修士論文

分類可能性予測システム

公立はこだて未来大学大学院　システム情報科学研究科
情報アーキテクチャ領域

早川 雄登

指導教員　　新美 礼彦

提出日　　2022 年 3 月 15 日

Master's Thesis

Prediction System of Classificatability

by

Yuto Hayakawa

Graduate School of Systems Information Science, Future University Hakodate
Media Architecture Field
Supervisor:　　Ayahiko Niimi

Submitted on March 15, 2022

Abstract–

Recently, data mining is being used in various domains. However, data mining has incurred an additional cost. This cost input is an investment in the knowledge gained from the data; however, the problem is that acquiring knowledge through data mining is uncertain. Data mining processes still require experts' knowledge. Therefore, this problem is more pronounced, especially when nonexperts want to perform data mining. In this study, we find the expectation that knowledge can be obtained from an unknown dataset. We defined classificatability as a measure of expectation in the analysis of unknown datasets as well as proposed and discussed methods for predicting classificatability using dataset meta-features. We also discussed the construction of a classificatability prediction system to obtain the classification performance of a classifier constructed using an unknown dataset. This study introduces the background of this research, current auto-machine learning technology, meta-learning, and knowledge evaluation metrics. Thereafter, it defines classificatability and proposes a method for predicting classificatability using meta-features. Because the training data and the data to be predicted in classificatability prediction are metadata sets, and each instance refers to a data, the predictive performance of classificatability prediction is significantly affected by the relationship between the training dataset and the data to be predicted. In this study, we present a method for better prediction using the minimum distance between data groups as a metric to study the effect and similarity between the training data group and the target data on the meta-feature space on classificatability prediction performance. Additionally, we discussed possible problems related to the format of the data to be predicted and the size of each dataset when the optimal dataset is prepared. The classificatability prediction proposed in this study can predict classificatability, and high prediction performance can be achieved by constructing a classificatability predictor using a group of training data with a minimum distance between the group and the target data.

Keywords: Data Mining, Classification, Preprocessing, Meta-Features, Classificatability

概 要：

データマイニングは近年様々な領域で利用されているが，データマイニングを行うために追加のコストが生じる．このコスト投入はデータから得られる知識への投資であるが，データマイニングによる知識の獲得は確実なものではないという問題がある．データマイニングのプロセスには依然として専門家の知識が不可欠である．そのため，特に非専門家がデータマイニングを行おうと考えた場合はこの問題がより顕著に現れる．そこで本研究では，未知のデータセットから知識が得られる期待度を求めることを目指した．私はデータマイニングの中でも分類タスクを対象に，未知のデータセットの分析における期待度を指す指標として分類可能性を定義し，データセットのメタ特徴からそれを予測する方法について提案した．そして，未知のデータセットを用いて分類器を構築した際の分類性能を簡易的に求める分類可能性予測システムの構築に関する議論をおこなった．本論文では，この研究に至った背景や現状の Auto Machine Learning 技術およびメタ学習に関する研究，知識評価指標について触れた上で分類可能性を定義し，メタ特徴を用いた分類可能性の予測方法を提案する．分類可能性予測における学習データおよび予測対象データはメタデータセットであり，その各事例はそれぞれデータセットを指すことから，分類可能性予測の予測性能は学習データ群と予測対象データとの関係性に大きな影響を受ける．さらに，最適なデータ群を用意したときに考えられる問題として予測対象データの形式に関わるものと各データセットのサイズに関するものを挙げ，その対策についても議論した．本論文で提案した分類可能性予測は分類可能性を予測することが可能であり，予測対象データとの群間最小距離が小さい学習データ群により分類可能性予測器を構築することで高い予測性能が実現可能である．

キーワード： データマイニング, 分類, 前処理, メタ特徴, 分類可能性

目次

第 1 章

序論

1.1 背景

近年，様々な領域でデータマイニングが活用されている．データマイニングは統計学に代表されるデータ解析を行う学問領域に属しているが，その活用範囲はこれらの学問領域の範疇を超えている．そのため，データマイニングはその専門家のみならず一般的な解析の方法として用いられていると考えられる．

特に近年の Digital Transformation (DX) 推進はデータマイニングの幅広い分野における活用に寄与していると考えられる．DX の推進により，従来はデータとして得られなかった事象がデータとして得られるようになり，それを分析する方法としてデータマイニングが用いられる．ここではその活用例についていくつか紹介する．

まず，金融工学における活用について紹介する．和泉らによると大規模データ解析による市場取引の高度化やブロックチェーン技術による決済のディジタル化などを含むフィンテックの潮流により金融分野と情報技術は強く結びついており，近年の利用動向や今後期待される利用方法について特集している [1]．この特集における金融分野のデータ解析は主にデータマイニングを指しており，金融分野においてデータマイニングはその発展に大きく寄与していると考えられる．

次に，図書館情報学における活用について紹介する．清田らによると「知識の利用と共有」,「言語メディア処理」,「Web インテリジェンス」の諸分野を研究対象に含む学問分野であるという点で AI 研究と深い関連があるとし，機械学習や知識ベースといった AI に関する分野を中心に図書館情報学における AI 研究との関係性について特集している [2].

