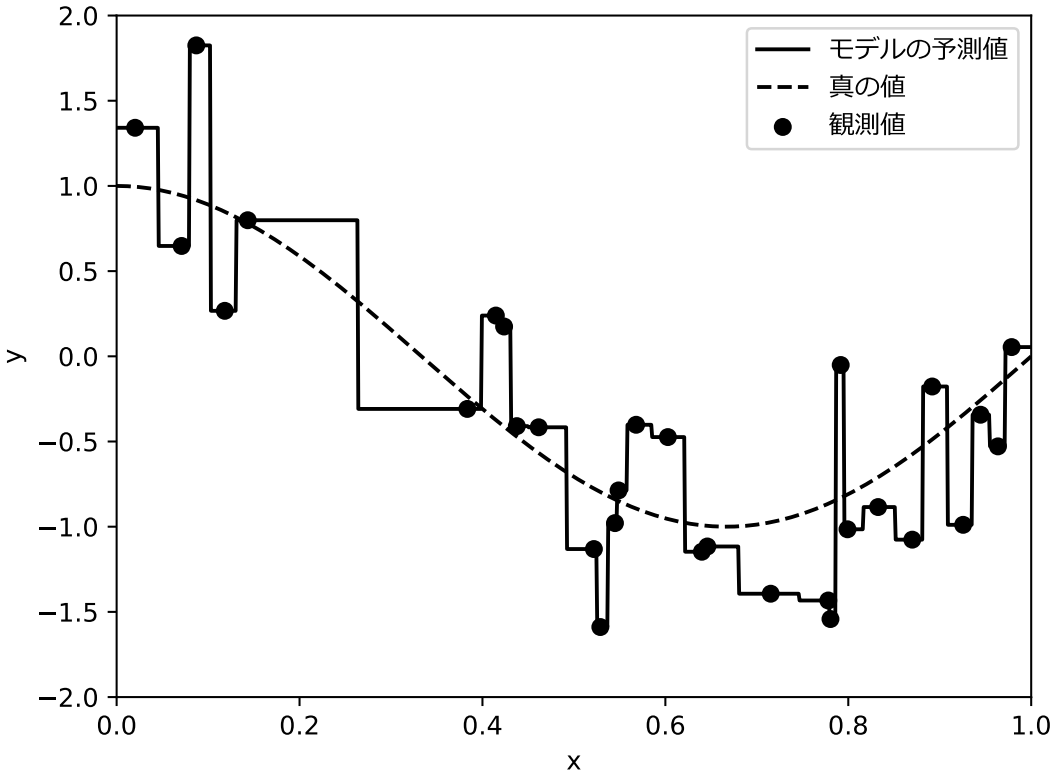


図 4.2　回帰分析結果（scikit-learn 利用）

4.2　自動化による分析支援

4.2.1　ユースケース

図 4.3 に，MALSS のユースケース図を示す．分析目的の決定，およびデータの取得は事前になされていることを前提としている．分析者が MALSS に対して行う操作は，1）分析目的の設定，2）データの設定，3）サンプルコードの出力，の 3 つのみである．MALSS は与えられた分析目的，およびデータに応じて適切な分析を行い，結果を分析レポートとして出力する．データハンドリングや機械学習のコアコンポーネントは，データ解析ライブラリ Pandas [26]，および scikit-learn を利用している．

MALSS の主要な機能は以下の 5 つからなる．

1. データの準備（特徴量エンジニアリング）
2. アルゴリズムの選択

3. 分析
4. 分析レポートの生成
5. サンプルコードの出力

以下において，これらの機能について述べる．

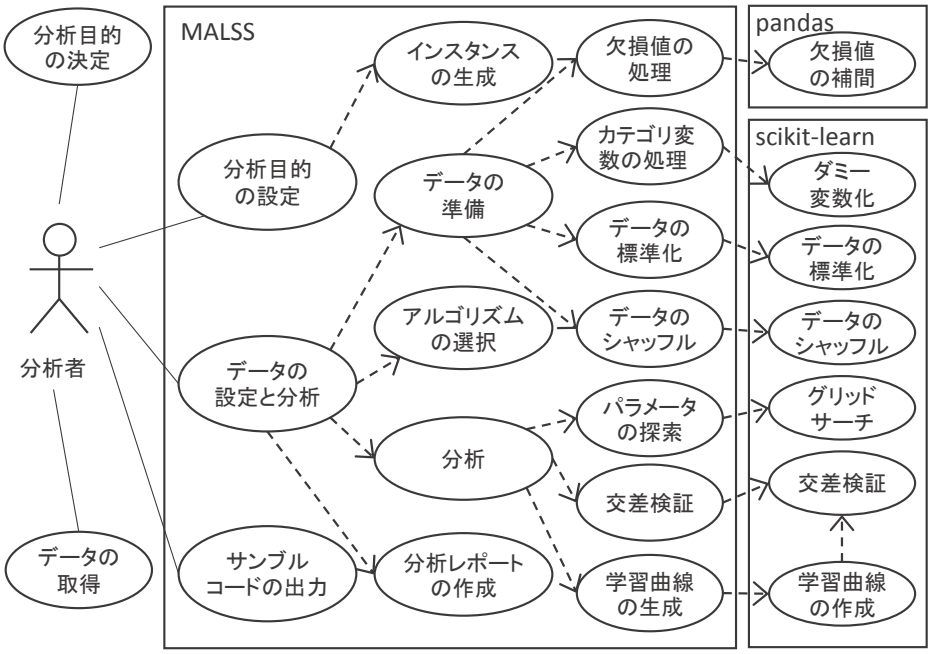


図 4.3　ユースケース図

4.2.2　データの準備（特徴量エンジニアリング）

3.3.3 項で述べた，MALSS が支援するデータ分析プロセスのうちの，特徴量エンジニアリングプロセスは以下の 4 つの機能から構成される．

1. 欠損値の処理
2. カテゴリ変数の処理
3. データの標準化
4. データのシャッフル

以下において，これらの機能について述べる．

欠損値の処理

分析対象となるデータには，ある項目の値が欠落していることがあり，これを欠損値とよぶ．欠損値は，値は分からなくとも，値が欠落していること自体が情報となる場合もある．しかし，機械学習のアルゴリズムが欠損値の入力に対応していない場合が多いため，欠損値は何らかの方法により補間することが一般的であり，欠損値の補間方法としては，

- 同項目の平均値などの統計量で置換する
- 何らかの定数で置換する
- 前後の値から補間する（時系列データ）

などがある [27]．MALSS では，データ種別によらず適用できる汎用的な補間手段として，統計量による置換方式を採用した．変数の種類に応じて，以下のように置換を行う．

- 実数型の変数は同項目の平均値
- 整数型の変数は同項目の中央値
- カテゴリ変数は同項目の最頻値

欠損値補間の有無は分析者が指定可能である．初期値は欠損値補間有とした．

カテゴリ変数の処理

機械学習のアルゴリズムは入力が数値であることを前提としているものが多く，カテゴリ変数をそのまま扱うことができない場合がある．このような場合には，ダミー変数を用いてカテゴリ変数を量的変数に変換することが一般的である [27]．

あるカテゴリ変数の項目 X が m 個のカテゴリをもつとき，この項目を m 個の新たな項目 $D_1, D_2, \ldots, D_m$ で表したものをダミー変数とよび，項目 X の変数 x が i 番目のカテゴリに属するとき，$D_i = 1, D_j = 0\ (j \neq i)$ である．実際には，m 個のダミー変数は冗長性をもつため，通常は $m - 1$ 個のダミー変数を用いる．

MALSS では，入力データがカテゴリ変数をもつとき，自動的にこれをダミー変数に変換する．

データの標準化

データの項目ごとの大きさやばらつきが大きく異なる場合，分析結果に悪影響を与えることがある．そこで，項目ごとに平均が 0，標準偏差が 1 になるように変換を行う標準化処理が一般的に行われる [25]．標準化の式は $z_i = (x_i - \mu)/\sigma$ で表される．ここで，z_i は

標準化後の変数，x_i は標準化前の変数，μ は変数 x の平均値，σ は変数 x の標準偏差である．

データや分析目的によっては，標準化を行うことが適切でない場合もあるため，標準化の有無は分析者が指定することができる．初期値は標準化有とした．

データのシャッフル

機械学習によるデータ分析ではモデルが与えられたデータに過剰に適応してしまう過学習を防ぐために，交差検証により汎化性能を評価する [27]．交差検証手法の一つである K-fold cross validation では，与えられたデータを k 等分する．そしてそのうちの 1 つをテストデータとし，残りを訓練データとして，訓練データでモデリングを行い，テストデータを用いて評価を行う．これをテストデータを替えながら k 回繰り返し，評価値の平均を求める．

データを分割する際に，元データの並び順に何らかの基準が存在する場合（予測すべきラベルでソートされているなど），訓練データとテストデータに偏りが生じてしまい，交差検証の結果に影響を及ぼす．また，入力データを一つずつ逐次的に読み込み，データが与えられるごとにパラメータを更新していくオンライン学習アルゴリズムも，データの並び順の影響を受ける．

以上のような影響を避けるために，予めデータをシャッフルするか否かを分析者が指定できるようにした．初期値はシャッフル有とした．

4.2.3 機械学習アルゴリズムの選択

データ分析に用いる機械学習アルゴリズムの選択は，分類か回帰かという分析タスク以外に，データサイズも考慮して行う必要がある．機械学習アルゴリズムによっては，データサイズが大きいときの計算コストが非常に大きく，当該アルゴリズムの使用が現実的でない場合もある．

scikit-learn のチュートリアル [*1] には，分析タスク，データ数に応じたアルゴリズムの選択指針がチャート形式でまとめられている．しかし，このチャートではデータのサンプル数しか考慮しておらず，データの項目の数が考慮されていない．機械学習アルゴリズムの計算コストはデータのサンプル数だけでなく項目数にも依存する場合があるため，本論文では，このチャートを参考にしながら，項目数も考慮した独自のアルゴリズム選択ルー

[*1] http://scikit-learn.org/stable/tutorial/ (2021/11/13 access)

ルを作成した．作成したアルゴリズム選択ルールを図 4.4 に示す．

アルゴリズム選択基準となるデータサイズの閾値により，分析に要する時間が大きく異なる．MALSS の利用対象者がデータ分析未習熟者であることを考慮すると，分析所要時間は長すぎないことが望ましい．そこで，処理時間が最も長くなる図 4.4 中の条件 (A) のときに，CPU 動作周波数 1.8GHz，4 スレッド，4GB RAM の PC で，約 30 分で分析が終了するように閾値を設定した．

機械学習アルゴリズムは分析者が追加，削除を行うことが可能である．

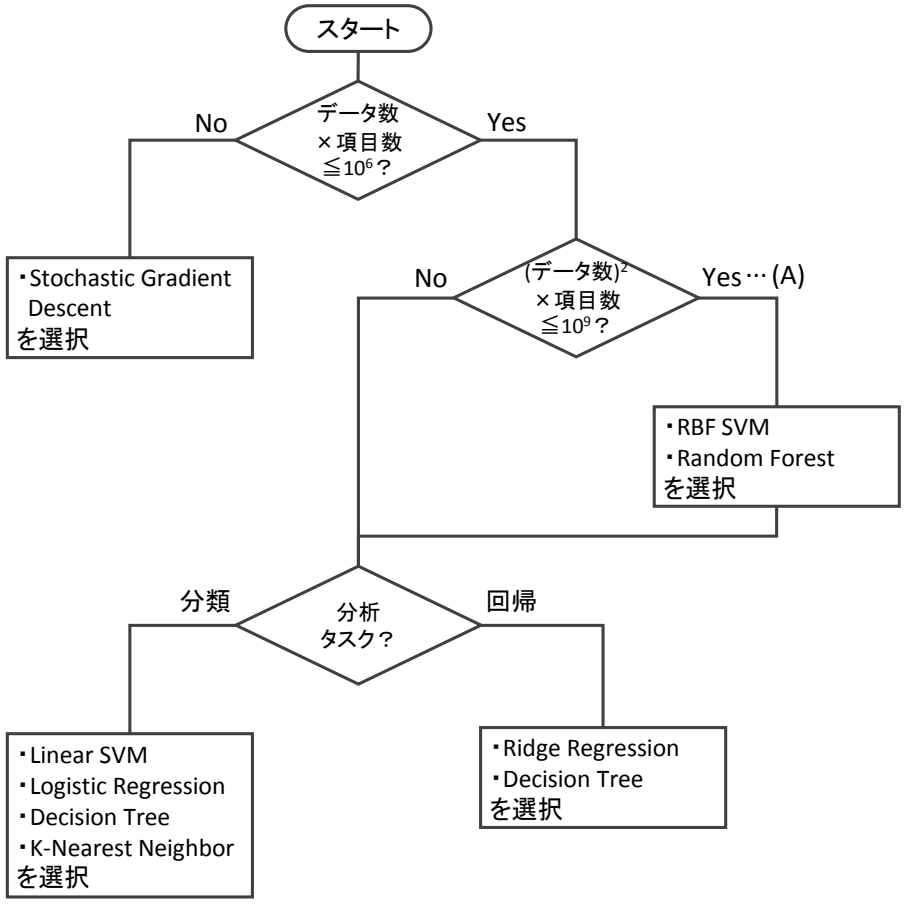


図 4.4　アルゴリズム選択ルール

4.2.4　分析

分析プロセスは，3.3.3 節で述べた，MALSS が支援するデータ分析プロセスのうち，モデリングのプロセスに相当する．分析プロセスは以下の 3 つの機能から構成される

1. ハイパーパラメータ調整
2. 交差検証
3. 学習曲線の生成

以下において，これらの機能について述べる．

ハイパーパラメータ調整

多くの機械学習アルゴリズムは性能を制御するためのパラメータ（ハイパーパラメータ）をもっており，分析データに応じて適切なハイパーパラメータを求める必要がある．ハイパーパラメータ調整には一般的にグリッドサーチが用いられる [28].

グリッドサーチはハイパーパラメータを適当な範囲で変化させ，ハイパーパラメータの値の組み合わせを格子点に見立て，格子点ごとにモデリングを行い性能を評価する．そして，最も性能の良い格子点のハイパーパラメータを採用する手法である．

性能評価の指標は様々なものがあり [28]，MALSS では分析者が指定することが可能である．初期値では，回帰タスクでは平均二乗誤差を用いる．分類タスクでは，精度 (accuracy) は，ラベルに偏りがある場合にモデルの性能を適切に評価することが困難であるため，F 値を用いる．

ハイパーパラメータの調整範囲，刻み幅は学習アルゴリズムごとに適切な値を初期値として設定したが，分析者が指定することも可能である．

交差検証

4.2.2 項で触れたように，過学習を防ぐために，交差検証による評価を行う．交差検証には K-fold cross validation の他に，分類タスクにおいて，ラベルごとのデータの比率を保ったまま k 等分する Stratified k-fold cross validation や，テストデータを 1 サンプルとし，残りを全て訓練データとしてデータ数分評価を繰り返す Leave-One-Out cross validation など，様々な手法が存在する [25].

MALSS では交差検証手法を分析者が指定することが可能である．初期値として，回帰タスクでは 5-fold cross validation を用いるが，分類タスクでは与えられたデータのラベルの比率に偏りがある場合に備え，Stratified 5-fold cross validation を用いる．

学習曲線の生成

横軸に分析データのサンプル数をとり，サンプル数を変化させた時の交差検証により求めた訓練誤差と汎化誤差を縦軸にプロットしたものを学習曲線とよぶ [25]．学習曲線はよ

り良い分析結果を得るために，より多くのデータを集めるべきなのか，より複雑なモデリングが可能な機械学習アルゴリズムを適用すべきなのか，などの指針を得るために有用である．MALSS では，分析レポートで利用するために，選択した機械学習アルゴリズムごとに学習曲線を作成する．

4.2.5 サンプルコードの出力

機械学習によるデータ分析の目的は，分析結果（モデル）を元に未知のデータから予測を行うことである．MALSS が支援対象としている分析者である Data Creative，および Data Developer は予測を行うためのシステムを自ら開発する必要がある．

そこで MALSS は，未知のデータに対し予測を行うためのプログラムのサンプルコードを自動で生成する．分析者はこのサンプルコードを見ることで予測プログラムの作成方法を習得するとともに，サンプルコードに必要な変更を行うことで，実際に開発するシステムへ実装することが可能となる．

4.2.6 MALSS の使用方法

4.1 節で示した例と同じ回帰分析を，MALSS を用いて行う場合を例に，MALSS の使用方法を説明する．MALSS を用いて分析を行う場合の Python コードの例を図 4.5 に示す．コードの行数は scikit-learn を用いた場合と同じであり，異なるのは，4 行目のライブラリインポート部と，11 行目の MALSS 利用部分のみである．MALSS はインスタンスの生成時に分析タスクを設定し (図の例では regression)，scikit-learn と同様に fit メソッドにデータを渡すだけで，分析タスク，データに応じた適切な手順で分析を自動で行うことができる．predict メソッドは，分析の結果最も性能の良いアルゴリズムを用いて予測を行う．

MALSS を利用することによって，テストデータの予測結果である汎化誤差が約 0.42 から約 0.23 へと改善しており，図 4.6 を見ても，真の値に近い適切なモデリングができていることが分かる．このように，MALSS を利用することで，容易に質の高い分析を行うことが可能となる．

```
from numpy import random, cos, pi, sort, newaxis
from sklearn.metrics import mean_squared_error as mse
import matplotlib.pyplot as plt
from malss import MALSS

random.seed(0)
true_fun = lambda X: cos(1.5 * pi * X)
X_train = sort(random.rand(30))
y_train = true_fun(X_train) + random.randn(30) * 0.4

clf = MALSS('regression').fit(X_train[:, newaxis], y_train)
print u'訓練誤差:', # 0.18385399836
print mse(y_train, clf.predict(X_train[:, newaxis]))

X_test = sort(random.rand(30))
y_test = true_fun(X_test) + random.randn(30) * 0.4
print u'汎化誤差:', # 0.229647776708
print mse(y_test, clf.predict(X_test[:, newaxis]))
```

図 4.5　MALSS を用いて回帰分析する Python コード例

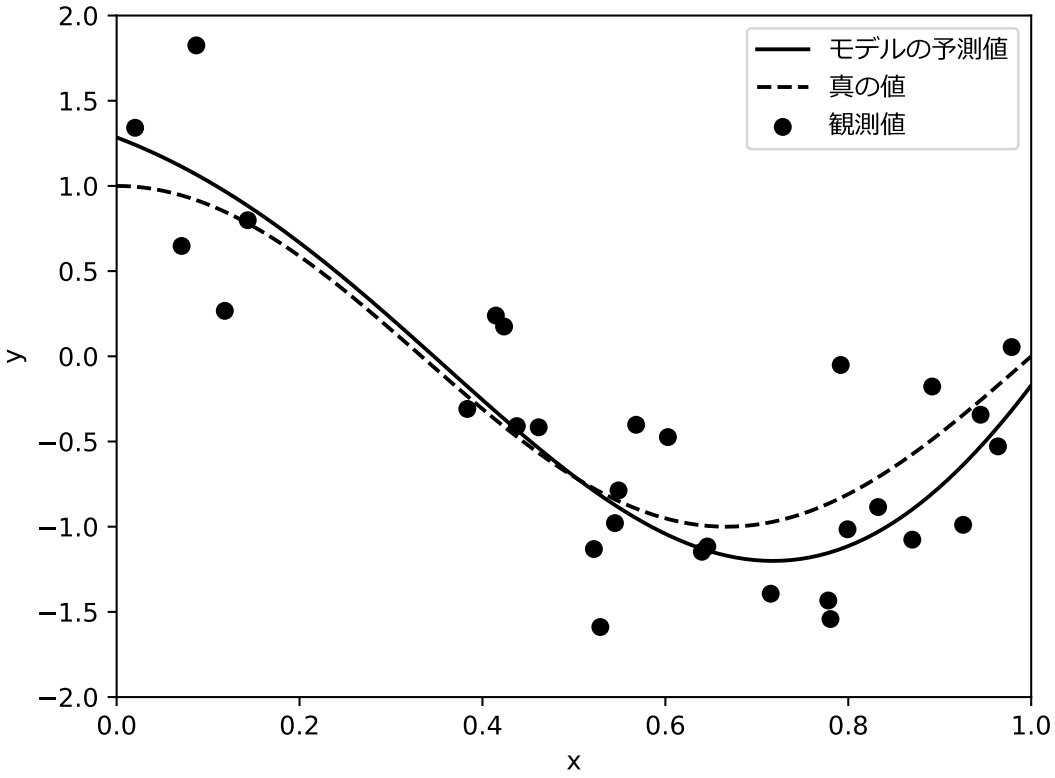


図 4.6　真の値と目的変数とモデルの予測値（MALSS 利用）

4.3　分析レポートによる知識習得支援

分析を自動化しただけでは，機械学習アルゴリズムだけでなく，分析プロセスもブラックボックス化してしまい，分析者の知識習得につながらない．そこで，分析レポートを作成することで，分析結果を明示するとともに，分析プロセスにおける留意点等も併記することで，分析者の知識習得を可能とする．

具体的には，分析レポートは HTML ファイルとして出力され，以下の項目から構成される．

1. 複数アルゴリズムの性能比較
2. 訓練データ概要
3. 各アルゴリズムの分析結果詳細

分析レポートの一例を図 4.7 から図 4.10 に示し，以下において，これらの項目について述べる．

複数アルゴリズムの性能比較

初めに，与えられた訓練データに応じて選択した複数の機械学習アルゴリズムについて，交差検証のスコアを表形式で提示する（図 4.7）．分析者はここで，分析を行った複数のアルゴリズムと，それぞれの交差検証のスコアを確認することができる．

また，分析者の知識習得を目的として，交差検証，および評価基準に関する説明を付記した．さらに，レポート中の専門用語には当該用語を解説している Web サイトへのリンクを付けた．

分析結果

アルゴリズム	交差検証のスコア（f1）
Support Vector Machine (RBF Kernel)	0.849
Random Forest	0.833
Logistic Regression	0.845

▼ 解説（クリックして中身を確認してください）

※交差検証：

- 機械学習では，モデル学習に使うデータ(訓練データ)に含まれない未知のデータに対して良い結果を出す能力，**汎化能力**が重要となります．
- モデルの学習と評価に同じデータを使うと訓練データに過度に適応(**過学習**)してしまい，訓練スコア(training score)は良く(訓練誤差(training error)は小さく)なりますが，汎化能力が低下してしまいます．
- 過学習を防ぐためには**交差検証**を行い，交差検証スコア(cross-validation score)(交差検証誤差(cross-validation error))により，汎化能力を評価します．
  代表的な交差検証法であるK-fold cross validationでは，まずデータセットをK個（default: 5）に分割します．そして，そのうちの1つをテスト用とし，残るK−1個でモデルを学習します．
  交差検証はK個に分割されたデータそれぞれをテストデータとしてK回検証を行い，得られた結果を平均して1つのスコアを得ます．
- 交差検証は様々な手法が提案されているので，目的に応じて適切な手法を選択してください．
  （デフォルトでは，回帰（regression）タスクでは5-fold cross validationが，分類（classification）タスクではStratified 5-fold cross validationが選択されています．）

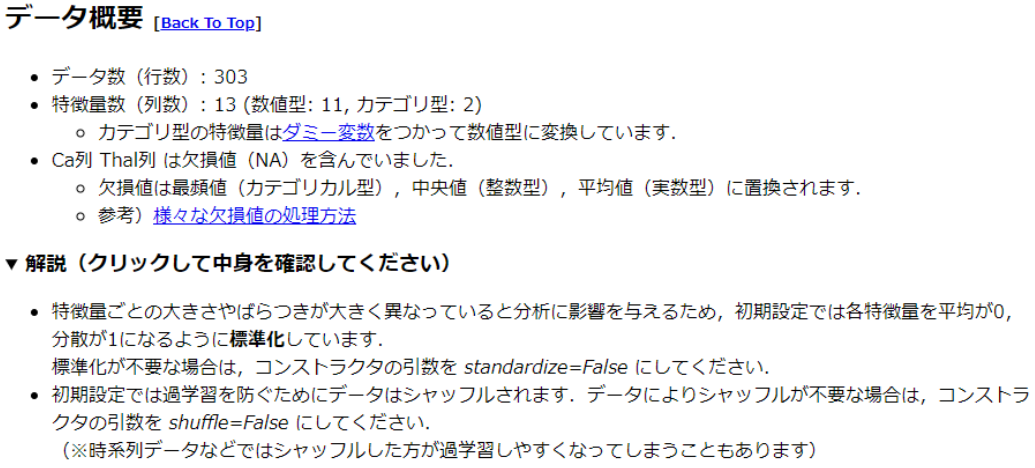
※スコア：

- 汎化能力の評価指標(スコア)には様々なものがあります(scoringオプション)．
- 精度(accuracy)は分類モデルを評価する代表的なスコアの一つですが，ラベルに偏りがあり，1%のデータのみが陽性の場合，常に陰性と予測するモデルの精度は99%ですが，このモデルは実用的ではありません．
- デフォルトでは，回帰（regression）タスクでは平均二乗誤差(mean squared error)(小さいほど良く0が最小)が，分類（classification）タスクではF値(f1 score)(大きいほどよく1が最大)が選択されています．

図 4.7　分析レポートの例（複数アルゴリズムの性能比較）

訓練データ概要

次に，データの概要を提示する（図 4.8）．データのサイズ（サンプル数，説明変数の数）の他に，欠損値の有無，欠損値がある場合は補間方法，データがカテゴリ変数を含む場合はダミー変数に変換したことを説明する．



データ概要 [Back To Top]

- データ数（行数）：303
- 特徴量数（列数）：13 (数値型: 11, カテゴリ型: 2)
  - カテゴリ型の特徴量はダミー変数をつかって数値型に変換しています．
- Ca列 Thal列 は欠損値（NA）を含んでいました．
  - 欠損値は最頻値（カテゴリカル型），中央値（整数型），平均値（実数型）に置換されます．
  - 参考）様々な欠損値の処理方法

▼ 解説（クリックして中身を確認してください）

- 特徴量ごとの大きさやばらつきが大きく異なっていると分析に影響を与えるため，初期設定では各特徴量を平均が0，分散が1になるように**標準化**しています．
  標準化が不要な場合は，コンストラクタの引数を *standardize=False* にしてください．
- 初期設定では過学習を防ぐためにデータはシャッフルされます．データによりシャッフルが不要な場合は，コンストラクタの引数を *shuffle=False* にしてください．
  （※時系列データなどではシャッフルした方が過学習しやすくなってしまうこともあります）

図 4.8　分析レポートの例（訓練データ概要）

各アルゴリズムの分析結果詳細

以降，選択したアルゴリズムごとに，分析結果詳細を提示する．図 4.9 と図 4.10 では，Support Vector Machine (RBF Kernel) の場合を例に説明する．

初めに，グリッドサーチによるハイパーパラメータの調整結果を表形式で示す（図 4.9）．最適なハイパーパラメータがグリッドの境界にある場合にはグリッドを変更すべき点，最適値の近傍で，グリッドの範囲を狭く，刻み幅を小さくした，詳細なグリッドサーチが有効である点を説明する．

つぎに，分析タスクが分類である場合は，ラベルごとの再現率・適合率・F 値・支持度を示す（図 4.9）．

最後に学習曲線を示す（図 4.10）．学習曲線の説明を行い，学習曲線から，モデルが訓練データに過剰に適応している状態（ハイバリアンス）なのか，モデルの性能が不十分な状態（ハイバイアス）なのかを判断する基準を提示するとともに，各状態に応じた対応策を示す．

Support Vector Machine (RBF Kernel) [Back To Top]

グリッドサーチによるパラメータチューニング結果

kernel	C	gamma	スコア（f1）	偏差
rbf	1	0.01	0.831	0.045
rbf	1	0.1	0.812	0.042
rbf	10	0.01	0.849	0.036
rbf	10	0.1	0.781	0.039

▼ 解説（クリックして中身を確認してください）

- 最適なパラメータがグリッドの端の値である場合，グリッドのレンジを変更してください．
- 最適なパラメータ付近でより細かいグリッドでパラメータチューニングを行うとさらに効果的です．

分類結果

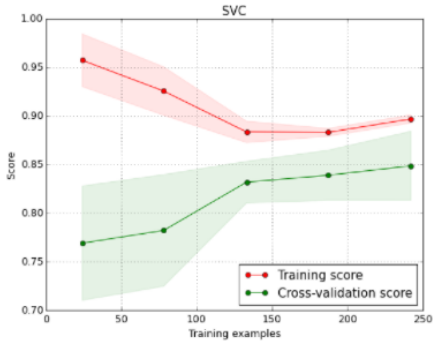
	precision	recall	f1-score	support
No	0.87	0.93	0.90	164
Yes	0.91	0.84	0.88	139
avg / total	0.89	0.89	0.89	303

※precision（適合率），recall（再現率），f1-score（F1値），support（正解ラベルのデータ数）

- 注）上記のスコアはクローズ評価（学習と評価が同データ）なので，モデルが過学習しているかもしれません．

図 4.9　分析レポートの例（アルゴリズムごとの分析結果詳細）

学習曲線（Learning curve）



▼ 解説（クリックして中身を確認してください）

学習曲線（Learning curve）

- 学習曲線はデータサイズを変えた時の訓練データでのスコア，交差検証のスコアをプロットしたものです．
- 学習曲線が以下のような場合，モデルは**ハイバリアンス**（オーバーフィッティング（過学習））であると言えます：
  - 学習データ増加に伴う交差検証のスコアの改善が飽和していない（改善し続けている）．
  - 訓練データのスコアと交差検証のスコアの差が大きい．
- 学習曲線が以下のような場合，モデルは**ハイバイアス**（アンダーフィッティング）であると言えます：
  - 訓練データのスコアでさえも悪い(誤差が大きい)．
  - 訓練データのスコアと交差検証のスコアの差が小さい．

ハイバリアンス(High variance)への対策：

- 特徴量選択や次元削減により特徴量の数を減らす．
- データ量を増やす．

ハイバイアス(High bias)への対策：

- 特徴量を増やす．
- より複雑なモデル（アルゴリズム）を利用する．
- データ量が多すぎて計算コストの問題から複雑なモデルが利用できない場合，データ量の削減が有効な場合があります．

※バイアス，バリアンス

図 4.10　分析レポートの例（学習曲線）

4.4　評価

本章では以下の 2 点を研究課題とする．

RQ1　機械学習によるデータ分析の未習熟者が，MALSS のみを利用して適切な分析を行うことができるか？

RQ2　MALSS のみを利用した分析を通じて，機械学習に関する知識を習得することができるか？

上記の研究課題に対する MALSS の有効性を評価するために，模擬データ分析実験，お

よび知識確認テストを行った.

4.4.1 分析の質の評価

MALSS によるデータ分析の質向上の有効性を評価するために，模擬データ分析実験を行った.

実験は機械学習によるデータ分析の未経験者，および初学者 10 名により行い，実験協力者を A グループ（5 名）と B グループ（5 名）に分けた．A グループは MALSS を利用してデータ分析を行い，B グループは既存のライブラリ等を利用して分析を行う．A グループは MALSS 以外の機械学習ライブラリの利用を禁止し，分析の際に MALSS の使用説明書，MALSS が出力する分析レポート，および分析レポートのリンク先以外からの情報取得も禁止した．B グループは，推奨ライブラリとして scikit-learn を提示したが，特に利用ライブラリの制限は設けなかった．また，機械学習によるデータ分析を行うための知識習得に有用な参考資料として slideshare [*2]上の資料や scikit-learn のチュートリアルなどを提示し，その他の資料の参照にも制限を設けなかった．今回，1 名だけ機械学習によるデータ分析の非未経験者 (初学者) がいたため，その実験協力者は B グループに割り当てた.

分析データは UCI Machine Learning Repository [*3]で公開されている Abalone Data Set を用いた．このデータセットはアワビの年齢を複数の身体特徴から推定する回帰タスクに用いることができる．実験ではデータの 80% を訓練用とし，20% をテスト用とした．分析者は訓練用データを用いて分析を行い，テスト用データに対して予測を行う．テスト用データは年齢データを削除しており，分析者はテスト用データに対する予測精度をその場で確認することはできない．予測性能の評価は平均二乗誤差 (mean squared error) により行った.

模擬データ分析実験の結果を図 4.11 に示す．今回，分析時間に 2 時間という制限を設けたが，A，B グループとも 1 名ずつ制限時間内に分析を終えることができなかったため除いている．棒グラフの値は各グループ 4 名の平均二乗誤差の平均値を示しており，エラーバーは平均二乗誤差の最大値と最小値を表している．A グループと B グループの平均値の差について，平均値に差がないことを帰無仮説とし，コルモゴロフ・スミルノフ検定（両側検定）を行ったところ，p 値は 0.009 であった．なお，コルモゴロフ・スミルノフ検定の実施にはブートストラップ法を用いる R 言語の Matching パッケージを利用し

[*2] http://www.slideshare.net/ (2021/11/13 access)
[*3] http://archive.ics.uci.edu/ml (2021/11/13 access)

た [29].

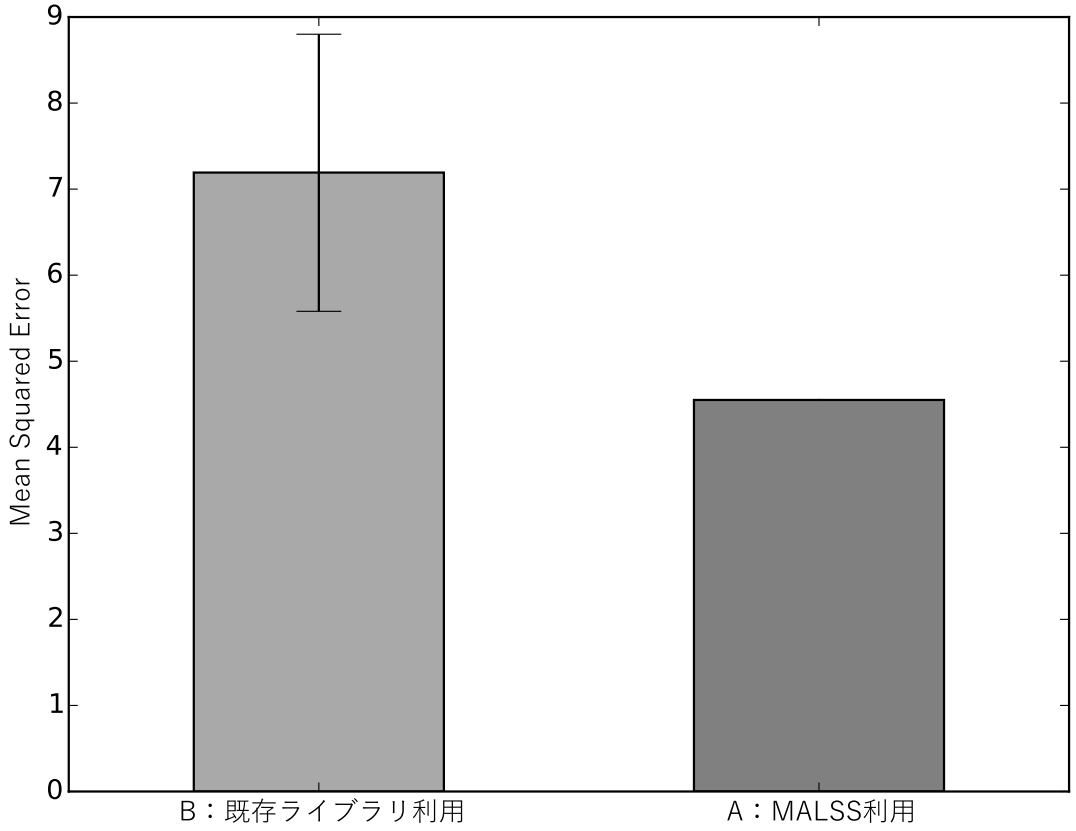


図 4.11　模擬データ分析実験結果

4.4.2　学習効果の評価

学習支援手法の評価としては，介入の効果を対象群と実験群の比較により定量化することが一般的に行われている [30]．本研究においても，MALSS による機械学習に関する知識習得の効果を確認するために，知識確認テストを行い，学習支援の効果を定量化した．

4.4.1 項で述べた模擬データ分析実験を行う前に，実験協力者に対し，機械学習によるデータ分析に関する知識の有無を問うテストを行った．テストの設問は A グループ，B グループ共通で，設問数は 14 問．MALSS が支援対象とする分析者が最低限習得すべき知識を問うように設計した．問題は，A グループは MALSS が作成する分析レポートの情報をもとに，B グループは有用な参考資料として提示した資料に記載されている情報をもとに正答できるようになっている．テスト回答にあたっては資料等を参照すること，お

よび勘で回答することを禁止した．

さらに，模擬データ分析実験終了後に再び同じテストを行い，スコアの変化から学習効果を測定した．実験協力者には模擬データ分析実験後にもテストを行うことは伝えていない．