しかしこのような状況に対し，データマイニングを行うために追加のコストが生じるという問題が考えられる．その理由として，データマイニングを行う際には分析対象のデータドメインに関する知識のほかに，データサイエンスに関する知識が必要となることが挙げられる．これは蓄積したデータを効率良く扱うためだけではなく，そもそもその知識がなければどのようにデータを収集しどう扱えばいいのかがわからないという事態に陥ることが想定されることに起因している．従って，その知識を有していない主体がデータマイニングを行

う際には，データサイエンスの知識を身につけた人材，いわゆるデータサイエンティストの育成や，データマイニングを行う外部組織へ委託する必要がある．

また，データマイニングにより期待された新たな知識が得られないことも考えられる．データマイニングによる知識発見プロセスは，その前提に分析対象データにその知識が含まれていることが求められる．しかし，データマイニングの対象データは，その性質が未知であることから，そこに知識が含まれているかを事前に調べることが困難である．

ここで，データマイニングに対するコスト投入が持つ意味について考える．このコスト投入はデータマイニングにより得られる知識に対する投資であると言える．しかし，前述の通り，データマイニングに対しかけたコストに見合う知識は必ずしも得られるとは限らず，分析できると考えられるデータに対しコスト投入を行ったとしてもそこに知識が含まれていなければ抽出できない．そのため，この投資における期待度を事前に予測する仕組みの実現は，このコスト運用の効率化に繋がると考えられ，ひいては更なるデータマイニングの活用も期待できる．

私は簡易的なデータマイニングシステムにより，分析対象のデータセットの分析に対する期待度を算出することにより，実際のコスト投入を行う際の参考になるのではないかと考えた．またその期待度の算出は，実際に分析する際のコストと比べてかなり安価に実現可能ならば望ましいだろう．

そこで，本研究ではデータマイニングタスクの中でも分類タスクに着目し，分析対象のデータセットのメタ特徴から分類タスクを適用した際の分類性能の期待度を求めることを目標とする．

1.2 目的

本研究の目標は，分類問題であると考えられる未知のデータセットを対象に，そのデータセットを用いて分類器を構築した際にどの程度の分類性能を示すかを簡易的に調べる方法を提案することである．

一般に，分類問題とはあるデータ列（以下，説明属性とする）からその事例がいずれかのクラスに属するかを予測する問題であり，説明属性を $\mathbb{X} = \{x_1, x_2, \cdots, x_n\}$（$n$ は説明属性の数），予測対象のクラス（以下，クラス属性とする）を y，またその予測値を $\hat{y}$ として，分類器 $\mathcal{M}$ を用いて $\hat{y} = \mathcal{M}(\mathbb{X})$ と予測する問題を指す．ここで，予測値 $\hat{y}$ と真値 y がデータセット内でどの程度一致しているかを分類性能と呼ぶ．

本研究ではこの目標に対し，あるデータセットに対し任意のアルゴリズムにより構築された分類器が示す性能がどの程度であるかを，そのデータセットに対する分類タスクにおける期待度として分類可能性を定義し，それをメタ特徴により予測することを目的とする．分類可能性は第 3 章で述べる通り，あるデータセットについて 7 つのアルゴリズムで構築された分類器において 5 つの指標を計算し，それぞれの指標における最大値がある閾値 θ を超えているか否かの 2 値の指標である．この指標の予測はそのデータセットのメタ特徴から行い，

詳細は第 4 章で述べる．

1.3 用語の整理

この節では，本論文における各用語の整理とその説明を行う．

本研究で “データセット” と記載されるものは大きく分けて 2 つある．1 つ目は分類可能性予測の対象や分類可能性予測器を構築する際の個々のデータセットである．特筆がなければ “データセット” という記載はこれを示す．2 つ目は分類可能性予測器の入力である，個々のデータセットのメタ特徴および分類可能性をまとめたデータセットである．これらは “メタ特徴・分類可能性データセット” と記載し，“メタ特徴・分類可能性データセット” の各事例は前述の “データセット” およびそのメタ特徴や分類可能性となる．また 1 つ目の “データセット” は本研究においていくつかをまとめたものとして扱うことが多いため，“データ群” と記載されることがある．その場合，“データ群” に含まれる個々のデータは “データセット” であり，それらから抽出されたメタ特徴と分類可能性を “メタ特徴・分類可能性データセット” である．なお，“データ” と記載するものは “データセット” を指しており，本論文で登場する “実データ” および “人工データ”，“半データ” は全て各 “データセット” にその出所を修飾した語である．

本研究における “予測性能” はあるデータセットについて構築された分類器がどの程度正確に予測することが可能であるかを一般に指す語である．分類タスクにおける “予測性能” は様々なものが提案されているが，それらについては第 3.1 節でいくつか紹介する．本研究において “データセット” に対する “予測性能” は “分類器評価指標” と記載しており，“分類可能性” の定義以後は “分類可能性” として利用している．そのため，第 4 章以後に記載される “予測性能” は特筆がなければ “分類可能性予測器” のものを指す．

“分類可能性” は第 3 章で定義するある “データセット” に対する指標である．またそれを予測することを “分類可能性予測” と記載し，予測するモデルは “分類可能性予測器” とする．また，それに関係する “モデル” は “予測器” を指す語として用いているため，“半データ学習モデル” は “半データ” を用いた “分類可能性予測器” である．“半データ” および “半データ学習モデル” については第 5.3 節で述べる．

“メタ特徴” は “データセット” の性質を示す特徴量である．一般な “メタ特徴” の利用については第 2.2 節で述べ，本研究で用いる “メタ特徴” は第 4.2 節で述べる．また，“メタ特徴” を用いた学習は “メタ学習” と呼称されるが，本研究における “メタ学習” は “分類可能性予測器” の構築であるため，本論文では “分類可能性予測器の構築” と表現する．

1.4 章構成

本論文は 7 章構成であり，その内訳は以下の通りである．

第 1 章では，本研究の背景を述べ，その背景に対して本研究で取り組んだ目的について述

べる．その後，本研究における 4 つの語についてまとめ，本論文がどのような構成であるかを示す．

第 2 章では，本研究に関連する研究や技術について 3 つの視点に基づいてまとめ，それらの中で本研究がどのような位置付けにあるかを述べる．

第 3 章では，本研究の 1 つ目の提案である分類可能性の定義を行う．分類可能性は既存の分類器評価指標の組み合わせから成っており，それらをどのように組み合わせるかを定義し，分類可能性を用いることに関する意義について述べる．

第 4 章では，本研究の 2 つ目の提案である分類可能性予測の定義を行う．まず第 3 章で定義した分類可能性を予測するシステムの概観を示し，分類可能性をデータセットのメタ特徴から予測する方法を述べる．そして，分類可能性予測器の構築方法を述べた後，実際に人工データ群および実データ群を用いた分類可能性予測の例を示す．

第 5 章では，分類可能性予測器が学習データ群から受ける影響について議論する．分類可能性予測器の構築に用いるデータ群と分類可能性予測の対象データとの間の関係性から分類可能性の予測性能は大きな影響を受けるため，学習データ群と予測対象データとの間の群間最小距離を用いて分類可能性予測を行う際の関係性の目安を示す方法について述べる．

第 6 章では，分類可能性予測システムにおける諸問題について 2 つの観点から議論する．分類可能性予測で得られる予測値はある閾値における分類可能性であるが，その予測値をどのように使うことで効果的に利用できるかということについて議論する．また，分類可能性予測システムでは対処できない問題も述べ，その対処法についても議論する．

第 7 章では，本研究で行ったことをまとめ，分類可能性予測システムを実用化するために今後どのような取り組みが必要であるかを今後の展望として示す．

第 2 章

関連研究

2.1 本研究の位置付け

この章では本研究の目標や用いた手法に関連する研究について 3 つの観点から紹介する．

1 つ目の観点は Auto ML 技術である．本研究の目標は未知データを分類問題として分析する際にどの程度の期待度が持てるかを求めるということであるが，データサイエンスに関する専門知識の要求や分析作業を低減するものとして Auto ML 技術が存在している．第 2.2 節では本研究と既存の Auto ML 技術を比較し，それらの技術では本研究の目的を満足に達成できない理由について述べる．

2 つ目の観点は知識評価である．本研究ではデータセットの分類タスクに対する期待度として分類可能性を定義し用いるが，これはそのデータセットを用いて構築したモデルの性能を元に議論している．学習モデルを含む知識の性能を評価する指標はいくつか存在しているが，第 2.3 節ではそれらについて簡単に紹介し，既存の知識評価指標と分類可能性の対比を述べる．

3 つ目の観点はメタ学習である．本研究ではメタ特徴を用いて分類可能性予測を行うが，メタ特徴を用いた学習を含むメタ学習という分野が存在している．第 2.4 節ではメタ学習に関する概観を示した上で，その中での本研究の位置付けを示す．

2.2 AutoML に関する研究

第 1 章で述べた通り，近年データマイニングの需要が高まっている．この需要の高まりに対し，Auto Machine Learning (AutoML) 技術の研究・開発が盛んである．

いくつか例を挙げると，オープンソースのものであれば Auto-WEKA[3] や Auto-Sklearn[4]，企業が開発・提供しているものであれば Google AutoML Tables[5] や IBM AutoAI[6] といったものがある．これらの AutoML フレームワークを用いると，データマイニングの処理をある程度自動化することが可能となるが，その後の結果の取り扱いについてはデータマイニングの知識が必要であると考えられる．また，企業が開発・提供しているも