知識確認テストの設問を図 4.12 に示す．また，知識確認テストの結果を図 4.13，および図 4.14 に示す．図 4.13 は，横軸が模擬データ分析実験前のテスト正答率を，縦軸が分析実験後の正答率を示している．ひとつのプロットがひとりの実験協力者のデータを表しており，分析前後の正答率が共に等しいサンプルは僅かにずらしてプロットしている．プロットが点線よりも上に位置していれば，模擬データ分析実験の前後でテストのスコアが向上したことを表している．図 4.14 は，正答率の向上値 (実験後の正答率から実験前の正答率を引いた値) を，グループごとに箱ひげ図で示したものである．図 4.13 において，実験協力者全体の模擬データ分析実験前後のテスト正答率の差について，符号検定による対応のある 2 つの母平均の差の検定（両側検定）を行ったところ，p 値は 0.00195 であった．また，図 4.14 において，A グループと B グループの正答率の向上値の差について，コルモゴロフ・スミルノフ検定（両側検定）を行ったところ，p 値は 0.357 であった．

機械学習によるデータ分析　知識確認テスト

機械学習に関する次の文章を読み，【ア】～【セ】に当てはまる語句を選択肢 1～30 から選んでください．
同じ選択肢を複数回選んでも構いません．
回答中は教科書や Web 等の資料を見ないでください．
分からないときは勘で埋めたりせず，30. 分からない を選んでください．

機械学習とは，「あるタスクを遂行するためのモデルを，与えられたデータ（学習データ）から構築すること」と定義されます．

機械学習は学習データの種類によって分類され，学習データに正解（こういう結果を出してほしいという情報）がついている場合の学習を，【ア　　】，正解がついていない場合の学習を【イ　　】とよびます．

さらに，【ア　　】において，正解が数値の場合を【ウ　　】問題とよび，正解がラベルの場合を【エ　　】問題とよびます．

機械学習では，学習データに含まれない未知のデータに対して良い結果を出す能力，
【オ　　】能力が重要となります．
学習データに過度に適合してしまい，【オ　　】能力が低下した状態を【カ　　】といいます．
【カ　　】への対策として，【キ　　】や【ク　　】検証などが挙げられます．

学習データの量に対するモデルの性能（学習データに対する性能と未知データに対する性能）をプロットしたものを【ケ　　】といいます．
【ケ　　】をみることで，モデルや特徴量が単純過ぎて誤差が大きい状態（【コ　　】）なのか，モデルが学習データに過度に適合している状態（【サ　　】）なのかなどを判断することができ，適切な対応をとることが可能となります．
モデルが【サ　　】な場合は，データ量を【シ　　】ことや，【ス　　】などが有効で，
モデルが【コ　　】な場合は，特徴量を【セ　　】ことや，より複雑なモデルを利用することなどが有効です．

選択肢

1. ROC 曲線，　2. 意思決定，　3. AUC，　4. オッカムの剃刀，　5. 回帰，
6. 過学習，　7. 学習曲線，　8. 機械学習，　9. 強化学習，　10. 教師あり学習，
11. 教師なし学習，　12. クラスタリング，　13. 交差，　14. 混同行列，
15. 次元削減，　16. 次元の呪い，　17. 正則化，　18. 説明変数，
19. データマイニング，　20. 特徴量選択，　21. ハイバイアス，
22. ハイバリアンス，　23. 汎化，　24. フィッティングカーブ，　25. 増やす，
26. 分類（識別），　27. ベイズの法則，　28. 減らす，　29. 目的変数，
30. 分からない

図 4.12　知識確認テスト

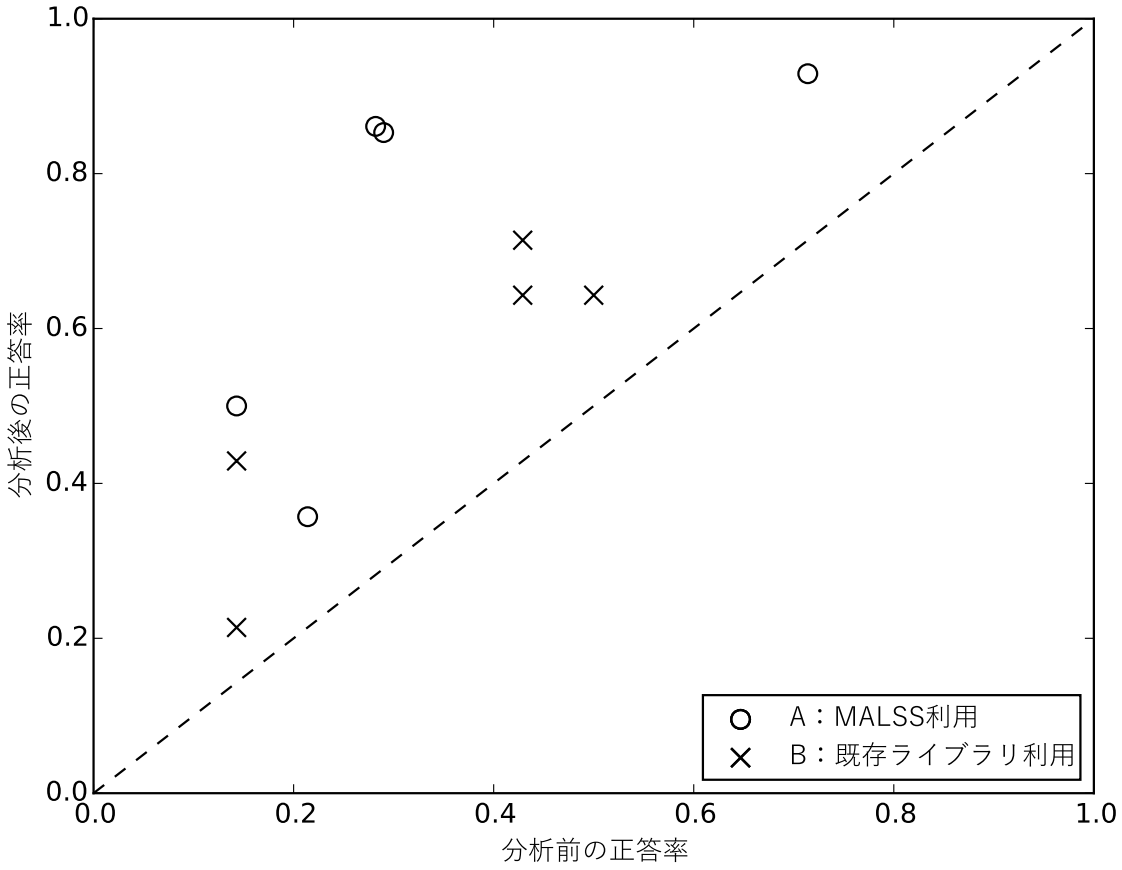


図 4.13　知識確認テスト結果

図 4.14　分析実験前後の知識確認テスト正答率差分

4.5　考察

4.5.1　研究課題に対する考察

RQ1　機械学習によるデータ分析の未習熟者が，MALSS のみを利用して適切な分析を行うことができるか？

模擬データ分析実験の結果（図 4.11），A グループは全員の平均二乗誤差が同一であった．これは，MALSS が一連の分析プロセスを自動化しているためであり，分析者の習熟度に依らず，均質な分析が可能であることを示している．

A グループの平均二乗誤差は，B グループの平均二乗誤差最小値よりも有意（有意水準 5%）に小さく，MALSS のみを利用して分析を行った場合に，機械学習によるデータ分析手法を学びながら既存ライブラリを利用して分析を行った場合よりも質の高い分析を安定して行うことが可能であると確認できた．

RQ2　MALSS のみを利用した分析を通じて，機械学習に関する知識を習得することができるか？

図 4.13 から，知識確認テストの結果，両グループとも分析の前後で正答率の向上が有意（有意水準 5%）に認められ，機械学習によるデータ分析に関する知識を習得できていることが分かる．図 4.14 を見ると，正答率の向上値は A グループの方が大きい傾向が認められるものの有意水準 5% で有意差は認められなかった．しかし両グループに差が認められないことから，MALSS のみを利用してデータ分析を行った場合に，機械学習によるデータ分析手法を学びながらデータ分析を行った場合と少なくとも同等の知識習得が可能であることが確認できた．今回，B グループはあらかじめ筆者が選定した参考資料を参照して分析を行っている．実際の分析においては分析者が自ら参照すべき資料を探す必要があるが，これは知識や経験の不足した分析者にとって容易ではなく，時間がかかるだけでなく不十分な知識を参照してしまう恐れもある（例えばあるアルゴリズムのライブラリの使用方法のみを説明した情報を参照し，過学習を防ぐための交差検証の必要性を理解せず分析を行ってしまうなど）．今回 MALSS のみを利用した A グループの知識習得度合いが，あらかじめ用意された資料を参照した B グループと同等以上であることは，実際の分析において MALSS を用いて分析を行うことにより効率的に知識習得可能であることを示している．

MALSS を利用した A グループの，模擬データ分析実験後の知識確認テストの正答率平均は 70% であった．4.4.2 項で述べた通り，知識確認テストの設問は MALSS が支援対象とする分析者が身に着けておくべき基本的な設問であり，全問正答できるようになることが望ましい．本評価は，知識や経験が不足している分析者が初めて MALSS を利用してデータ分析を行った際の知識習得の程度を評価したものである．実際の実務において，MALSS を繰り返し利用し分析業務を遂行していく中で，知識習得の度合いが向上していくと考えられる．継続利用時の知識習得度合いの評価は今後の課題である．

4.5.2　妥当性への脅威

本項では，本研究の結果の妥当性に影響を及ぼす恐れのある事項について述べる．

3.3.4 項で述べたように，MALSS はプログラミングスキルの高い分析者を支援対象として想定している．しかし，本評価実験の実験協力者は全員が Python の初学者であった．既存の機械学習ライブラリを利用し分析を行った B グループは，自動で分析を行う MALSS を利用した A グループと比較して，プログラムの記述量が多くなると考えられ

る．このため，Python 初学者であるという条件は B グループに対し不利に働いている可能性がある．今後は，プログラミングスキルと分析の質との関係についても検討を行いたい．

また，本評価実験は，実験協力者が各グループ 5 名と少人数であるため，サンプル数が増えた時に評価結果が異なる可能性がある．模擬データ分析実験における A グループと B グループの差は，有意水準 5% で有意差が認められたが，知識確認テストについては，A，B グループとも模擬データ分析実験前後でスコアが向上し，スコア変化分は MALSS を利用した A グループの方が大きいものの，両グループの間に差がないという帰無仮説を有意水準 5% で棄却することはできなかった．今後，より大規模な評価実験を行うことが求められる．

4.6 第 4 章におけるまとめ

本章では，知識・経験が十分でない分析者のデータ分析作業支援を目的として，自動化による分析支援，および分析レポート作成による知識習得支援を行う機械学習支援システム MALSS を提案した．模擬データ分析実験，および知識確認テストにより，MALSS を利用したデータ分析により，機械学習によるデータ分析手法を学びながら分析した場合よりも質の高い分析を行いながら，機械学習によるデータ分析に必要な知識を習得することが可能であることを示し，MALSS の有効性を確認した．

今後の課題としては，教師なし学習など対象分析タスクの拡大や，分析結果に応じた動的な分析支援などが挙げられる．

第 5 章

グラフィカルユーザーインタフェースを用いたインタラクティブな分析支援

5.1　分析レポートによる支援における課題

分析レポートを用いた知識習得支援の課題として，サブプロセスの実行結果に応じて分析の実施内容を変更することが困難であることが挙げられる．図 3.1 や図 3.2 に示すように，機械学習を用いたデータ分析の各プロセスは繰り返し実施される．図中では全てのプロセスを結ぶ一つのループとして描かれているが，実際には一部のプロセス，あるいはあるプロセス内のサブプロセスのみを繰り返すことも多い．分析終了後に分析レポートを生成する MALSS では分析プロセス全体を自動化しているため，MALSS を複数回実行することでこれらのプロセス全体を繰り返し実施することは可能であったが，一部のサブプロセスのみの繰り返し処理を支援することができない．

3.3.3 で述べた特徴量エンジニアリングとプロトタイピングのプロセスは，さらに次のようなプロセスから構成される．

- 特徴量エンジニアリング
  - データの標準化
  - 欠損値の処理
  - カテゴリ変数の処理
  - 次元削減

- プロトタイピング
  - アルゴリズム選択
  - モデルの学習（訓練）
  - （交差）検証
  - 評価
  - ハイパーパラメータ調整

このなかで，カテゴリ変数の処理，次元削減，ハイパーパラメータ調整のプロセスにおいて以下のような課題が存在する．

カテゴリ変数の処理

MALSS は，質的変数であるカテゴリ変数を，ダミー変数を用いて量的変数に変換する．しかし，カテゴリ変数は数値データとして表されることもあり（例：国名を数字 3 桁の国名コードで表す）．このようなカテゴリ変数を数値データとして扱うとモデルの性能に悪影響を及ぼすことがあるが，これを自動的に判定し適切な変換を行うことは容易ではない．従来の MALSS ではこのような数値データとして表されるカテゴリ変数をカテゴリ変数として扱うことができず，分析者自身が MALSS を利用する前に変換しておく必要があった．

ハイパーパラメータ調整

MALSS は，グリッドサーチによるハイパーパラメータ調整をサポートしている．ハイパーパラメータ調整では，各ハイパーパラメータの探索範囲と，探索範囲内の探索点数の決定が課題となる．探索範囲と探索点数を十分広く・多くすると，ハイパーパラメータの組合せの数が膨大となる．そこで実用においては，モデルの学習・評価結果を見ながら分析者が探索範囲と探索点数を調整していく．従来の MALSS でも探索範囲と探索点数を変更することは可能であったが，データの準備とモデリングのプロセスが完了したのちに，分析レポートを読んでその必要性を判断し，探索範囲と探索点数の変更が必要な場合には，再度，特徴量エンジニアリングとプロトタイピングのプロセスを最初から実行する必要があった．

次元削減

モデルの学習に用いる特徴量の数を削減する次元削減は，モデルが学習データに過度に適合している過学習状態のときに効果的であり，モデルの学習・評価結果に応じた適切な

実行要否の判断が重要である．既存の AutoML ツールのように，モデルの状態によらず特徴量選択を試行し，モデルの評価結果によりその妥当性を評価するアプローチも合理的である．しかし 3.3.5 項で述べた通り，MALSS は分析者の基本的な分析に関する知識の習得を支援することを目的としているため，モデルの学習・評価結果を考慮することなく完全に自動化してしまうことは望ましくない．従来の MALSS では次元削減の機能をサポートしておらず，分析者が分析レポートを読んで次元削減が必要と判断した場合，MALSS を実行する前の訓練データに対し適切に次元削減の処理を行う必要があった．

5.2 グラフィカルユーザーインタフェースを用いたインタラクティブな分析支援

5.2.1 要件

本章では，従来の MALSS の機能を拡張することで，前節で述べた，サブプロセスの実行結果に応じて実施内容を変更することが困難であるという課題を解決することを目的とする．そのために MALSS が実装すべき機能要件，および実装する機能以外の要件，すなわち非機能要件は以下の通りである．機能要件 1，2 により特徴量エンジニアリングとプロトタイピングプロセスの全体および一部分の繰り返し試行を支援することが可能となる．

機能要件 1　サブプロセスごとに分析を実行可能であること．

機能要件 2　サブプロセスの実行結果に応じて分析者の入力を受け付け，プロセス実施内容を変更可能であること．

非機能要件 1　従来手法よりも適切に分析者のデータ分析を支援可能であること．

非機能要件 2　従来手法である分析後の分析レポート提示と同等に，機械学習に関する知識習得を支援可能であること．

次項より上記要件を満たし課題を解決するための機能設計について述べる．

5.2.2 機能設計

グラフィカルユーザーインタフェース

分析者の入力の受け付け（機能要件 2）を可能とするために提案手法ではグラフィカルユーザインタフェース（GUI）を採用する．従来の MALSS が提供するコマンドライン

インタフェース（CLI）でも分析者の入力を受け付けることは可能であるが，GUI を採用することでより簡便な入力の受け付けが可能となる．GUI はクロスプラットフォームアプリケーションフレームワーク Qt の Python バインディングである PyQt [*1]を用いて開発した．Python のネイティブアプリケーションにすることで，従来の MALSS 同様，Python だけで知識習得と分析遂行支援を行うことが可能となる．

画面レイアウト

提案手法では，図 5.1 に示すように，GUI 左側にコンテンツビューを，右側にメインビューを設け，サブプロセスごとの分析を実行可能とする（機能要件 1）．分析のサブプロセスごとにコンテンツを分け，当該コンテンツに関連する情報のみをメインビューに表示する．サブプロセスが完了するとコンテンツビューに新たなコンテンツを追加し，メインビューは新たに追加されたコンテンツに関連する情報に変更される．

図 5.1　カテゴリ変数の処理（コンテンツビューの「データの確認」に対応）．

[*1] https://www.riverbankcomputing.com/software/pyqt/ (2022/2/10 access)

分析課題に対する設計

■カテゴリ変数の処理　図 5.1 は「データの確認」というコンテンツでカテゴリ変数の処理を行う GUI 画面である．「データの確認」のひとつ前のコンテンツ「入力データ」で，特徴量は量的変数とカテゴリ変数に大別されること，カテゴリ変数はそのままでは分析に用いることが困難であり，MALSS ではダミー変数を用いてカテゴリ変数を量的変数に変換することを説明する．「データの確認」コンテンツでは，メインビュー上部で訓練データの一部を分析者に提示し，中央部でカテゴリ変数の中には数値データとして表されているものもあることを教示する．そしてメインビュー下部では，特徴量ごとにカテゴリ変数か量的変数かを分析者に選択させることで，入力された訓練データに応じて適切にカテゴリ変数の処理を行うことを可能とする．

■ハイパーパラメータ調整　ハイパーパラメータの調整は図 5.2 に示すように，「結果の確認」のコンテンツに対応するメインビューで実施する．ハイパーパラメータの組み合せごとにモデルの評価スコアを提示し，最もスコアの高いモデルのハイパーパラメータが，ハイパーパラメータ試行範囲の端点であった場合に，ハイパーパラメータの調整が必要であることを示す．分析者は結果を確認したうえで，必要に応じてハイパーパラメータの調整を行い，「Re-analyze」ボタンを押下することで再度分析を実行することができる．図 5.2 の例では，分析者は決定木アルゴリズムの木の深さ（*max_depth*）というハイパーパラメータを，いくつからいくつまで何分割にするかを調整することができる．