のは利用するためのコストが必要であり，やはり対象としているデータの価値がそのコストを上回っていることを確認する意義があると考えられる．

また，統計学において非専門家でも扱えるツールを開発する研究がある．Jun らは研究対象のドメインの知識を持っていても統計に関する知識のないユーザを対象とした，統計的テストの選択と実行を自動化するドメイン固有言語とランタイムシステムである Tea を提案，開発した [7]．その研究の背景には，統計学において，膨大な量的手法が存在しているため，特定の問題にどの統計検定を使用すべきか特定することの難しさが認識されているという問題がある．さらに，この問題は統計の知識を持たない人にとっても統計的手法が一般的な作業となっていることで深刻化しているとされている．そこで，Tea は統計的検定の選択と実行を自動化し，最小限のプログラミング経験しかないユーザでも一般的なデータ分析ワークフローに直接統合できるように設計されている．

Tea は統計学を対象としたものであるが，同様の状況がデータマイニングにおいても生じている．その状況に対し，Amazon Web Services (AWS) は，プログラムを記述することなく直感的に学習モデルを構築できる “Sagemaker Canvas” を発表した [8]．このサービスでは，データのクリーニングと結合，内部での数百のモデルの作成，最もパフォーマンスの高いモデルの選択，新しい個別予測またはバッチ予測の生成を，機械学習の専門知識なしに行えるとされており，本研究の目的に合致していると考えられる．しかし，前述の通り企業が開発・提供しているものは利用するためのコストが必要であることに加え，このサービスにおいても予測するに足るデータが用意されていなければ予測することが困難であり，その原因がデータにあるかどうかということを確かめることは困難であると考えられる．

これらの AutoML ソフトウェアやサービスに対し，分類可能性予測は予測対象のデータセットが分類予測により予測が可能であるかという目安を提示することで利用者にこれらのサービスの利用を促進させる使い方もできると考えられる．その場合のユースケースは，まず利用者は分類可能性予測システムに対し予測対象データを与え，得られた分類可能性が良好なものであればこれらのサービスを使う段階に進み，分類可能性が良好なものでなければ更なるデータの追加や要因の分析等に進むことが考えられる．

2.3 知識評価に関する研究

学習器，特に分類器における性能の評価指標はいくつか存在している．それらについては分類可能性の定義に関係しているため第 3.1 節で述べることとする．これらはモデルのデータへの適合度を評価する代表的な指標であり，主に分類モデルに対しどの程度望ましい分類が行われているかを判断する指標として用いられている．

また，モデル選択基準という指標が存在する．それらはなんらかのモデルのパラメタ推定と構築の後に予測を行う際に，複数の候補となるモデルの中からどのモデルを用いることが最適であるかを調べるための指標である．代表的なものとして，赤池情報量基準（Akaike Information Criterion, AIC）やベイズ情報量基準（Bayesian Information Criterion, BIC），

最小記述長（Minimum Description Length，MDL）といったものが挙げられ，これらはモデルの複雑さとデータ適合度とのバランスを測る指標であることから，なるべく少ないパラメタでより良い適合度を実現するモデルを選択するために用いられる．

これらの指標は特定のモデルやそのパラメタとデータの相性を判断する指標であると言えるが，それは検証に用いたある分類器構築アルゴリズムとあるハイパーパラメタを用いて構築されたモデルで上手く分析できたか否かを示す指標であると考えられる．ここで，本研究の目的を考えると，検証に用いたある条件においての適合度ではなく，なんらかの方法を用いてそのデータを分析した際にどの程度の適合度を見込めるかを示す必要がある．そのため，これらの既存の知識評価指標ではこの目的に適していない．また，モデル選択基準ではあるデータセットに対して構築されたモデルの良し悪しをモデル間で相対的に比較することは可能であるが，その絶対的な良し悪しを考えることは困難である．

分類可能性はあるデータセットに対しいくつかのアルゴリズムにより構築した分類器で得られた性能から最も高いものを選択する．それにより，単にそのデータセットに対し特定のアルゴリズムとパラメタを利用して構築した際の分類器の性能を用いた場合と比較して，分類タスクにおいてどの程度の性能が見込めるかを広範に示すことができると考えられる．

2.4 メタ学習に関する研究

メタ学習（Meta-Learning）は Learning to Learn とも呼ばれ，様々な機械学習アプローチが広範な学習タスクでどのように動作するかを体系的に観察し，このメタ的なデータから学習することで新たなタスクを他の方法よりも効率的に学習することができるようになる学問領域および技術である [9]．Weng は彼女の記事 [10] の中で，特に画像処理を対象とした深層学習におけるメタ学習について，Vinyals が示したメタ学習の一般的なアプローチの分類を表 2.1 のように示した上で具体的なアルゴリズムをいくつか提示した．ここで，S は学習データ，$\mathbf{x}$ は特徴ベクトル，y はラベル，$P_\theta(y|\mathbf{x})$ は $\mathbf{x}$ が 与えられたときの y に属する確率，f_θ はパラメタ θ を持つ分類器，k_θ は x_i と x の類似度を測るカーネル関数である．