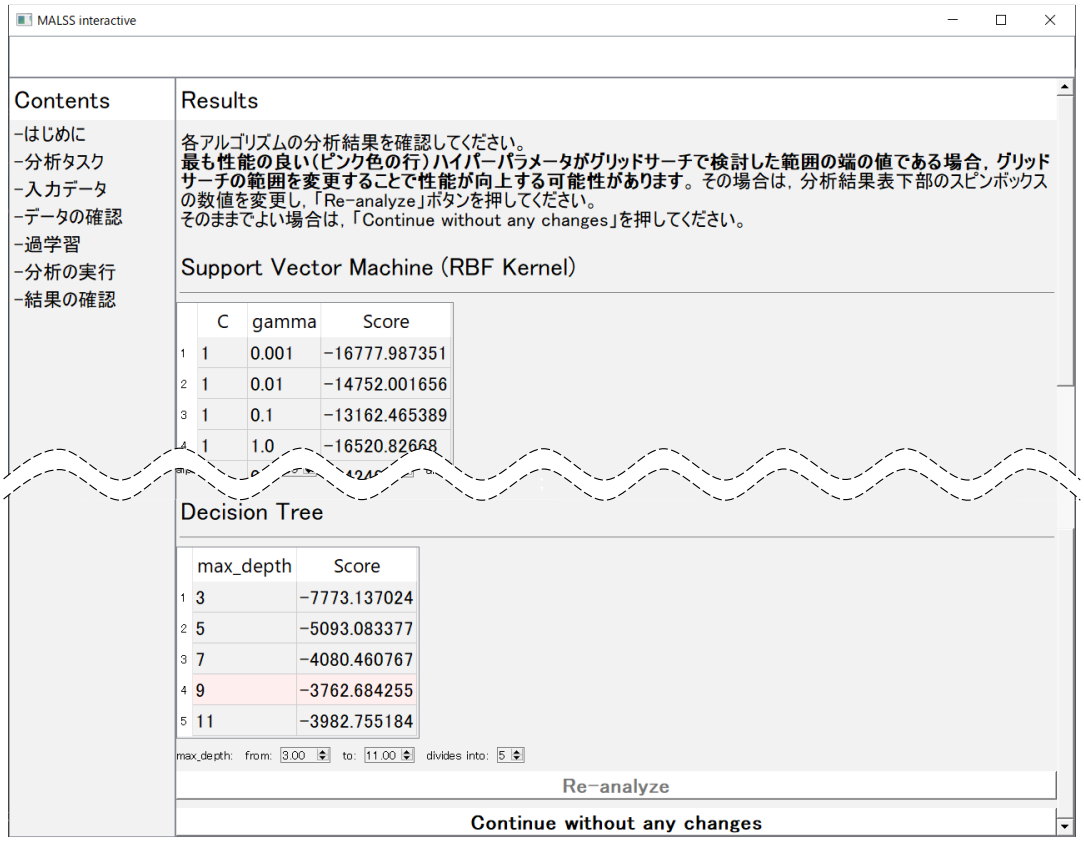


図 5.2　ハイパーパラメータの調整（コンテンツビューの「結果の確認」に対応）.

■次元削減　MALSS では，モデルの性能向上への貢献が小さい特徴量を削除する特徴量選択による次元削減を行う．次元削減は特徴量選択と，主成分分析などを用いて高次元の特徴量を低次元の特徴量に変換する特徴抽出に大別される．特徴量選択は特徴抽出と比較して，手法および結果の理解が容易であるため，MALSS では特徴量選択による次元削減を採用した．特徴量の選択には，特徴量の重要度を表す指標である Permutation feature importance (PFI) [31] を用いる．PFI の計算には OSS のライブラリ，rfpimp [*2]を利用し，PFI が負となる特徴量を削除する．

特徴量選択は，モデルが学習データに過度に適合している過学習状態のときに効果的であり，モデルの学習・評価結果に応じた適切な実行要否の判断が重要である．図 5.3 は次元削減実施後の「特徴量選択」コンテンツに対応するメインビューである．「バイアスとバリアンス」のコンテンツで，作成したモデルの状態を確認するためには学習曲線を作成することが有効であることを教示している．そして「学習曲線」のコンテンツで分析者に

[*2] https://github.com/parrt/random-forest-importances (2022/2/10 access)

学習曲線を提示する．分析者はモデルの状態を確認したうえで，特徴量選択による次元削減を行う（あるいは行わない）という判断を下すことができる．分析者が特徴量選択を実施した場合，「特徴量選択」のコンテンツで選ばれた特徴量を提示し，選ばれた特徴量を用いて再度モデルの学習と評価のプロセスを行うよう分析者に促す．

PFI を用いた特徴量選択は有効であるものの，多くの初学者向けの教科書では扱っていないため，基本的な機械学習の知識の範疇を超えると判断し，特徴量選択の手法や計算結果（特徴量ごとの PFI の値など）の提示は行わないこととした．特徴量選択手法の効果的な知識習得支援方法については今後の課題である．

図 5.3　特徴量選択による次元削減（コンテンツビューの「特徴量選択」に対応）．

5.3　評価

本章では，機械学習を用いたデータ分析における特徴量エンジニアリングとプロトタイピングのプロセスにおいて，サブプロセスの実行結果に応じて分析の実施内容を変更する

ことが困難であるという従来の MALSS の課題に対し，サブプロセスごとの分析を実行可能とし（機能要件 1），サブプロセスの実行結果に応じて分析者の入力を受け付けプロセス実施内容を変更可能とする（機能要件 2）手法を提案する．

本節では，機能要件 1，2 の実現により非機能要件 1，2 が満たされることを確認するための評価実験について述べる．非機能要件が満たされることを確認するために，模擬データ分析実験および知識確認テストを行った．

5.3.1 模擬データ分析実験

非機能要件 1 について検証するために，模擬データを用いて実際に分析を行う実験を行った．

実験協力者

模擬データ分析実験の協力者は 12 名であり，全員 1 年以上のプログラミング経験を有するデータ分析の未経験者あるいは初心者である．Python の未経験者を A グループ（7 名）に，経験者を B グループ（5 名）に分けた．A グループは提案手法である GUI を備えた MALSS を利用してデータ分析を行い，B グループは従来の MALSS を利用して分析を行う．

模擬データ

実験は数値データの目的変数（*Target*）を予測する回帰タスクであり，実験に用いる模擬データは以下の手順により作成した．

データの特徴量の数は 21 個（*Num1, ..., Num19, Cat1, Cat2*）で，量的変数（*Num*）が 19 個，カテゴリ変数（*Cat*）が 2 個である．1〜3 個目の量的変数 x_1〜x_3 は標準正規分布に従う乱数であり，目的変数 y と以下の関係をもつ．

$1 \leq i \leq 1000$ **のとき**

$$y_i = 10.3x_{1,i} + 51.9x_{2,i} + 9.68x_{3,i} + 10 + N(20, 1)$$

$1001 \leq i \leq 2000$ **のとき**

$$y_i = 15.1x_{1,i} + 24.4x_{2,i} + 54.7x_{3,i} - 10 + N(20, 1)$$

$2001 \leq i \leq 3000$ **のとき**

$$y_i = 92.6x_{1,i} + 43.1x_{2,i} + 68.0x_{3,i} + N(20, 1)$$

ここで $x_{1,i}$ は変数 x_1 の i 番目のデータであることを示す．また，$N(\mu, \sigma^2)$ は，平均 μ，分散

σ^2 の正規分布に従う乱数を表す．上式に基づくデータは scikit-learn の *make_regression* メソッドにより生成した．4〜19 個目の量的変数 x_4〜x_{19} は，区間 $[0, 1)$ の一様分布に従う乱数とした．つまり，これらの変数は予測に寄与しない特徴量であり，分析を行うにあたり，これらの変数を特徴量選択により除去することが重要となる．

1 個目のカテゴリ変数 x_{20} は下式に従い生成した後，ランダムに 10% のサンプルの位置を入れ替えた．

$$x_{20,i} = \begin{cases} 1 & (1 \leq i \leq 1000) \\ 2 & (1001 \leq i \leq 2000) \\ 3 & (2001 \leq i \leq 3000) \end{cases} \tag{5.1}$$

上式に示すように，変数 x_{20} は数値データとして表されるが，カテゴリ変数として扱うことが適切である．2 個目のカテゴリ変数 x_{21} は，1〜4 の数値をランダムに割り振った．つまり，カテゴリ変数のうち予測に寄与するのは x_{20} のみである．数値データとして表されている変数をカテゴリ変数として扱うべきか否かは，ドメイン知識等に基づき決定する必要がある．本実験では事前に変数 x_{20} および x_{21} はカテゴリ変数であることを実験協力者に明示した．

データ作成後シャッフルし，1000 サンプルを訓練用，2000 サンプルをテスト用とした．実験参加者は訓練用データを用いて予測モデルを作成し，テスト用データの予測結果を提出する．モデルの性能評価基準は平均二乗偏差とした．

実験方法

模擬データ分析実験では初めに実験協力者に以下の情報を提示した．

- 実験タスクは与えられたデータに含まれる変数 *Target* の値を予測する回帰タスクであり評価指標は平均二乗偏差である．
- *Target* 以外の変数 *Num1*，...，*Num19*，*Cat1*，*Cat2*（5.3.1 の x_1〜x_{20} に対応）は予測に利用する変数で，*Cat1* と *Cat2* はカテゴリ変数である．
- 訓練用データを用いてモデルを作成し，テスト用データを用いて予測した *Target* の推定値を提出する（実験協力者はテスト用データの真の *Target* の値を知ることはできない）．
- データ分析に費やす時間は 1 時間を上限とする．

さらに，A グループは提案手法である GUI を備えた MALSS を，B グループは従来の MALSS を利用して分析を行うよう指示し，MALSS の利用方法を指示した．両グループとも，MALSS 以外の機械学習ライブラリの利用は禁止し，分析の際に MALSS の使用説

明書，MALSS の出力する知識習得のための情報，および分析レポートのリンク先（B グループのみ）以外からの情報取得も禁止した．

最後に，実験協力者から提出されたテスト用データの *Target* の推定値を用いて，筆者が，各実験協力者が作成したモデルの性能を評価した．

模擬データは数値データで表されたカテゴリ変数を正しくカテゴリ変数として扱い，特徴量選択による次元削減を行うことが適切であるように作成している．提案手法では，5.2.2 項で述べた「カテゴリ変数の処理」機能と「次元削減」機能により，GUI を用いてこれらの処理を容易に実行可能となっている．さらに，モデルの性能向上に有効なハイパーパラメータ調整処理も，従来の MALSS であれば分析全体を何度も繰り返す必要があったが，提案手法では「ハイパーパラメータ調整」機能により分析プロセスの途中で，モデルの性能評価結果を確認しながら，ハイパーパラメータ調整処理のみを繰り返すことができる．

このように，分析者がこれらのプロセスの必要性を正しく認識し，適切に実行することで，より性能のよいモデルを作成することができるようになっており，模擬データ分析実験の実験結果（モデルの性能）を比較し，提案手法を用いて分析を行う A グループの方が作成したモデルの性能が高ければ，非機能要件 1 が満たされていることを確認することができる．

実験結果

実験結果を図 5.4 に示す．縦軸は予測値と正解値の平均二乗偏差を表し，エラーバーは 95% 信頼区間を表す．A グループと B グループの平均値の差について，平均値に差がないことを帰無仮説とし，コルモゴロフ・スミルノフ検定（両側検定）を行ったところ，p 値は 0.002 であり，有意水準 5% で有意に A グループが作成したモデルの方が性能が良いと言える．

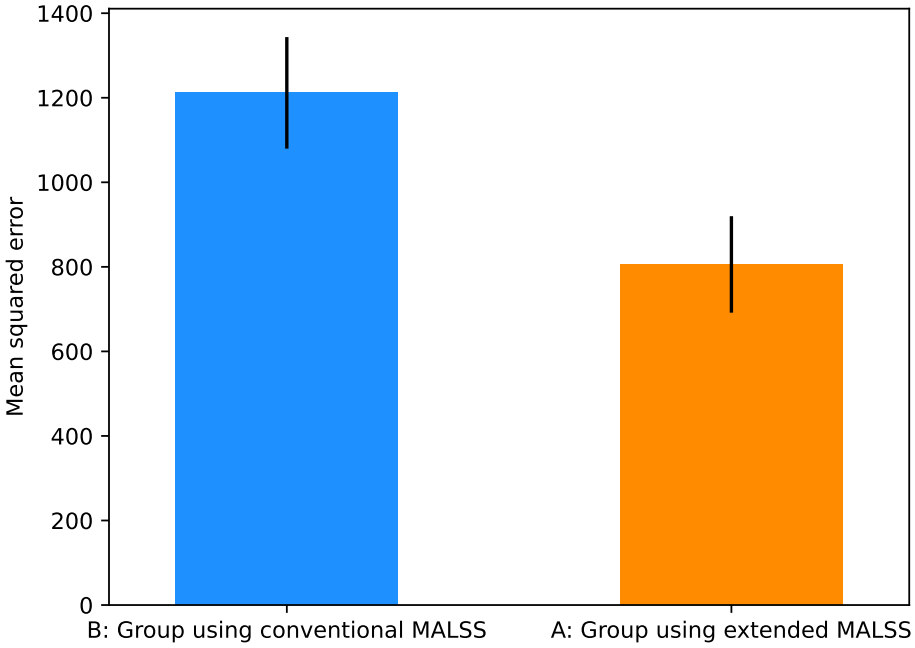


図 5.4　模擬データ分析実験結果.

5.3.2　知識確認テスト

非機能要件 2 について検証するために，模擬データ分析実験を通じた分析者の知識習得度合いを，知識確認テストにより評価した.

評価方法

5.3.1 で述べた模擬データ分析実験を行う前に，実験協力者に対し，機械学習を用いたデータ分析に関する知識の有無を問うテストを行った．テストの設問を図 5.5 に示す．設問は A，B グループ共通で，設問数は 15 問，30 個の選択肢から回答を選ぶ形式である．問題は，4.3 で示した知識習得支援範囲の知識に関して，提案手法の GUI 上で提示される情報，および従来の MALSS の分析レポートに記載されている情報をもとに正答できるように設計した．ただし，設問 1 から 4 は上述の知識習得支援範囲に含まれていない．これは従来の MALSS では特徴量エンジニアリングとプロトタイピングのプロセスを全て自動化する都合から，分析の目的は決定していることを想定しており，分析タスクに関する知識は習得済みであることを前提としていたためである．提案手法ではサブプロセスごとの分析実行が可能であるため，教師あり学習や教師なし学習，分類や回帰といった分析タスクに関する情報を提示し分析者の知識習得を支援しながら，分析者に実行するタスクを

選択させることが可能となる．従来の MALSS の分析レポートは設問 1 から 4 に関する情報を含まないため，B グループの実験協力者が不利にならないよう，実験方法説明資料に上記設問に関連する情報を記載した．テスト回答にあたっては，資料等を参照することと，勘で回答することを禁止した（分からない設問は分からないという選択肢を選ぶように指示した）．

さらに，模擬データ分析実験終了後に再び同じテストを行い，スコアの変化から学習効果を測定した．暗記力の評価になってしまわないよう，実験協力者には，模擬データ分析実験後に再び同じテストを行うことは伝えていない．

1. （A）は機械学習タスクの1つで、正解（こういう結果を出してほしいという情報）がついている学習データに基づき、入力と出力の関係を学習するタスクです。
2. （B）は、ラベルがついていない学習データから、何らかの関係性／パターンを発見する機械学習のタスクです。
3. Aにおいて、数値を予測するタスクを（C）タスクと呼びます。
4. Aにおいて、ラベルを予測するタスクを（D）タスクと呼びます。
5. データに含まれる変数は、数値変数とカテゴリ変数の2つに大別されます。多くの機械学習アルゴリズムは直接カテゴリ変数を扱うことができないため、（E）を用いてカテゴリ変数を数値変数に変換することがよく行われます。
6. （F）能力とは，学習したモデルが，学習データに含まれない未知のデータに対して適切に対応できる能力です。
7. モデルの性能は，学習前に人手で設定するハイパーパラメータに大きく依存します。ハイパーパラメータを最適化するために、（G）がよく用いられます。
8. 学習したモデルが学習データだけに過度に適応してしまい、未知のデータに対する性能が低下してしまった状態を（H）と呼びます。
9. 機械学習アルゴリズムには、Hを防ぐための、（I）パラメータというハイパーパラメータを持つものがあります。
10. モデルの、未知のデータに対する性能を推定するために、（J）検証法がよく用いられます。
11. H状態のことを、（K）とも呼びます。
12. H状態の反対の状態 (モデルがデータの特性を十分に表現できていない状態)を（L）と呼びます。
13. Kの場合、データを増やす、特徴量の数を（M）などが性能向上のために有効です。
14. Lの場合、特徴量の数を（N）、アルゴリズムをより複雑なものに変更することが、性能向上のために有効です。
15. （O）を作成し確認することで、モデルがKな状態なのか、Lな状態なのかを判断することができます。

選択肢：

1. ROC曲線,	2. ダミー変数,	3. 近傍探索,	4. 回帰,	5. グリッドサーチ,
6. 過学習,	7. 学習曲線,	8. 機械学習,	9. 強化学習,	10. 教師あり学習,
11. 教師なし学習,	12. クラスタリング,	13. 交差,	14. 混同行列,	15. 次元削減,
16. 次元の呪い,	17. 正則化,	18. 説明変数,	19. データマイニング,	20. 特徴量選択,
21. ハイバイアス,	22. ハイバリアンス,	23. 汎化,	24. フィッティングカーブ,	25. 増やす,
26. 分類/識別,	27. ベイズ,	28. 減らす,	29. 目的変数,	30. 分からない

図 5.5　知識確認テストの設問.