表 2.1 メタ学習の一般的なアプローチ

	モデルベース	計量ベース	最適化ベース
キーアイデア	RNN および記憶	計量学習	最急降下法
$P_\theta(y|\mathbf{x})$ モデルの構築方法	$f_\theta(\mathbf{x}, S)$	$\sum_{(\mathbf{x}_i, y_i)} k_\theta(\mathbf{x}, \mathbf{x}_i) y_i$	$P_{g_\phi(\theta, S^L)}(y|\mathbf{x})$

表中で，メタ学習は “モデルベース”，“計量ベース”，“最適化ベース” の 3 種類に大別されている．モデルベースのアプローチは $P_\theta(y|\mathbf{x})$ の形式を仮定せず，他のメタ学習モデルや内部アーキテクチャによりパラメタ更新がされるものである．計量ベースのアプローチは最近傍アルゴリズムやカーネル密度推定のように，未知ラベル y は既知のラベルセットの加重和から計算されるものである．最適化ベースのアプローチは，深層学習アルゴリズムにおける

勾配の後方伝播を最適化するものである．

以下に，それぞれの分類の研究例を挙げる．

モデルベースのアプローチには，“Memory-Augmented Neural Networks” (MANN) を用いたものがある．MANN はアルゴリズムを学習する Memory を持つ Neural Network である Neural Turing Machine の一種であり，通常の Reccurent Neural Network（RNN）や Long Short Term Memory（LSTM）のような内部メモリのみを持つニューラルネットワークとは異なる仕組みにより，わずかなサンプルにより新しいタスクに適応可能である．Santoro らは MANN において Memory に保存されている表現へのアクセス性を上げるために，データセットとラベルの情報を関連付けて学習を行う方法を提案した [11]．ここで，前述のモデルベースのアプローチの説明に対応付けると，そのラベルを持ったデータセットの情報をモデル内部の表現として持つことで単にラベルと表現のみを保持する場合に比べて MANN の性能を向上させている．

計量ベースのアプローチには，“Convolutional Siamese Neural Network” (CSNN)[12] がある．Siamese Neural Network (SNN) は 2 つのネットワークで構成され，入力データサンプルのペア間の関係を学習する．これらのネットワークは同じ重みとネットワークパラメタを共有している．CSNN は SNN を画像分類を行う方法として用いており，2 つの画像が同じクラスに属する確率を出力するように学習させ，学習セット内全ての画像とテスト画像で最も高い確率を示した画像のクラスをテスト画像のクラスとするものである．ここで，前述の計量ベースのアプローチの説明に対応付けると，CSNN はテスト画像に対し学習データセット内から最近傍のものを選び，それと同じクラスに属すると判断する．

最適化ベースのアプローチには，“Model-Agnostic Meta-Learning” (MAML)[13] がある．MAML は勾配降下で学習する任意のモデルと互換性のある，かなり一般的な最適化アルゴリズムである．パラメタ θ を持つモデル f_θ に対し，タスク τ_i とそれに関連するデータセット $(\mathcal{D}^{(i)}_{train}, \mathcal{D}^{(i)}_{test})$ が与えられたとき，1 つ以上の勾配降下ステップにより下式 2.1 のようにモデルパラメタを更新できる．この式は 1 ステップの更新を表しており，$\mathcal{L}^{(0)}$ は id (0) のミニバッチデータを用いて計算された損失である．

$$\theta'_i = \theta - \alpha \nabla_\theta \mathcal{L}^{(0)}_{\tau_i}(f_\theta) \tag{2.1}$$

上式は 1 つのタスク (τ_i で示される) に対して最適化されるが，様々なタスクで良好な一般化を達成するために，タスク固有の微調整がより効率的になるように最適な θ^* を見つけたいと考える．ここで，id (1) の新しいバッチデータをサンプリングすると，損失は $\mathcal{L}^{(1)}$ と呼ばれ，これはミニバッチ (1) に依存する．$\mathcal{L}^{(0)}$ と $\mathcal{L}^{(1)}$ の上付き文字は異なるバッチを示すだけで，同じタスクの同じ損失目的を指している．そこで MAML は図 2.1 に示すアルゴリズムに従って θ^* を発見する．

ここで，MAML を前述の最適化ベースのアプローチの説明に対応付けると，学習モデルの持つパラメタ θ をタスクのバッチ τ_i を用いて繰り返し更新し，様々なタスクで用いるため

Algorithm 1 Model-Agnostic Meta-Learning

Require: $p(\mathcal{T})$: distribution over tasks
Require: α, β: step size hyperparameters
1: randomly initialize θ
2: **while** not done **do**
3: Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
4: **for all** $\mathcal{T}_i$ **do**
5: Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ with respect to K examples
6: Compute adapted parameters with gradient descent: $\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
7: **end for**
8: Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$
9: **end while**

図 2.1 MAML の一般形 (原著 [13] より)

に良好な一般化がされた θ^* を発見することで，タスク固有の微調整を効率化させている．

ここまでは，近年のメタ学習の概観について示すために Weng の記事で紹介されていたものを取り上げていた．しかし，この記事は前述の通り特に画像処理を対象とした深層学習におけるメタ学習に関するものである．そのため，次に本研究とより関係性の近い研究について述べる．