知識確認テストの結果を図 5.6，および図 5.7 に示す．図 5.6 は，横軸が模擬データ分析実験前のテストの正答数を，縦軸が分析実験後の正答数を示している．ひとつのプロットがひとりの実験協力者のデータを表しており，プロットが点線よりも上に位置してい

れば，模擬データ分析実験の前後でテストのスコアが向上したことを表している．図 5.7 は，グループごとの正答数の平均値を棒グラフで示したものである．エラーバーは 95% 信頼区間を表す．A，B グループそれぞれについて，模擬データ分析実験前後の正答数の差について，符号検定による対応のある 2 つの母平均の差の検定（両側検定）を行ったところ，p 値は，A グループは 0.0156，B グループは 0.0625 であった．また，A グループと B グループの正答数の向上値の差について，コルモゴロフ・スミルノフ検定（両側検定）を行ったところ，p 値は 0.32 であった．したがって，有意水準を 5% としたとき，提案手法の MALSS を利用した A グループは実験の前後で有意にテストのスコアが向上したと考えられる一方，A グループと B グループの正答数向上の値に差がないという帰無仮説を棄却することはできない．

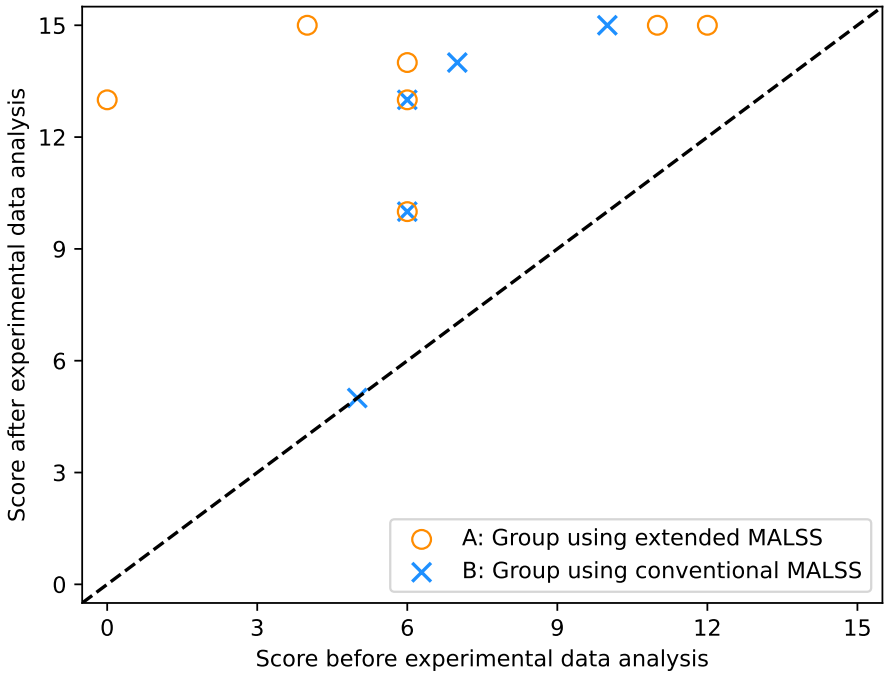


図 5.6　知識確認テスト結果（散布図）.

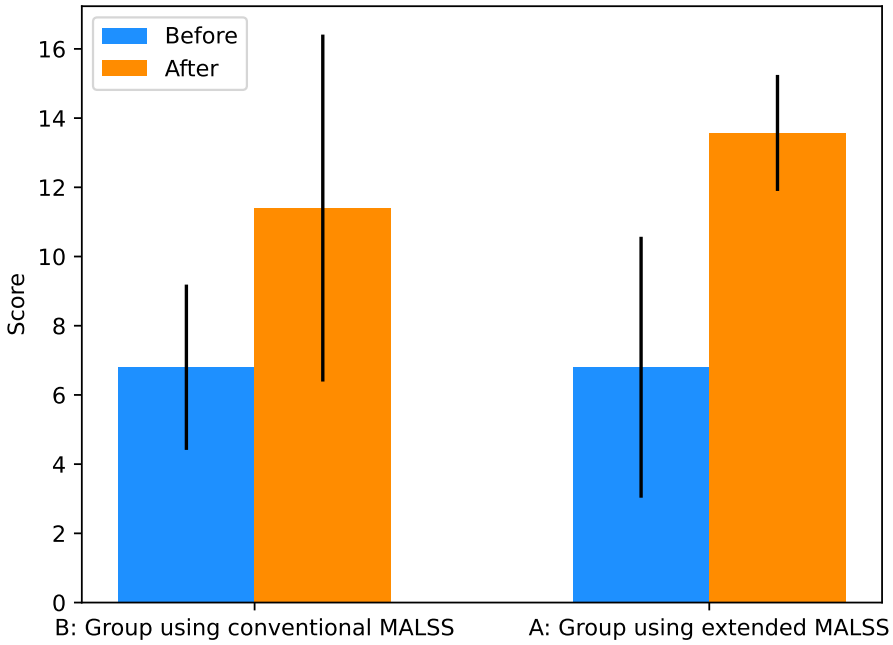


図 5.7　知識確認テスト結果（棒グラフ）.

5.4　考察

模擬データ分析実験の結果（図 5.4）から，提案手法により作成したモデルの性能は従来手法のモデルの性能よりも有意（有意水準 5%）に良く，より適切な分析支援が可能となっていることから，非機能要件 1 が満たされていることが確認できる.

また，知識確認テストの結果（図 5.6）から，提案手法の MALSS を利用した A グループは有意（有意水準 5%）にテストのスコアが向上したものの，両グループの正答数向上の値（図 5.7）に有意（同 5%）な差は認められなかった．しかし，提案手法の目的は訓練データやモデルの学習・評価結果に応じた分析支援を可能とすることであり，知識習得については従来の MALSS と同等であることを非機能要件としているため，非機能要件 2 は満たされていることが確認できた．GUI 化による知識習得効果改善の有無についてはより大規模な検証が必要であり今後の課題である.

以上から，訓練データやモデルの学習・評価結果に応じた分析支援を可能とした提案手法の MALSS は，従来の MALSS と同等以上の知識習得効果を備え，従来の MALSS よりも質の高い分析が可能であることを確認できた.

模擬データ分析実験における，各分析プロセスの実施結果は表 5.1 のようになる．セル

内の数値は，各グループの実験協力者のうち，当該処理を実施するという判断をした実験協力者の人数を表している．カテゴリ変数の処理については，処理を実施するという判断をした 7 人全員が正しく *Cat1*，*Cat2* のみをカテゴリ変数として指定していた．ハイパーパラメータ調整については，実施した 5 人全員が，最も性能の良いモデルのハイパーパラメータが，ハイパーパラメータ試行範囲の端点でないように調整することができていた．特徴量選択については，実験協力者は特徴量選択処理を行うという判断をするのみであり，実際にどの特徴量が選択されるかは MALSS が利用するライブラリ rfpimp により決定される．今回の模擬データ分析実験では，予測に寄与しない変数のうち，変数 *Num5*〜*Num14*，*Num17*，*Num18* と *Cat2* を 4 つのダミー変数に変換したうちの 2 変数が rfpimp により削除された．

表 5.1　各分析プロセスを実施した実験協力者数.

	B グループ 従来の MALSS	A グループ 提案手法の MALSS
カテゴリ変数の処理（ダミー変数の利用）	0/5	7/7
ハイパーパラメータ調整	0/5	5/7
特徴量選択	3/5	5/7

従来手法の MALSS を利用したグループでは，特徴量選択は 5 人中 3 人が実施できたが，ダミー変数を用いたカテゴリ変数の量的変数への変換と，ハイパーパラメータ調整を実施できた実験協力者はいなかった．これは先行研究 [32] の結果と同様である．分析レポートを分析実行後に確認し，適切にこれら分析プロセスの必要性を判断し，必要なプログラミングを行うことは，分析未習熟者には困難であるためと考えられる．

一方，提案手法の MALSS を利用したグループでは，特徴量選択とハイパーパラメータ調整は 7 人中 5 人が，カテゴリ変数の処理は 7 人全員が実施することができた．GUI を備え，分析の途中でモデルの学習・評価結果を提示しながら分析プロセスの説明を行うことで，分析未習熟者であっても，適切にこれらの分析プロセスを実施することが可能となる．3 つの処理をすべて実施するという判断をした実験協力者は 7 人中 3 人であり，全員提案手法の MALSS を利用した実験協力者であった．彼らの模擬データ分析実験の結果（作成したモデルの性能）は，提案手法の MALSS を筆者の想定通りに利用した場合と同じであった．特徴量選択あるいはハイパーパラメータの調整のいずれかをを実施しなかった 4 名について，いずれの実験協力者も，知識確認テストにおいて特徴量選択またはハイパーパラメータ調整に関連する設問に正答しており，分析プロセスの実施・不実施と知識

確認テストの結果に特異な傾向は認められなかった．

知識確認テストの結果について，B グループの結果，および 4 章の知識確認テストの結果をみると，分析レポートや Web 上の情報をもとに知識を習得した実験協力者は，模擬データ分析実験前後のテストスコアに強い正の相関（相関係数 0.72）がみられる一方，GUI 上の情報提示により知識を習得した実験協力者には強い相関はみられない（相関係数 0.37）．従来の知識習得方法では，習得できる知識量が予め備えている知識量に依存する一方，提案手法では，事前の知識量に依らず必要な知識を身に着けることが可能となっている．

5.4.1　妥当性への脅威

本項では，本研究の結果の妥当性に影響を及ぼす恐れのある事項について述べる．

知識習得度合いの評価について，模擬データ分析実験後に知識確認テストを行うことを実験協力者に伝えないことで，暗記力の評価になることを避けているが，長期的な知識の定着度合いについては別途評価を行う必要がある．ただし，分析業務に従事する際に MALSS を利用し続けることは反復学習を行うこととなり，知識の定着に有効であると考えられる．

また，本論文で実施した評価実験は模擬データを用いた分析であるため，提案手法が実際の業務で必要となるデータ分析を適切に遂行することができるか，分析者が実際の業務で必要となる知識を習得することができるかについてはさらなる評価が必要である．

5.5　第 5 章におけるまとめ

本章は，機械学習を用いたデータ分析手順に関する知識の不足した分析者の，分析遂行とデータ分析に関する知識習得を同時に支援する Python ライブラリ MALSS の機能を拡張し，教師あり学習において，入力される訓練データやモデルの学習・評価結果に応じて分析を支援する機能を開発した．先行研究の MALSS では，分析自動化により分析支援を行い，分析後に分析レポートを作成することで，分析者の知識習得を支援した．しかし，分析手順の一部においては完全な自動化が困難であり，訓練データやモデルの学習・評価結果に応じて適切な分析手段を選択する必要がある．そこで提案手法では，グラフィカルユーザーインタフェースを備え，訓練データやモデルの学習・評価結果に応じた情報の提示と，提示された情報に基づく分析者のアクションを受け付ける分析支援を行うことで，分析者が適切な分析手段を選択することを支援する．模擬データ分析実験および知識

確認テストにより，提案手法の MALSS は先行研究の MALSS と同等以上の知識習得支援を可能としつつ，分析プロセスの途中で分析者の判断に基づく入力を受け付け分析内容を変更することで，より質の高い分析が可能であることを確認した.

第 6 章

教師なし学習を用いたデータ分析の支援

6.1　教師なし学習と用いたデータ分析支援における課題

3.1 節でも述べたように，多くの AutoML OSS は教師あり学習を対象としており，教師なし学習をサポートするものは多くない．これは，教師なし学習がデータに正解として扱える情報を含まず，分析結果の妥当性を分析者が主観的に評価をするプロセスを分析プロセスの中に含むためであると考えられる．

しかし，教師なし学習においても，誤った分析プロセスによる分析は分析に失敗や誤った意思決定につながる恐れがあり，適切な支援を行うことが重要となる．適切な支援を必要とする例を図 6.1 に示す．図 6.1 は，教師なし学習の一つであるクラスタリング分析の結果を図示したものである．$x1$，$x2$ という 2 次元の変数をもつデータを，代表的なクラスタリングアルゴリズム，k-means アルゴリズムを用いて，3 つのクラスターに分割している．図中には，一つの大きなクラスターと，その右側に二つの小さなクラスターがあることを容易に目視することができる．しかし，k-means アルゴリズムの結果は，データを適切にクラスタリングすることができていない．これは，k-means アルゴリズムが，各クラスターが同一サイズの超球であることを仮定しているためである．もし分析者がこのようなアルゴリズムがもつ前提条件などの正しい知識をもっていない場合，不適切な分析結果や誤った結論を得てしまう恐れがある．

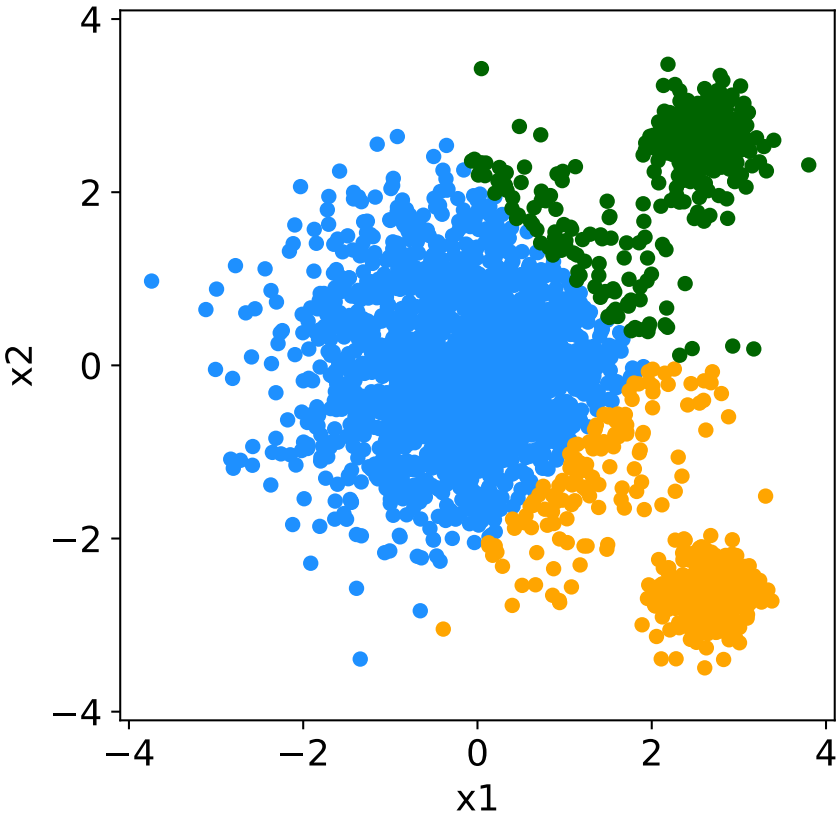


図 6.1　誤ったクラスタリング結果の例．k-means アルゴリズムは大きさの異なるクラスターを適切に分割することができない．

このような課題に対し，本章では，教師なし学習の一つであるクラスタリング分析について，自動化による分析支援と，分析に関する知識習得支援を行う方法を提案する．

6.2　クラスタリング分析の支援

6.2.1　クラスタリング分析の自動化

クラスタリング分析

MALSS は，クラスタリング分析の主要な目的の一つである，適切なクラスター数の推定を自動化することで，クラスタリング分析を支援する．クラスタリング分析は，データを何らかの類似度指標に基づき類似したいくつかのグループに分割するタスクである．クラスタリング分析は，データサンプルが正解となるラベル（所属するグループなど）をもたないため，教師なし学習に分類される．一般的にクラスター数は未知であり，多くのク

ラスタリングアルゴリズムは分析者が事前に推定されるクラスター数を指定する必要がある．分析後に，分析者は分析結果を可視化するなどして，結果の妥当性を解釈する必要がある．

クラスター数推定

適切なクラスター数を推定するための様々な指標が提案されている．これら複数の指標の多数決により適切なクラスター数を推定する手法も存在する．NbClust [33] は 30 個のクラスター数推定指標の多数決により最適なクラスター数を推定する，R 言語の OSS ライブラリである．NbClust は AutoML OSS ではないため，分析者は自身でデータ分析のパイプラインを構築する必要がある．

MALSS は NbClust と同様に，複数のクラスター数推定指標の多数決により最適なクラスター数を推定する．MASS は 4 つの主要なクラスター数推定指標，Silhouette スコア [34]，Davies Bouldin スコア [35]，Calinski Harabasz スコア [36]，Gap 統計量 [37] を用いる．最初の 3 つの指標は MALSS が機械学習アルゴリズムのバックエンドエンジンとして利用している既存のライブラリ，scikit-learn を用いて計算する Gap 統計量は著者が実装したものを用いる．

クラスタリングアルゴリズム

我々はクラスタリングアルゴリズムとして最も基本的な k-means クラスタリングと階層型クラスタリングを用いる．より発展的なアルゴリズムを利用することも可能であるが，MALSS はデータ分析の未習熟者が基本的なデータ分析を実行することと，基本的なデータ分析に関する知識を習得することを目的としているため，意図的に基本的なアルゴリズムを選択している．

k-means アルゴリズムは [38, 39] クラスター内の分散が最小になるようにあらかじめ定められた k 個のクラスター重心を決定する（データサンプルは最近傍のクラスター重心に属するものとする）．距離の計算にはユークリッド距離が用いられる．階層型クラスタリング [40] は，最近傍のクラスターのペアを結合していく（凝集型），あるいは大きなクラスターを分割していく（分割型）ことで，木構造の階層型クラスターを生成する．MALSS は k-means クラスタリングアルゴリズムは scikit-learn のものを，階層型クラスタリングアルゴリズムは Python の数値解析ライブラリである SciPy [41] が提供するものを利用する．k-means クラスタリングは分析者があらかじめ推定されるクラスター数を設定する必要がある．階層型クラスタリングはクラスター数を設定する必要はなく，得られた木構造の階層型クラスターから，所望のクラスター数のクラスターが得られるよう

に木の切断点を決定する．SciPy には指定されたクラスター数に応じて切断済みのクラスターを求める機能が実装されている．

MALSS によるクラスタリング分析

MALSS は事前に探索するクラスター数の範囲が設定されており（分析者はこれを変更することもできる），その範囲の中で推定クラスター数を変えながら上記 2 種類のクラスタリングアルゴリズムを繰り返し実行する．そして実行のたびに 4 つのクラスター数推定指標を計算し，当該推定クラスター数におけるクラスタリングの妥当性を評価する．最後に，複数のクラスター数推定指標におけるクラスター数推定結果から，多数決により最適なクラスター数を推定する．

MALSS のクラスタリング分析自動化全体のワークフローは図 6.2 のようになる．教師あり学習の分析支援の時と同様に，前処理の段階において，MALSS は入力データの標準化，欠損値の処理，カテゴリ変数の数値変数への変換を行う．次に，モデルの学習を行い，最適なクラスター数を推定する．最後に分析者の知識習得を支援するための分析レポートを生成する．

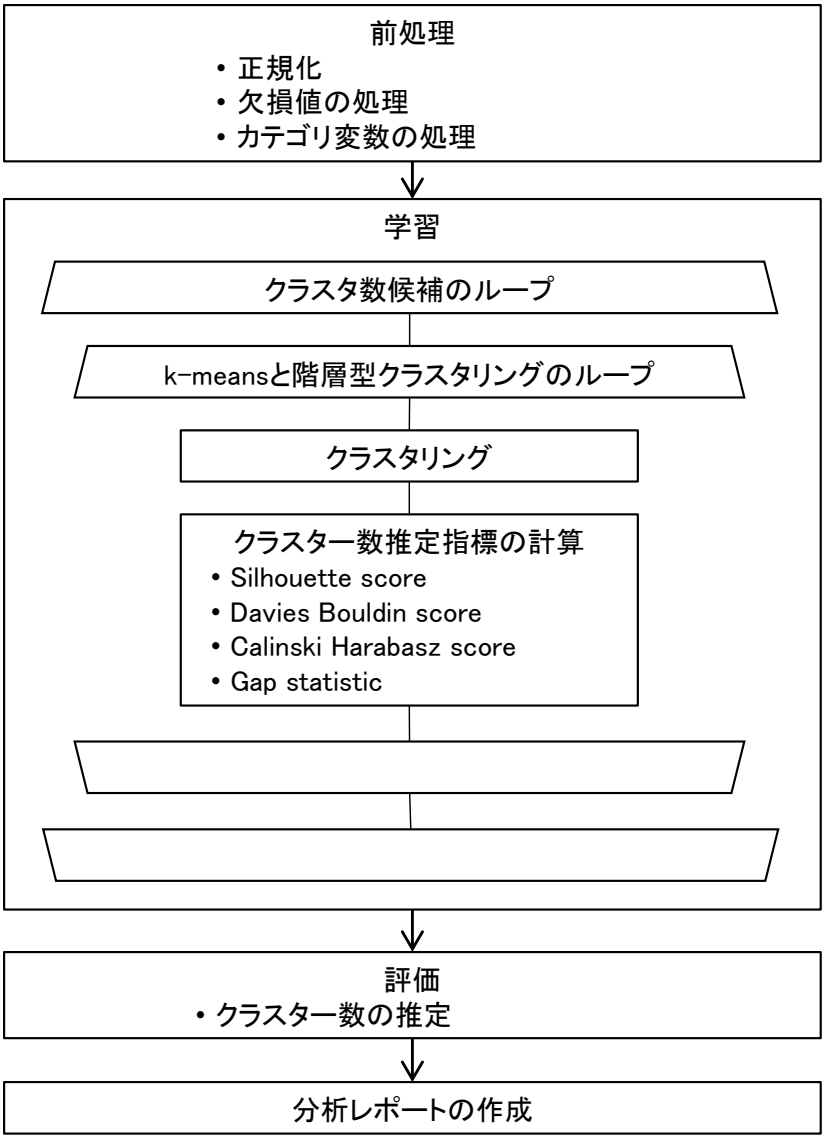


図 6.2　クラスタリング分析自動化のワークフロー.