Arjmand らは最適な時系列予測アルゴリズムを選択するアプローチを行った [14]．気象予測において，地理的特性はアルゴリズムの性能や予測精度に有効な要因の一つであるのに対し，その分野においてメタ学習モデルはあまり考慮されていなかったため，彼らの研究では気象学的特性が最適な予測器を決定するフレームワークの開発を行った．彼らは外生的な気象変数をメタ特徴として用い，6 種類の時系列予測アルゴリズムの候補の中から最適な予測器を選択することを目的としていた．実験では最適な予測器が異なると考えられた 10 の地理的なゾーンから採取されたサンプル群に対して予測が行われ，結果的に歪度や尖度といった気象データの統計的なメタ特徴が各サンプル群に最適な予測器の割り当てに重要な役目を果たした．

Garcia らは最適な分類器構築アルゴリズムを選択するアプローチを行った [15]．この研究はメタ特徴としてクラスタリング尺度を用いていることが特徴的である．クラスタリング尺度は，データセットに対しクラスタリングを行った結果，その構造の良し悪しを評価する指標であり，コンパクト性や分離性などの情報を指す．クラスタリング尺度を含めたメタ特徴を用いて，分類器構築アルゴリズムを推薦するシステムを構築し，データセット上で全ての分類器をテストするのに加えて大幅な計算コスト減を実現した．

本研究では，上述の研究と同様にデータセットの性質としてメタ特徴を用いて予測を行う．これらの研究と異なる部分は予測対象が分類可能性であることだが，その狙いには Garcia

らの研究と同様に計算コストの削減も含まれている．分類可能性については第 3 章で，分類可能性予測については第 4 章で述べる．

第 3 章

分類可能性

3.1 既存の分類器評価指標

この章では分類可能性の定義を行うが，ここではまず既存の分類器評価指標について考える．第 2.3 節で言及したとおり分類器の評価を行うために様々な指標が存在しているが，ここでは Seliya の論文 [16] を参考に例示する．この論文では 9 種類の分類性能評価指標（うち 2 種類はさらに 8 種類の指標を内包）を紹介しており，その内訳は以下の通りであった．

- Accuracy and Predictive Values (A-PV)
- F-measure (FM)
- Geometric Mean (GM)
- Area Under the ROC Curve (ROC)
- Area Under the Precision-Recall Curve (PRC)
- Logarithmic Score (LGS)
- Brier Inaccuracy (BRI)
- Divergence (DVG)
- Kolmogorov-Smirnov Statistic (KSS)

A-PV，FM および GM はそれぞれ分類器の決定閾値 $[0, 1]$ 毎に計算が可能である．
また，A-PV および KSS はそれぞれ以下が内包されている．

- Accuracy (ACR)
- Misclassification Rate (MCR)
- True Positive Rate (TPR)
- True Negative Rate (TNR)
- False Positive Rate (FPR)
- False Negative Rate (FNR)
- Positive Predictive Value (PPV)

- Negative Predictive Value (NPV)

これらの指標について概括的に説明する．A-PV から KSS の全ての指標は 2 クラス分類問題における分類器の評価指標である．それらの多くは 2x2 混合行列の True Positive(TP)，True Negative(TN)，False Positive(FP)，False Negative(FN) より算出される．

A-PV はある決定閾値における TP，TN，FP，FN の比となっており，FM は TPR と PPV より，GM は TPR と TNR より算出される．ROC は TPR と FPR，PRC は TPR と PPV をそれぞれ決定閾値毎に計算しプロットしたものの Area Under the Curve である．

LGS，BRI および DVG は全て混合行列の各値を用いずに予測結果がどの程度正解ラベルと一致しているかを示す指標である．それぞれ，LGS および BRI はクロスエントロピー，DVG は統計量に基づいている．

KSS は A-PV と同様に混合行列から得られた値を用いて算出する．その際に，Kormogorov-Smirnov 検定により正負例の 2 群間の距離が最大となる決定閾値を探し，その決定閾値における ACR をはじめとした各評価値を算出する．

なお，この論文の目的は分類器評価指標間の関係に対する考察を行うことにより，これらに対する理解を分析者に提供することで，独立した側面から分類器の性能評価を行えるような指標選択を用意することであった．論文中では 35 のデータセットと C4.5 を用いて構築した決定木に対する因子分析により，3 つの因子それぞれから指標を選択することを提案している．その一例として，(1) ROC (2) BRI (3) KSS-ACR が挙げられている．

ただし，この結果はそれぞれの指標の性質を理解したうえで用いることが前提であり，本研究のユースケースに適しているとは考え難い．そこで本研究ではこれらの指標の組み合わせにより新たな指標である分類可能性を定義する．