クラスタリング分析を行う場合の MALSS のコード記述例を図 6.3 に示す．分析者は，分析目的（clustering）を指定して MALSS のインスタンスを生成し，*fit* メソッドにデータを渡すだけで，容易にクラスタリング分析を行うことができる．また，*predict* メソッドに未知のデータを渡すことで，新たなデータサンプルが属するクラスターを求めることができる．

```
from malss import MALSS

model = MALSS(task='clustering')
model.fit(data, dname='report_dir')
pred = model.predict(data_new)
```

図 6.3　MALSS によるクラスタリング分析のコード記述例.

6.2.2　クラスタリング分析に関する知識習得支援

本研究では，第 4 章と同様に，分析完了後に分析レポートを分析者に提示することで，分析者のクラスタリング分析に関する知識習得を支援する．我々は，レポートの記述内容をクラスタリング分析のワークフローをカバーする以下の 4 つのカテゴリに分類した.

- Clustering
- Preprocessing
- Algorithms
- Interpretation

Clustering カテゴリは，クラスタリング分析は分析結果の主観的な妥当性評価が必要であるなど，クラスタリング分析に関するトピックをカバーする．Preprocessing カテゴリは，データの標準化や特徴量選択などデータの前処理に関するトピックを扱う．Algorithms カテゴリは，クラスタリングアルゴリズムの特徴と，利用時の注意点を記述する．Interpretation カテゴリは，分析結果を解釈する際の注意点について述べる.

各カテゴリは複数の内容を記述する．記述内容は当該分野で著名な複数の教科書を参考に選定した．すべての記述内容を表 6.1 に示す.

表 6.1　分析レポート記載項目

カテゴリ	レポート記載項目
Clustering	• データは正解ラベルを含まないため，分析者は分析結果の妥当性を主観的に判断する必要がある． • 多くのクラスタリングアルゴリズムは推定されるクラスター数を事前に指定する必要がある． • 最適なクラスター数を推定するための様々な指標が提案されている．
Preprocessing	• 特徴量のスケールの違いの影響を低減するためデータを標準化する必要がある． • 数値データとして表されている特徴量もカテゴリデータとして扱う必要があることもある． • 特徴量の数が多い場合特徴量選択や次元削減を検討する必要がある．
Algorithms	• クラスターが同じサイズの超球でない場合 k-means アルゴリズムは不適切な結果となる場合がある． • 階層型クラスタリングは凝集型と分割型に大別される． • 階層型クラスタリングはデンドログラムという木構造で可視化できるという利点をもつ．
Interpretation	• デンドログラムの水平方向の近さはデータの類似性を表さないので注意が必要である． • 階層型クラスタリングはデータが階層構造をもつか否かに関わらず常に階層構造の結果を返す． • デンドログラムの形状はデータの非類似性の定義に大きく依存する．

分析レポートの一部を図 6.4 と図 6.5 に示す．レポートでは，最初にクラスター数推定結果を示す（図 6.4）．そして，クラスタリング分析について，入力データの詳細，各アルゴリズムの分析結果詳細を示す．これらの記述を確認することで，分析者は MALSS がどのようにデータを分析し，また分析の際にどのような点に気を付けなければいけないかという知識を習得することができる．さらに，レポート中の専門用語には当該用語を解説している Web サイトへのリンクを付けた．階層型クラスタリングの結果詳細説明部（図 6.5）では，デンドログラムを用いてクラスタリングの結果を可視化し確認することができる．

分析結果

アルゴリズム	Estimated number of clusters			
	Gap統計量	Silhouetteスコア	Davies-Bouldinスコア	Calinski and Harabaszスコア
K-Means	3	2	2	2
Hierarchical Clustering	3	3	3	3

多数決から推定されるクラスタ数は**3**です.

クラスタリング [Back To Top]

- クラスタリングとは，**データをある共通の特徴をもつ部分集合（クラスタ）に分割する**分析手法です.
- クラスタリングは一般的に正解が与えられない（教師なし学習）ため，**人が分析結果の妥当性を判断する必要があります**.
- 多くのクラスタリングアルゴリズム（MALSSが採用するものも）は，想定されるクラスタ数を事前に人が設定してやる必要があります． しかし，真のクラスタ数は一般に未知です.
  そこでMALSSでは，あらかじめ与えられた複数のクラスタ数候補に対して分析を行います.
- 適切なクラスタ数を判断する様々な指標が提案されています．MALSSでは，**複数の指標から多数決でクラスタ数を推定**します.
- MALSSがクラスタ数の判断に用いる指標は， Gap統計量， Silhouetteスコア， Davies-Bouldinスコア， Calinski and Harabaszスコア の４つです.

図 6.4　分析レポートの一部.

階層的クラスタリング (Hierarchical clustering) [Back To Top]

- 階層的クラスタリングは，k-meansクラスタリングと並んで最もポピュラーなクラスタリングアルゴリズムの１つです．
- 階層的クラスタリング手法は，**データ一つ一つを順次併合していく凝集型（agglomerative）**と，**データ全体を順次分割していく分割型（divisive）**に大別されます．
  MALSSでは凝集型階層的クラスタリングをサポートしています．
- 階層的クラスタリングには，分析結果を**デンドログラム（dendrogram）**という樹形図で可視化できるという利点があります．
  しかし，デンドログラムを見る際には，以下の点に注意する必要があります．
  - デンドログラムの縦軸はクラスタ間の距離を表すが，横軸の描画方法には任意性があり**近くのデータと類似しているとは限らない**．
  - デンドログラムは常に階層構造を示すが，**データに本当にそのような階層構造が存在するとは限らない**．
  - デンドログラムの形状はクラスタ間の距離の計算方法（linkage criteria）に強く依存します．
    MALSSではlinkage criteriaに完全連結法（complete linkage method）を採用しています．

デンドログラム（Dendrogram）

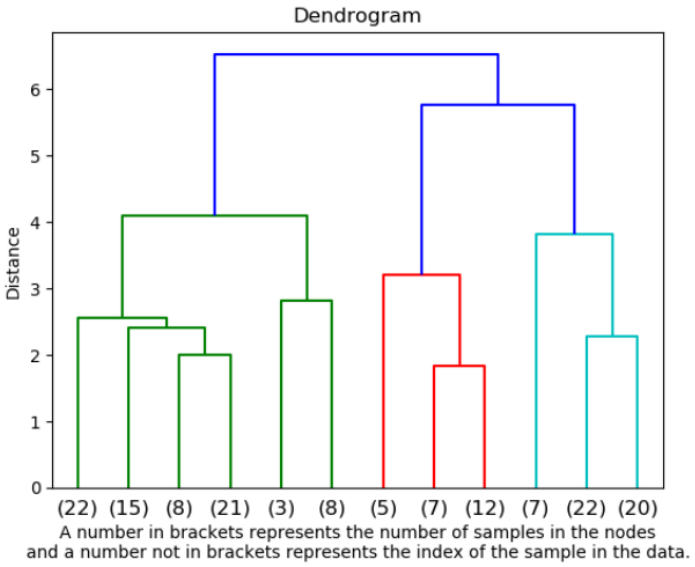


図 6.5　分析レポートの一部．

6.3 評価

本章における研究課題は，MALSS を用いることでクラスタリング分析を適切に実施することができるか（RQ1）と，分析者が分析レポートを読むことでクラスタリング分析に関する知識を習得することができるか（RQ2）である．これらの問いに答えるために，模擬データ分析実験と知識確認テストを行った．

6.3.1 クラスタリング分析自動化に関する評価

MALSS を用いたクラスタリングデータ分析の質を評価するために，scikit-learn に含まれる 3 つのオーブンデータセット，iris，wine，breast cancer を用いた．これらデータセットの詳細を表 6.2 に示す．これらのデータセットは識別タスクに用いられることを想定しているため，各サンプルには属するクラスのラベルが与えられている．そこで，クラスタリング分析により推定したクラスター数が実際のクラス数と一致するかを分析の質として評価することとする．scikit-learn に含まれる，識別タスクに用いることができるテーブルデータ形式のデータセットは，現在のところこの 3 つで全てである．

表 6.2　評価データセット

Datasets	Iris	Wine	Breast cancer
クラス数	3	3	2
クラス内サンプル数	50	59, 71, 48	212, 357
総サンプル数	150	178	569
次元数	4	13	30
特徴量種別	real, positive	real, positive	real, positive

評価結果を表 6.3 と表 6.4 に示す．表 6.3 には比較対象として，H2O，NbClust でクラスター数推定を行った結果も示している．表 6.4 に示している精度は，クラスタリング結果と実際のクラスとの一致度である．

表 6.3　評価結果

データセット	クラス数	推定クラスター数			
		H2O	NbClust		MALSS
			Default	Standardized	
Iris	3	2	3	3	3
Wine	3	3	2	3	3
Breast cancer	2	4	4	4	2

表 6.4　クラスター数推定指標ごとの推定結果

データセット	アルゴリズム	クラス数	推定クラスター数				精度
			Gap statistic	Silhouette	Davies-Bouldin	Calinski and Harabasz	
Iris	K-means	3	**3**	2	2	2	0.877
	Hierarchical		**3**	**3**	**3**	**3**	0.787
Wine	K-means	3	4	**3**	**3**	**3**	0.972
	Hierarchical		5	**3**	8	**3**	0.837
Breast cancer	K-means	2	10	**2**	**2**	**2**	0.905
	Hierarchical		4	**2**	**2**	**2**	0.631

6.3.2　学習効果の評価

学習効果の評価には Amazon Mechanical Turk（MTurk） [42] を利用した．MTurk は様々な実験をアウトソースすることができるクラウドソーシングのマーケットプレイスである．

MTurk で応募してきた実験協力者は 46 名であった．アンケートにより，その内 39 名が実験協力者として適正であると判断した．MALSS はデータサイエンティストや機械学習エンジニアのデータ分析を支援することを目的としているため，実験協力者はコンピュータサイエンス系の仕事に従事する者に限定し，さらにプログラミング経験の無い者は実験協力者から除外した．39 名中 17 名の実験協力者はプログラミング経験が 1 年未満の初心者であった．実験のテストの中には品質コントロールのための設問を設け，評価実験にまじめに取り組んでいないと判断される実験協力者は除外している．実験協力者には報酬として 8 ドルを支払った．

実験の手順は以下のとおりである．始めに，実験協力者に対し，クラスタリング分析に関する知識の有無を問うテストを行った．テストの設問は選択式で設問数は 14 問である．

設問の内容は複数の著名な教科書 [40, 43, 25], を参考にし，分析レポートを参照することで正答できるように設計されている．テスト回答にあたっては分析レポート以外の資料等を参照すること，および勘で回答することを禁止した．次に，実験協力者に対し MALSS が作成したクラスタリング分析を行った際に生成された分析レポートを注意深く読むよう指示した．最後に，実験協力者に対し最初と同じテストをもう一度行い，分析レポート参照前後のテストのスコアを比較することで，知識習得の程度を評価した．実験協力者は分析レポート読了後に再び同じテストを行うことを知らされていない．テストの設問を図 6.6 に示す．実際には実験協力者にはこれに品質コントロールのための設問 2 問を加えた設問を提示している．なお，実験協力者の募集は英語で行っており，分析レポートおよびテストの設問は全て英語を用いている．

1. Clustering is the task of _____.
A. reducing the number of random variables under consideration by obtaining a set of principal variables
B. finding the most frequent and relevant patterns in large datasets
C. grouping a set of data samples in such a way that the data samples in the same group are more similar to each other than to those in other groups
D. identifying to which of a set of categories a new observation belongs
E. I don't know.

2. Clustering belongs to _____ learning because the data samples have no labels.
A. supervised, B. unsupervised, C. semi-supervised, D. self-organized, E. I don't know

3. Many clustering algorithms require the user to _____ in advance.
A. set the number of clusters
B. annotate the data samples so that the algorithms can learn to classify information
C. reduce the number of dimensions in the dataset
D. make polynomial features of given degrees
E. I don't know

4. The _____ is an index that estimates the optimal number of clusters.
A. F-measure, B. Jaccard index, C. Hamming distance, D. Gap statistic, E. I don't know

5. If _____ than that of the other attributes, the effect of the attribute is ignored.
A. the scale of an attribute is much larger, B. the scale of an attribute is much smaller, C. I don't know

6. Therefore, _____ is commonly used to normalize the range of attributes of data.
A. standardization, B. feature selection, C. dimensionality reduction, D. average pooling, E. I don't know

7. _____ is commonly used to encode categorical features into numerical features before clustering.
A. Neighbor embedding, B. Feature embedding, C. Target encoding, D. One-hot encoding, E. I don't know

8. _____ clustering is one of the most commonly used clustering algorithms.
A. K-means, B. Nearest neighbor, C. t-SNE, D. Logistic, E. I don't know

9. If data has too many features, clustering algorithms severely degrade their performance due to _____.
A. the no free lunch theorem, B. the ugly duckling theorem, C. the curse of dimensionality, D. Laplace's demon, E. I don't know

10.To address the curse of dimensionality, _____ is one of the effective options.
A. feature selection, B. normalization, C. k-means++, D. multiple imputation, E. I don't know

11.The K-means clustering algorithm assumes that _____.
A. the data has hierarchical structure, B. the clusters are spherical and of equal size, C. all attributes are scaled individually, D. the data has no outliers, E. I don't know

12.Hierarchical clustering algorithms are broadly divided into _____ approaches.
A. linear and non-linear, B. stochastic and deterministic, C. agglomerative and divisive, D. batch and online, E. I don't know

13.The results of hierarchical clustering can be visualized by using a _____.
A. directed graph, B. disjunctive graph, C. dendrogram, D. histogram, E. I don't know

14.Proximity along the horizontal axis of the dendrogram _____.
A. represents the similarity of two observations, B. doesn't represent the similarity of two observations, C. I don't know

図 6.6　知識確認テストの設問.