3.2 分類可能性の定義

ここでは分類可能性の定義を行う．本研究では，ある閾値における分類可能性を正判定が誤判定に対しその閾値に対応した割合を超えることを保証する指標と定義し，5 つの分類器評価指標を 7 つのアルゴリズムにより構築された分類器で算出し，各指標における最大値が閾値 θ を超えているか否かを示す二値の指標とする．

この節では，分類可能性の定義について 3 つの小節に分けて述べる．まず分類可能性定義に用いる分類器評価指標について述べ，次にこれらの分類可能性評価指標を算出する分類器の構築アルゴリズムについて述べる．最後に θ の定め方とその一例を示す．

3.2.1 分類可能性定義に用いる分類器評価指標

分類可能性は第 3.1 節で述べた分類器評価指標のうち，“Accuracy and Predictive Values” における以下の指標を用いて定義する．

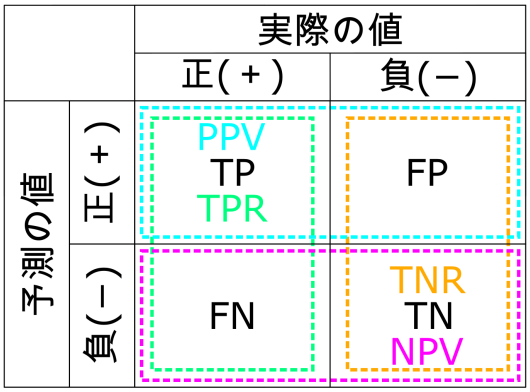

図 3.1 混合行列における各指標

1. Accuracy (ACR)
2. True Positive Rate (TPR)
3. True Negative Rate (TNR)
4. Positive Predictive Value (PPV)
5. Negative Predictive Value (NPV)

これらの指標は全て混合行列から得られる．図 3.1 は混合行列における各指標の対応を示している．この図では各指標と色が対応しており，破線で示している 2 つの値を用いて各指標は計算される．図に示した通りこれらの指標は全て正解と不正解の比率となっており，TPR は TP と FN，TNR は TN と FP，PPV は TP と FP，NPV は TN と FN に対応している．なお，ACR のみ TP および TN と FP および FN と全ての値を用いて計算される．

第 3.1 節で述べた指標に “F-measure” (FM) があるが，この指標は下式により求められ，正例に関する分類性能を示す指標 (FM_+ と表記) である．

$$FM_+ = \frac{2 \cdot TPR \cdot PPV}{TPR + PPV} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{3.1}$$

また，負例に関する FM(FM_- と表記) は以下のように求められる．

$$FM_- = \frac{2 \cdot TNR \cdot NPV}{TNR + NPV} = \frac{2 \cdot TN}{2 \cdot TN + FP + FN} \tag{3.2}$$

ここで，3.1 と 3.2 で用いられている指標は，TPR，PPV，TNR，NPV であることから，これらの 4 つの指標を用いることで FM_+ および FM_- の両方を求めることができる．しかしこれら単体では分類器の全体的な分類性能を示していないため，全数に対する正誤率である ACR を用いて 5 つの指標を用いることとする．

3.2.2 分類可能性定義に用いる分類器構築アルゴリズム

分類可能性はあるデータセットにおける特定のアルゴリズムで構築された分類器の指標を示すものではなく，そのデータセットの分類タスクにおける期待度を示す指標である．そのため，あるデータセットに対し一般的に用いられる様々なアルゴリズムを用いて構築された分類器の性能を求める必要がある．

一般的な分類タスクで用いられる分類器構築アルゴリズムの選択は，Tavasoli による記事 [17] をもとに行った．この記事ではよく用いられている分類器構築アルゴリズムとして以下の 7 つが挙げられている．

- Logistic Regression (LR)
  説明属性から各事例があるクラスに属する確率を回帰により予測する
- Decision Tree (DT)
  各説明属性のうち情報利得が最大となる属性により事例を分割する作業を繰り返すことにより構築された木を用いて予測する
- Support Vector Machine (SVM)
  説明属性空間において各事例をクラスで分割する超平面を構築する点により予測する
- Naive Bayes (NB)
  各説明属性がある値を取る場合にその事例がどのクラスに属するかという条件付き確率を全ての属性で考慮することで予測する
- K-Nearest Neighbor (KNN)
  周囲 K 個の事例からその事例のクラスを予測する
- Random Forest (RF)
  学習データからランダムサンプリングにより生成したサブサンプルにより学習した複数の Decision Tree を用いて予測する
- Gradient Boost (GB)
  Decision Tree を繰り返し構築する際に以前構築した木の誤判定を加味して構築することにより誤判定を抑えつつ予測する

これらは全て異なる方策を用いるため，これらのいずれかにより構築された分類器が高い性能を示すデータセットは少なくとも 1 つの方策により分類が可能であると考えられる．そのためこれらのアルゴリズムで構築された分類器の分類性能から各分類性能における最大値を用いて分類可能性を定義する．