実験結果を図 6.7 および図 6.8 に示す．図 6.7 は，横軸が分析レポート確認前のテスト正答率を，縦軸が確認後の正答率を示している．ひとつのプロットがひとりの実験協力者のデータを表しており，分析前後の正答率が共に等しいサンプルは僅かにずらしてプロットしている．プロットが点線よりも上に位置していれば，模擬データ分析実験の前後でテストのスコアが向上したことを表している．図 6.8 は，左のバーは全実験協力者のテスト

スコアの平均を，真ん中のバーは 1 年以上のデータ分析経験を有する実験協力者のスコアの平均を，右のバーはデータ分析経験が 1 年未満の実験協力者のスコアを平均を表している．エラーバーは 95% 信頼区間を表している．実験協力者全体のテスト正答率の差について，対応のある母平均の差の t 検定（両側検定）を行ったところ，p 値は 5.82×10^{-7} であった．また，1 年以上のデータ分析経験を有する実験協力者群と，1 年未満の実験協力者群それぞれについて，分析レポート前後の正答率の差について，符号検定による対応のある 2 つの母平均の差の検定（両側検定）を行ったところ，p 値は，前者が 0.523，後者が 2.75×10^{-4} であった．

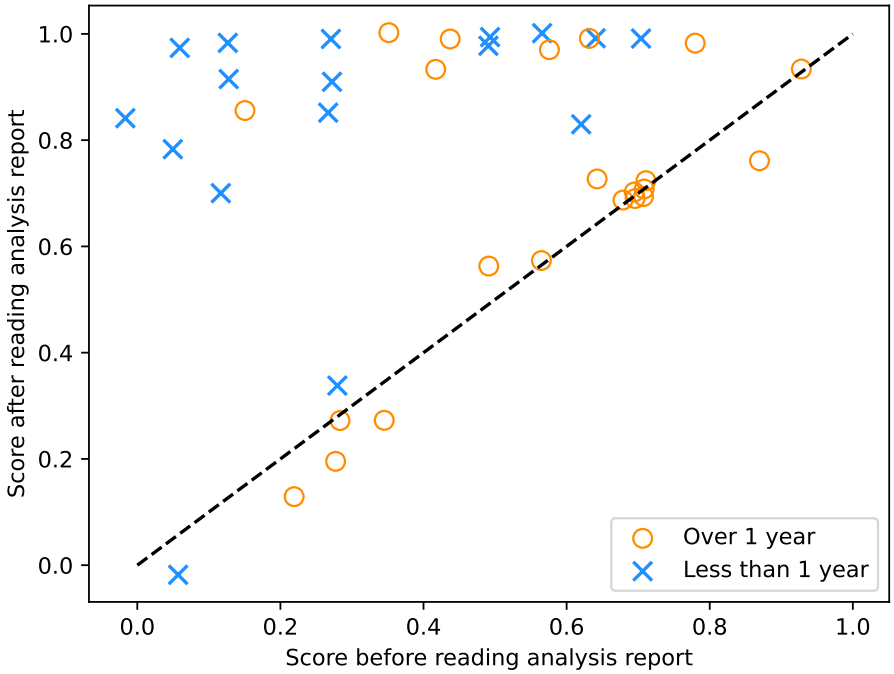


図 6.7　知識確認テストの正答率（散布図）

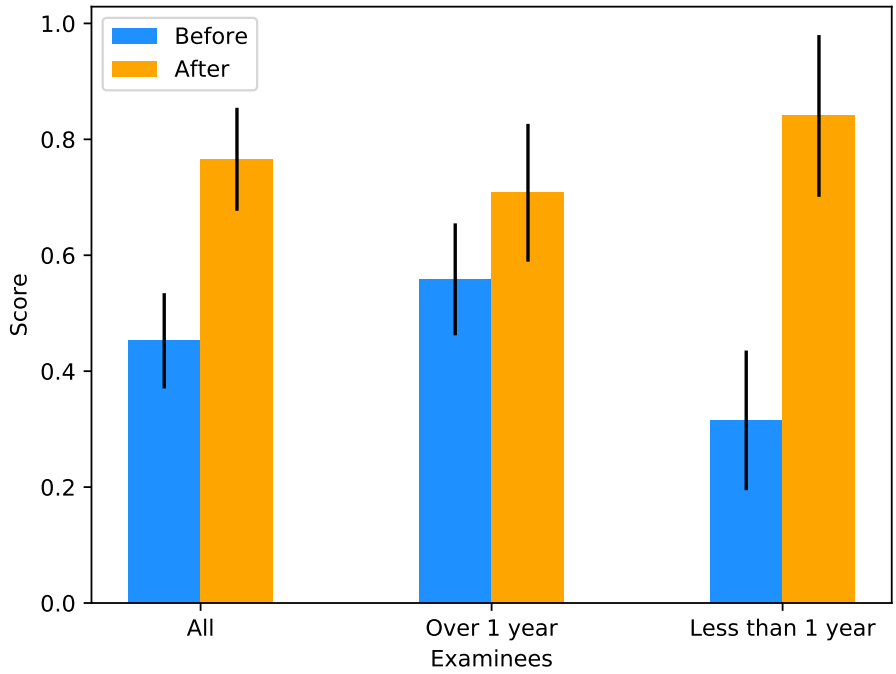


図 6.8　分析レポート参照前後の知識確認テストのスコア（青：レポート参照前，橙：参照後）.

6.4　考察

MALSS を用いることでクラスタリング分析を適切に実施することができるか（RQ1）について，表 6.3 および表 6.4 より，MALSS は 3 つすべてのデータセットに対して実際のクラス数と等しいクラスター数を推定することができており，クラスタリング分析の分析プロセスを自動化し，複数のクラスター数推定指標の多数決により適切なクラスター数推定が可能であることが示された．表 6.4 に示すように，3 つすべてのデータセットのクラスター数を正しく推定することができた単独の指標およびアルゴリズムは存在しない．このことは，複数指標を用いた多数決によるクラスター数推定が推定の頑健性を向上させていることを示している．H2O は 3 つの内 2 つのデータセットのクラスター数を誤って推定している．H2O は単一の指標を用いてクラスター数を推定しており，このため推定性能が低くなっていると考えられる．NbClust の結果を見ると，分析前に特徴量のスケールを標準化することで，推定精度が 1/3 から 2/3 に向上している．これはデータの前処理が重要であることを表している．表 6.4 の精度は全体的に高く，MALSS のクラスタリング分析は推定クラスター数が実際のクラス数と一致しているだけでなく，妥当なクラスタリングができていることが分かる．

次に，分析者が分析レポートを読むことでクラスタリング分析に関する知識を習得することができるか（RQ2）について，図 6.7 および図 6.8 より，分析レポート参照後にスコアが向上しており，実験協力者は分析レポートを参照することによってクラスタリング分析に関する知識を習得できていることが分かる．

図 6.8 をみると，データ分析未習熟者のスコア（右側）の向上分は経験者のスコア（真ん中）の向上分よりも大きい．これは，MALSS が特に未習熟者の知識習得支援に効果的であることを示している．事前知識の少ない実験協力者の方が知識習得効果が高いことは，適応型学習システムにおける Albacete らの研究でも確認されている [44]．これは，経験者のもつ事前知識が，分析レポートから知識を習得することを妨げたのではないかと推察される．経験者は自身の知識を過信し，未習熟者と比較して，分析レポートの内容を真剣に読もうとしなかったのではないかと考えられる．

本研究の結果の妥当性に影響を及ぼす可能性のある事項として，実験参加者による実際のクラスタリング分析を行っていないことが挙げられる．しかし，MALSS はデータ分析プロセス全体を自動化するため，一般的なプログラミングスキルをもつユーザーであれば用意に MALSS を利用した分析を行えることを確認している [32]．実際に分析を行うことが知識習得に及ぼす影響の調査については今後の課題である．

また，図 6.7 をみると，分析レポート確認前後で知識確認テストのスコアが悪化している実験協力者が存在しており，その数は特に経験者の方が顕著に多い．これは，4 章や 5 章では見られなかった結果である．今回，実験協力者をクラウドソーシングを利用して募集しており，実験協力者に報酬を支払っている点が，4 章，5 章のときと条件が異なる．報酬を提示したことにより，特に経験者が報酬目的に参加し，真面目に取り組まなかったケースを，クオリティコントロール設問では適切に排除できなかった可能性が考えられる．

6.5　第 6 章におけるまとめ

本章では，教師あり学習を支援対象としていた MALSS を拡張し，既存の AutoML ツールにおけるサポートが十分でない，クラスタリング分析の実行を支援することを可能とした．拡張した MALSS は，クラスタリング分析の主要な目的の一つである，最適なクラスター数推定を，分析プロセスの自動化により支援する．さらに，MALSS は自動化された分析の後に分析レポートを生成することで，クラスタリング分析の未習熟者がクラスタリング分析に関する知識を習得することを支援する．提案手法の有効性を検証するために，複数のオープンデータセットを用いた模擬データ分析実験と，クラウドソーシング

サービスを利用して募集した実験協力者に対する学習効果確認テストを行った．模擬データ分析実験の結果，提案手法は既存手法と比べても高精度に適切なクラスター数推定が可能であることが確認できた．また，知識確認テストの結果，MALSS が分析終了後に作成する分析レポートは，特に未習熟者がクラスタリング分析に関する知識を習得するのに効果的であることが示された．

第 7 章

提案手法の支援範囲についての考察

本章では，第 4 章，第 5 章，第 6 章で提案した手法により．第 3 章で述べた MALSS のスコープにおいて，何が支援され，何が現状の MALSS では支援の範囲外となるのかを整理する．

7.1 MALSS が支援するスキルの範囲

第 3 章で述べたように，MALSS は，機械学習を用いたデータ分析プロジェクトの実務経験が無い，あるいは不十分であるにも関わらず，データ分析の実務に従事することを求められている分析者を支援することを目的としている．このような分析者に求められるスキルセットについて，データサイエンティスト協会は，データサイエンティストスキルチェックリストを公開している [*1]．

データサイエンティスト協会はデータサイエンティストに求められるスキルを，ビジネス力，データサイエンス力，データエンジニアリング力の 3 つに分けて定義している．MALSS はこのなかで，データサイエンス力の習得を支援することを目的としている．データサイエンス力はさらに，スキルカテゴリ，スキルレベルで細分化されており，スキルレベルは，業界を代表するレベル（シニアデータサイエンティスト），棟梁レベル（フルデータサイエンティスト），独り立ちレベル（アソシエイトデータサイエンティスト），見習いレベル（アシスタントデータサイエンティスト），データサイエンティスト以前に分類されている．MALSS は，データサイエンティスト以前，あるいは見習いレベルとしての知識・経験が不足しているデータサイエンティストが，見習いレベルとして十分な知

[*1] https://www.datascientist.or.jp/common/docs/skillcheck_ver4.00_simple.xlsx（2022/2/10 access）

識を習得するのを支援することも目的としている．

見習いレベルに求められるスキルカテゴリは表 7.1 に示すとおりである．このなかで，3.3.3 項で述べた「データの準備」と「モデリング」のプロセスに関連するのは，「予測」「グルーピング」「データ加工」「時系列分析」「学習」「自然言語処理」「画像・映像認識」「音声認識」「パターン発見」のスキルカテゴリである．さらに，MALSS はテーブルデータ形式を対象とし，教師あり学習と教師なし学習のクラスタリングタスクを支援対象としているため，「予測」「グルーピング」「データ加工」「学習」が MALSS が支援するスキルカテゴリとなる．

表 7.1　見習いレベルに求められるスキルカテゴリ（データサイエンティストスキルチェックリストより著者が作成）．

スキルカテゴリ	対象プロセス	対象タスク
基礎数学		
データの理解・検証		
意味合いの抽出，洞察		
予測	○	○
推定・検定		
グルーピング	○	○
性質・関係性の把握		
サンプリング		
データ加工	○	○
データ可視化		
時系列分析	○	
学習	○	○
自然言語処理	○	
画像・映像認識	○	
音声認識	○	
パターン発見	○	

上述の MALSS が支援するスキルカテゴリにおけるスキルチェック項目は表 7.2 の通りである．チェック項目について，MALSS が自動化により分析遂行を支援できるもの，情報提示により知識習得を支援できるものに○をつけている．3.3.5 項で述べたように，MALSS は深層学習を支援対象外としているため，支援範囲のスキルチェック項目数は 27 個である．このうち，MALSS が分析遂行か知識習得のいずれかにより支援できる項目の数は 19 個である．データサイエンティスト協会では，見習いレベルに求められるスキル

レベルとして，チェック項目の 70% を目安としている．27 項目の 70% は 18.9 項目であり，MALSS は支援可能な範囲において，見習いレベルに求められるスキルの習得を支援することが可能となる．

表 7.2　見習いレベルに求められるスキルチェック項目（データサイエンティストスキルチェックリストより著者が作成）．

スキルカテゴリ（サブカテゴリ）	チェック項目	分析遂行支援	知識習得支援
予測（回帰・分類）	単回帰分析において最小二乗法、回帰係数、標準誤差、決定係数を理解し、モデルを構築できる	○	
予測（回帰・分類）	重回帰分析において偏回帰係数と標準偏回帰係数、重相関係数について説明できる		
予測（回帰・分類）	線形回帰分析は量的な変数を予測し、ロジスティック回帰分析は二値の質的な変数を予測する手法であることを説明できる		○
予測（評価）	ROC 曲線、AUC(Area under the curve)、を用いてモデルの精度を評価できる		
予測（評価）	混同行列（正誤分布のクロス表）、Accuracy、Precision、Recall、F 値、macro 平均、micro 平均、重み付き平均といった評価尺度を理解し、精度を評価できる	○	○
予測（評価）	RMSE、MAE、MAPE、決定係数といった評価尺度を理解し、精度を評価できる	○	○
グルーピング（グルーピング）	教師なし学習のグループ化（クラスター分析）と教師あり学習の分類（判別）モデルの違いを説明できる		○
グルーピング（グルーピング）	階層クラスター分析と非階層クラスター分析の違いを説明できる		○
グルーピング（グルーピング）	階層クラスター分析において、デンドログラムの見方を理解し、適切に解釈できる		○
データ加工（データクレンジング）	外れ値・異常値・欠損値とは何かを理解し、指示のもと適切に検出と除去・変換などの対応ができる	○	○
データ加工（データ加工）	標準化とは何かを理解し、適切に標準化が行える	○	○
データ加工（データ加工）	名義尺度の変数をダミー変数に変換できる	○	○
データ加工（特徴量エンジニアリング）	数値データの特徴量化（二値化／離散化、対数変換、スケーリング／正規化、交互作用特徴量の作成など）を行うことができる	○	○
学習（機械学習）	機械学習にあたる解析手法の名称を 3 つ以上知っており、手法の概要を説明できる		○
学習（機械学習）	機械学習のモデルを使用したことがあり、どのような問題を解決できるか理解している（回帰・分類、クラスター分析の用途など）	○	○
学習（機械学習）	「教師あり学習」「教師なし学習」の違いを理解している		○
学習（機械学習）	過学習とは何か、それがもたらす問題について説明できる		○
学習（機械学習）	次元の呪いとは何か、その問題について説明できる		
学習（機械学習）	教師あり学習におけるアノテーションの必要性を説明できる		
学習（機械学習）	観測されたデータにバイアスが含まれる場合や、学習した予測モデルが少数派のデータをノイズと認識してしまった場合などに、モデルの出力が差別的な振る舞いをしてしまうリスクを理解している		
学習（機械学習）	機械学習における大域的な説明（モデル単位の各変数の寄与度など）と局所的な説明（予測するレコード単位の各変数の寄与度など）の違いを理解している		
学習（機械学習）	ホールドアウト法、交差検証（クロスバリデーション）法の仕組みを理解し、学習データ、パラメータチューニング用の検証データ、テストデータを作成できる	○	○
学習（機械学習）	時系列データの場合は、時間軸で学習データとテストデータに分割する理由を理解している		○
学習（機械学習）	機械学習モデルは、データ構成の変化 (データドリフト) により学習完了後から精度が劣化していくため、運用時は精度をモニタリングする必要があることを理解している		
学習（機械学習）	ニューラルネットワークの基本的な考え方を理解し、出力される「ダイアグラム」の入力層、隠れ層、出力層の概要と、活性化関数の重要性を理解している		
学習（機械学習）	ライブラリを使ってサポートベクターマシンによる分析を実行・評価できる	○	○
学習（機械学習）	決定木をベースとしたアンサンブル学習による分析を、ライブラリを使って実行でき、その結果を正しく解釈できる	○	○
学習（深層学習）	深層学習（ディープラーニング）モデルの活用による主なメリットを理解している（特徴量抽出が可能になるなど）		

7.2 MALSS が支援する分析課題対応

MALSS の開発動機は第 3 章で述べたように，知識や経験が不足した分析者が機械学習を用いたデータ分析を行う際に，誤った分析プロセスにより不適切な分析結果から誤った意思決定をする恐れがあるという実用上の課題を解決することである．そのため MALSS を利用する意義は，必ずしも分析により適切な分析結果が得られることだけではなく，適切な分析結果が得られていないことに気がつけるようになることもそのひとつである．第 4 章から第 6 章では，適切な分析結果を得るための手法について提案してきた．本節では，適切な分析結果が得られない要因を挙げ，それらにおいて提案手法の MALSS がどのように対応できるのか，あるいはどこからは現状の MALSS の支援範囲外となるのかを明らかにする．

MALSS を用いてデータ分析を行った際に適切な分析結果が得られない要因を，タスク，データ，分析に大別し，それぞれについて MALSS の支援状況をまとめると表 7.3 のようになる．

タスク起因の要因としては，行うべき分析タスクが，MALSS が支援する教師あり学習とクラスタリングタスクでないケースが考えられる．3.3 節で述べたように，これらのケースは現在の MALSS では支援対象外としており，分析者は対象外であることに容易に気がつくことができる．

データ起因の要因はさらに，データ種別，データ量，データの質に分けられる．データ種別の要因として，画像や音声，自然言語など，テーブルデータ形式以外のデータ種別は MALSS の支援対象外としており分析者は分析前にそうと気がつくことができる．データ量の要因として，データが適切な分析を行うには少なすぎるケースが考えられる．分析者は，MALSS が生成した学習曲線と，学習曲線の読み取り方を確認することで，作成したモデルの性能が不十分である原因がデータ不足である可能性に気がつくことができる．データの質に起因するものとして，教師あり学習において教師データであるラベルの品質が低い，識別タスクにおいてラベルの偏りが大きい，外れ値と疑われる値が多く存在する，ノイズが大きい，分析に有効な説明変数が不足している，ことが挙げられる．ラベルの品質，偏り，外れ値，ノイズの要因については一般に対応が困難な課題として知られており，常に有効な対策というものは存在せず，パターン化が困難であるため，現在の MALSS では解決を支援することはできない．これらの課題については，分析者がより知識や経験の豊富な分析者と協力し，ドメイン知識なども活用し対応していくことが必要となる．しかし，このようなケースにおいても，学習曲線や複数の評価指標（Recall，Precision 等）を

確認することで，モデルの性能が不十分であることに気がつくことができるため，誤った意思決定をしてしまうことを防ぐことは可能である．説明変数の不足については，データ量の要因と同様に，学習曲線を確認し，作成したモデルがハイバイアスな状態にあることを確認することで，性能が不十分であることの要因として有効な説明変数が不足していることに気がつけるよう，必要な情報を提示している．

分析に起因する要因としては，用いるアルゴリズムの能力が不十分であるケースが挙げられる．3.3 節で述べたように，MALSS は分析の初心者が基本的な分析のプロセスを実施し，基本的な分析プロセスに関する知識を習得することを支援するため，アンサンブル学習や深層学習など，高度で複雑な分析手法を採用していない．したがって，データの性質が複雑であり，これらの表現力の高いアルゴリズム，分析手法が必要なケースには MALSS では対応することができない．しかしこのようなケースにおいても，学習曲線を確認することで，性能不足の要因がアルゴリズムの性能である可能性に気がつくことができるように情報提示を行っている．

以上のように，現状の MALSS は，解決が困難で対応がパターン化されていない課題については支援対象外としている．これらの課題については MALSS 利用の前後で分析者が他者と協力しドメイン知識を活用し解決にあたることを想定しているが，今後技術の発展によりこれらの課題に対する対応がパターン化された場合には，そのパターンを MALSS に取り入れていくことが可能であると考える．また，支援が困難な課題についても，学習曲線を提示し，モデルの状態を把握することで，分析者がモデルの性能が不十分な場合の改善策を挙げられるよう必要な情報を提示するという形で支援を行っている．このように，MALSS を用いることで，知識や経験が不十分な分析者が，基本的な分析のプロセスを遂行するだけでなく，基本的な分析だけではモデルの性能が不十分であることに気がつくことができるようにすることで，分析初心者がその責務を果たすことができると考える．

表 7.3　適切な分析結果が得られない要因と MALSS の対応.

種別	要因		支援方法
タスク	教師あり学習，クラスタリングタスク対応不可		支援対象外であることに気がつくことができる
データ	データ種別	テーブルデータ形式以外	支援対象外であることに気がつくことができる
	量	データ不足	学習曲線を確認し要因に気がつくことができる
	質	ラベルの品質が低い ラベルの偏り 外れ値 ノイズ	学習曲線や評価指標を確認し分析結果が 不十分であることに気がつくことができる 要因の特定，解決にはドメイン知識などが必要
		有効な説明変数の不足	学習曲線を確認し要因に気がつくことができる
分析	アルゴリズムの能力不足		学習曲線を確認し要因に気がつくことができる

第 8 章

結論

8.1 本研究の成果

本研究では，機械学習を用いたデータ分析に関する知識が不十分な分析者の分析業務遂行を支援することを目的として，分析の実行と，分析に関する知識習得を同時に支援する方法の検討を行った．技術の進歩により扱えるデータの量と種類が急速に増加したことで，収集したデータを分析するために機械学習技術が注目を集めている．機械学習技術のニーズが急増したため，特にビジネスの分野では，経験や知識の不足した担当者が分析に従事せざるを得ないことがある．近年はオープンソースソフトウェアのライブラリを利用することで，高度な機械学習アルゴリズムを容易に利用することが可能になっている．しかし，データ分析に関する適切な知識が不足している場合，誤った分析手順により間違った分析結果を得たり，間違った分析結果から誤った意思決定を行う恐れがある．

このような課題に対し，本研究ではデータ分析の知識を習得しながら実際に分析を行うことのできるツール MALSS（Machine Learning Support System）を開発した．MALSS は機械学習を用いたデータ分析のプロセスを自動化することで，知識や経験が不足した分析者でも適切な手順で分析を遂行することを可能とする．しかし，分析の自動化は分析プロセスの中身をブラックボックス化してしまうため，分析者が分析に関する知識を習得し，ツールの支援範囲を超えた発展的な分析を独力で行うことができるようにならない．そこで MALSS では分析中または分析後に，機械学習を用いたデータ分析に関する情報を提供し，分析者の知識習得を支援する．

第 1 章では，上述のように研究の背景を明らかにし，本研究の目的と意義について述べた．

第 2 章では，機械学習を用いたデータ分析の遂行支援と，知識習得支援の従来技術を俯瞰した．

第 3 章では，従来技術の課題について説明し，課題に対する本研究のアプローチを述べるとともに，本研究のスコープを明確にした．本研究では，データサイエンティストの中で特に，データ形成者やデータ分析者とよばれる，機械学習を用いたデータ分析を主要タスクとする分析者をターゲットとし，構造化データであるテーブルデータを入力とする，教師あり学習の回帰タスクと分類（識別）タスク，教師なし学習のクラスタリングタスクを支援する．上記タスクを行うための一連のデータ分析プロセスの中で，プロセスの定型化が比較的容易であり，自動化による支援の効果が高い，特徴量エンジニアリングとプロトタイピングのプロセスを支援対象とする．本研究は，知識や経験が不足した分析者の分析遂行支援と知識習得支援を目的としていることから，支援する技術レベルは基本的なものにフォーカスする．

第 4 章では，分析の自動化と分析レポートによる知識習得支援の提案および有効性の検討を行った．提案手法は教師あり学習をターゲットとし，データの前処理，モデルの学習，評価のプロセスを自動化することで分析遂行を支援する．さらに，分析終了後に分析レポートを生成し分析者に提示することで，分析者の教師あり学習に関する知識習得を支援する．分析レポートには分析結果だけでなく，データ分析プロセスの手順や，各プロセスにおける実行時の注意点についても記載することで，分析者がこれらのプロセスに関する適切な知識を習得することを支援する．

提案手法の有効性を検討するために，模擬データ分析実験と知識確認テストを行った．模擬データ分析実験では実験協力者を MALSS のみを用いてデータ分析を行うグループと，MALSS 以外のツールおよびデータ分析に関するテキストを用いて分析を行うグループとに分け，回帰モデルを作成する分析を行い，作成したモデルの性能を評価した．さらに，模擬データ分析実験の前後に知識確認テストを行い，テストの点数の増加量から分析に伴う知識習得度合いを評価した．模擬データ分析実験の結果，MALSS を利用したグループは比較対象グループよりも性能の良いモデルを作成することができ，MALSS を利用することで適切なデータ分析を実施可能であることを確認できた．また，知識確認テストの結果では，模擬データ分析実験前後でテスト正答率が向上しており，MALSS を利用したデータ分析を通して，機械学習を用いたデータ分析に関する知識を習得できていることを確認できた．このように本提案手法を用いることで，経験や知識の不足した分析者であっても，分析に関する知識を習得しながら，適切な手順で分析を遂行することが可能と

なる．

第 5 章では，グラフィカルユーザーインタフェース（GUI）を用いて，データ分析プロセスの途中結果に応じて分析実施内容を変更する手法の提案および有効性の検討を行った．分析自動化と分析レポートによる分析支援では，分析プロセス全体を自動化するため，分析プロセスの途中結果に応じて分析内容を変更することが困難であるという課題があった．そこで提案手法では，分析のサブプロセスごとの実行を可能とし，GUI を介して，サブプロセスの実行結果に応じて分析者の入力を受け付け，プロセスの実施内容を変更可能なように MALSS を拡張した．提案手法では分析完了後の分析レポートではなく，分析の途中に GUI 上に必要な情報を提示することで，分析者の知識習得を支援する．

本章においても，提案手法の有効性を検証するために，第 3 章と同様の模擬データ分析実験と知識確認テストを行った．模擬データ分析実験の結果，提案手法の MALSS を用いたグループの作成したモデルの性能は，第 3 章で提案した MALSS を用いたグループの作成したモデルの性能よりも良く，GUI を介して分析プロセスの途中結果に応じて分析内容を変更することで，より適切な分析支援が可能なことを確認した．また，知識確認テストにより，GUI を用いた分析支援により，分析レポートと少なくとも同等の知識習得が可能であることを確認した．

第 6 章では，提案手法の支援範囲を教師なし学習のクラスタリング分析に拡張した．機械学習を用いたデータ分析プロセスを自動化する AutoML は様々な OSS ツールが提案されているが，その多くが教師あり学習にフォーカスしている．提案手法では，教師なし学習の代表的なタスクの 1 つである，クラスタリング分析を支援する．クラスタリング分析の主要な目的の一つである，最適なクラスター数推定を，分析プロセスの自動化により支援し，分析レポートによりクラスタリング分析に関する知識習得を支援する．

提案手法の有効性を検証するために模擬データ分析実験と，クラウドソーシングを利用した知識確認テストを行った結果，提案手法は既存手法と比較しても高精度にクラスター数推定が可能であることを確認し，分析レポートは，特に知識の不足した分析者の知識習得に効果的であることを確認した．

第 7 章では，第 4 章から第 6 章までの提案を踏まえ，第 3 章で述べた MALSS のスコープにおいて，何が支援され，何が現状の MALSS では支援の範囲外となるのかを整理した．

MALSS はデータサイエンティスト協会が提案するデータサイエンティストのスキル

チェックリストにおいて，見習いデータサイエンティストに求められるデータサイエンス力に関する分析遂行および分析に関する知識習得を支援することができる．また，MALSS は，MALSS を用いてデータ分析を行うことで適切な分析結果が得られるよう支援するだけでなく，適切な分析結果が得られないケースにおいても，分析結果が不十分であることに気がつくことができるよう支援することによって，分析者が与えられた責務を果たすことを支援することができる．

以上のことにより，提案した，機械学習を用いたデータ分析と分析に関する知識習得を同時に支援するツール MALSS を用いることで，経験や知識の不足した分析者が，分析に関する知識を習得しながら，適切な手順で分析を遂行することを支援することが可能になるといえる．

8.2 今後の課題と展望

本研究を通して得られた課題としては，第一に分析支援範囲の拡充が挙げられる．本研究では，教師あり学習における回帰と分類タスク，および教師無し学習におけるクラスタリングタスクの支援方法について検討を行った．しかし，教師あり学習においては画像や音声，言語など非構造化データの入力に対しては適用範囲外である．また，教師なし学習においては，異常検知や頻出パタンマイニングなど，クラスタリング以外のタスクに対応していない．

画像や音声，言語などのメディア処理については，同じ分類タスクであっても，例えば画像処理であれば物体認識，物体検出，シーン認識のように様々なタスクが存在し，これらを広く支援対象とすることは容易ではない．また，後述するようにこれらの分野では昨今深層学習を用いたアプローチが大きな成果を上げている．深層学習については現在黎明期にあたり日進月歩で新しい技術が提案されており，どこまでを初学者の支援範囲とすべきかの決定が難しいという問題も存在する．

教師なし学習タスクに対する支援については，教師あり学習と異なり正解となるデータをもたない教師なし学習においては，分析結果を分析者自身が解釈する必要があるため，分析プロセスのなかの何をどのように支援すべきかをよく検討する必要がある．クラスタリング分析においては，主要な分析目的の一つである，クラスター数の推定を支援対象とした．その他の教師なし学習タスクにおいても有効な支援対象を検討していく必要がある．

第二の課題は習得を支援する知識レベルの設定についてである．これには 3 つの側面が存在する．

1 つ目は，初心者が習得すべき知識範囲の変化への対応である．近年，深層学習技術の発展により機械学習分野は目覚ましい進展を遂げており，毎日のように革新的な技術が提案されている．これらの技術はその有効性が確認されるとすぐさま OSS のライブラリに追加される．そのため，初心者であっても扱えると期待される技術の範囲は日々拡大している状況である．MALSS は現在，初心者が基本的な機械学習に関する知識を習得することの支援を目的としており，深層学習を用いた分析は支援対象外としている．しかし今後は，ビジネスの現場において初心者に期待される技術範囲に合わせて支援範囲を見直していく必要がある．

2 つ目は，分析者の知識水準に応じた知識習得支援手法の検討である．本研究の評価実験を通じて，一口に分析初心者といってもその知識水準には大きな差があることが明らかになった．分析者の知識水準が異なれば，適切な知識習得支援のために提示すべき情報も異なるため，幅広い知識水準に応じた知識習得支援手法の検討が必要となる．

3 つ目は，分析者個人の成長に合わせた支援方法の検討である．本研究により MALSS を利用することで分析者の分析に関する知識習得を支援できることを明らかにした．上述のように，知識水準が変化すれば，適した知識習得支援内容も変化するため，MALSS を繰り返し利用することで変化した知識水準に応じた知識習得支援を行うことが望ましい．

参考文献

[1] Khalid Salama, Jarek Kazmierczak, and Donna Schut. Practitioners guid to mlops: A freamework for continuous delivery and automation of machine learning. *Google Could White paper*, 2021.

[2] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. Data scientists in software teams: State of the art and challenges. *IEEE Transactions on Software Engineering*, 44(11):1024–1038, 2018.

[3] McKinsey Global Institute. Big data: The next frontier for innovation, competition, and productivity. Technical report, McKinsey Global Institute, 2011.

[4] トーマス H. ダベンポート. データ・サイエンティストほど素敵な仕事はない. In **ダイヤモンド・ハーバード・ビジネス・レビュー**, pages 84–95. ダイヤモンド社, 2013.

[5] Harlan Harris, Sean Murphy, and Marck Vaisman. *Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work*. O'Reilly Media, Inc., 2013.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[7] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas,

Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[8] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. Efficient and robust automated machine learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2962–2970. Curran Associates, Inc., 2015.

[9] Randal S. Olson, Ryan J. Urbanowicz, Peter C. Andrews, Nicole A. Lavender, La Creis Kidd, and Jason H. Moore. *Applications of Evolutionary Computation: 19th European Conference, EvoApplications 2016, Porto, Portugal, March 30 – April 1, 2016, Proceedings, Part I*, chapter Automating Biomedical Data Science Through Tree-Based Pipeline Optimization, pages 123–137. Springer International Publishing, 2016.

[10] H2O.ai. *H2O*, 5 2019. H2O version 3.24.0.2.

[11] T. Swearingen, W. Drevo, B. Cyphers, A. Cuesta-Infante, A. Ross, and K. Veeramachaneni. ATM: A distributed, collaborative, scalable system for automated machine learning. In *2017 IEEE International Conference on Big Data*, pages 151–162, Dec 2017.

[12] Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, and Bernd Wiswedel. Knime: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer, 2007.

[13] Janez Demšar, Tomaž Curk, Aleš Erjavec, Črt Gorup, Tomaž Hočevar, Mitar Milutinovič, Martin Možina, Matija Polajnar, Marko Toplak, Anže Starič, Miha Štajdohar, Lan Umek, Lan Žagar, Jure Žbontar, Marinka Žitnik, and Blaž Zupan. Orange: Data mining toolbox in python. *Journal of Machine Learning Research*, 14:2349–2353, 2013.

[14] データサイエンティスト協会. 設立の背景と目的. http://www.datascientist.or.jp/about/. 参照 Feb. 21,2015.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, page

1097–1105. Curran Associates Inc., 2012.

[16] B. F. Skinner. Teaching machines. *Science*, 128(3330):969–977, 1958.

[17] Nuri Kara and Nese Sevim. Adaptive learning systems: Beyond teaching machines. *Contemporary Educational Technology*, 4(2):108–120, 2013.

[18] Lijia Chen, Pingping Chen, and Zhijian Lin. Artificial intelligence in education: A review. *IEEE Access*, 8:75264–75278, 2020.

[19] 新井 稔也 and 山田 和宏. パナソニックにおける AI 人材強化. *Panasonic Technical Journal*, 64, 5 2018.

[20] Colin Shearer. The crisp-dm model: The new blueprint for data mining. *Journal of Data Warehousing*, 5(4):13–22, 2000.

[21] 国立研究開発法人産業技術総合研究所. 機械学習品質マネジメントガイドライン. **サイバーフィジカルセキュリティ研究センターテクニカルレポート**, **訂補** *2* **版**, 2021.

[22] 大城信晃, マスクド・アナライズ, 伊藤徹郎, 小西哲平, 西原成輝, and 油井志郎. ***AI・データ分析プロジェクトのすべて［ビジネス力×技術力＝価値創出］***. 技術評論社, 2020.

[23] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. **深層学習**. KADOKAWA, 2018.

[24] Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth Publishing Company, Belmont, California, U.S.A., 1984.

[25] Kevin P Murphy. *Machine learning: a probabilistic perspective*. The MIT Press, Cambridge, MA, 2012.

[26] Wes McKinney. Data structures for statistical computing in python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56, 2010.

[27] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer New York Inc., New York, NY, USA, 2001.

[28] Willi Richert and Luis Pedro Coelho. **実践 機械学習システム**. O'Reilly Japan, Inc., 2014.

[29] Jasjeet S. Sekhon. Multivariate and propensity score matching software with automated balance optimization: The matching package for r. *Journal of Statistical Software*, 42(7):1–52, 2011.

[30] Elena Verdú, Luisa M. Regueras, María Jesús Verdú, Juan Pablo De Castro,

and María Ángeles Pérez. An analysis of the research on adaptive learning: The next generation of e-learning. *WSEAS Transactions on Information Science and Applications*, 5(6):859–868, 2008.

[31] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001.

[32] 鴨志田 亮太 and 坂本 一憲. MALSS：未習熟者の機械学習によるデータ分析を支援するツール. **電子情報通信学会論文誌** *D*, J99-D(4):428–438, 2016.

[33] Malika Charrad, Nadia Ghazzali, Veronique Boiteau, and Azam Niknafs. NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software, Articles*, 61(6):1–36, 2014.

[34] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987.

[35] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, April 1979.

[36] T. Caliński and J Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974.

[37] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[38] Stuart P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

[39] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif., 1967. University of California Press.

[40] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*, chapter 14.3, pages 501–528. Springer, 2 edition, 2009.

[41] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern,

Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake Vand erPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1. 0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[42] Amazon. *Amazon Mechanical Turk*, 2005. accessed 31 March 2020.

[43] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer, 2014.

[44] Patricia L Albacete and Kurt A VanLehn. Evaluating the effectiveness of a cognitive tutor for fundamental physics concepts. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 22, 2000.

謝辞

本論文は，執筆者が国立情報学研究所トップエスイープロジェクトおよび電気通信大学で行った研究をまとめたものです．本研究を遂行するにあたり，数多くの方々にご指導とご援助を賜りました．特にお世話になった方々をここに記し，深い感謝の意を表します．

本研究全般にわたり，主任指導教員として研究のまとめ方の方向性から論文構成の細部まできめ細かくご指導いただいた電気通信大学大学院情報理工学研究科　石川冬樹客員准教授に深く感謝申し上げます．

早稲田大学グリーン・コンピューティング・システム研究機構 坂本一憲客員准教授には，トップエスイープロジェクトにおいて，指導員として本研究テーマの立ち上げに際し多くのご指導を頂いたことを深く感謝申し上げます，

さらに，本論文をまとめるにあたり，全体の構成検討や記載内容の拡充のために有益なご指導を頂いた電気通信大学大学院情報理工学研究科　大須賀昭彦教授，電気通信大学大学院情報システム学研究科　田中健次教授，電気通信大学大学院情報理工学研究科　柏原昭博教授，西野哲朗教授に深く感謝いたします．

また，本研究を遂行するにあたり，評価実験に快くご協力頂いた皆様に厚く御礼申し上げます．

最後に，執筆者の研究活動および社会人学生生活を精神的に支えてくれた家族に心から感謝します．

上記以外にも，執筆者の研究活動を支えていただいた全ての方々に感謝いたします．ありがとうございました．

関連論文

1. **鴨志田 亮太**, 坂本 一憲: ”MALSS：未習熟者の機械学習によるデータ分析を支援するツール”, 電子情報通信学会論文誌 D, J99-D(4):428–438, 2016. （第 4 章に関連）
2. **鴨志田 亮太**, 石川 冬樹, ”教師あり機械学習の実行と分析者の知識習得を同時に支援するツールの提案”, コンピュータソフトウェア, 採録決定済み, 2021 年. （第 5 章に関連）
3. **Ryota Kamoshida**, Fuyuki Ishikawa: ”Automated Clustering and Knowledge Acquisition Support for Beginners”, Procedia Computer Science, Vol. 176(2020), pp.1596–1605. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES2020. （第 6 章に関連）

著者略歴

鴨志田 亮太（かもしだ りょうた）

2001 年早稲田大学電気電子情報工学科卒．2003 年同大大学院理工学研究科修士課程修了．2003 年日立製作所 中央研究所入所．2018 年電気通信大学大学院情報理工学研究科入学．2021 年電気通信大学大学院情報理工学研究科単位取得退学.