なお，これらのアルゴリズムで用いたパラメータは全て Python のオープンソース機械学習ライブラリである Scikit-Learn のデフォルト値を用いた．

3.2.3 分類可能性における閾値

分類可能性を計算する際の閾値 θ にはそのデータセットに対し求めたい性能の目安を設定する．

例として，はじめて分類可能性の定義を行った論文 [18] では $\theta = 0.75$ を提示した．この論文ではその理由として，“TP，TN が FP，FN に対し 3 倍以上となるような分類が可能となる” としたが，以下ではその詳細について述べる．

分類可能性定義に用いた指標はそれぞれ下式 3.3 から 3.7 により計算される．

$$ACR = \frac{TP + TN}{TP + FP + FN + TN} \tag{3.3}$$

$$TPR = \frac{TP}{TP + FN} \tag{3.4}$$

$$TPR = \frac{TN}{TN + FP} \tag{3.5}$$

$$TPR = \frac{TP}{TP + FP} \tag{3.6}$$

$$TPR = \frac{TN}{TN + FN} \tag{3.7}$$

また，3.3 は，$T = TP + TN$ および $F = FP + FN$ を用いて下式のように変形できる．

$$ACR = \frac{T}{T + F} \tag{3.8}$$

式 3.4 から 3.8 より，分類可能性定義で用いられている 5 つの指標は全て正解数 C に対する不正解数 IC との和との比となっており，これらの指標を一般化すると下式 3.9 となる．

$$Measurement = \frac{C}{C + IC} \tag{3.9}$$

分類可能性はこれらの指標全てで閾値 θ を超えているかという定義であるが，あるデータセットから得られたある指標を $Measurement$ とした時，式 3.9 より θ に対し下式の関係となる．

$$\begin{aligned} Measurement = \theta < \frac{C}{C + IC} \\ \theta(C + IC) < C \\ \theta IC < (1 - \theta)C \\ \frac{\theta}{1 - \theta} IC < C \end{aligned} \tag{3.10}$$

式 3.10 より，C と IC の関係は θ と $1-\theta$ の比となることが示された．表 3.1 に θ と C/IC の比率を示す．$\theta = 0.75$ の場合 C は IC の $0.75/0.25 = 3$ 倍以上となり，全ての指標においてこれが保証されることとなる．

表 3.1　閾値 θ と比率の対応表

θ	比率	θ	比率
0.60	1.50	0.80	4.00
0.61	1.56	0.81	4.26
0.62	1.63	0.82	4.56
0.63	1.70	0.83	4.88
0.64	1.78	0.84	5.25
0.65	1.86	0.85	5.67
0.66	1.94	0.86	6.14
0.67	2.03	0.87	6.69
0.68	2.13	0.88	7.33
0.69	2.23	0.89	8.09
0.70	2.33	0.90	9.00
0.71	2.45	0.91	10.11
0.72	2.57	0.92	11.50
0.73	2.70	0.93	13.29
0.74	2.85	0.94	15.67
0.75	3.00	0.95	19.00
0.76	3.17	0.96	24.00
0.77	3.35	0.97	32.33
0.78	3.55	0.98	49.00
0.79	3.76	0.99	99.00

分類タスクにおける最適な比率を示す閾値を一意に定めることは困難である．1 つの方法として，いくつかの閾値による分類可能性を提示することによりその中から分析者が用いたい閾値における分類可能性を選択するというものがあるが，これは本研究のユースケースであるデータサイエンスの非専門家がデータの分類における性能を簡易的に調べるという場合においては適しているとは言えない．

分類可能性予測において，低い閾値を用いた場合は予測された分類可能性が正であるにも関わらず高い分類性能が実現できない可能性が高くなり，高い閾値を用いた場合は予測された分類可能性が負であるにも関わらず高い分類性能が実現できる可能性が高くなる．本研究の目的の背景にはデータマイニングの活用機会を向上させ，データマイニングの社会における寄与を促進するというものがあるため，機会損失につながると考えられる後者は回避した

い．しかし，前者の場合，データマイニングを行いたいと考えた主体が予測された分類可能性を元に必要コストの見積もりをおこなった際に，より大きなコストが要求される恐れがある．

なお，複数閾値の組み合わせを含めた予測された分類可能性の使い方については，第 6.2 節で言及する．

3.3 分類可能性を用いる意義

あるデータセットに対する分類可能性は，特定のアルゴリズムにより構築された分類器の性能を示すものではない．その理由は，この研究の目的はあるデータセットの分類タスク適用への期待度を求めることであるためだ．

論文 [19] では，分類可能性ではなく Classification And Regression Tree (CART) および Multi Layer Perceptron (MLP) により構築された分類器の各予測指標 Accuracy と F1 Value をメタ特徴から回帰により予測した．この論文における実験は，上述の 2 手法により構築された分類器の予測指標を Lasso 回帰，Ridge 回帰，ElasticNet を適用し，それらのパラメタを α のみ変化させることでそれぞれ最も性能が良かったモデルについて議論した．表 3.2 に実験結果を示す．

表 3.2 回帰分析の結果

データ名	アルゴリズム	R^2	$RMSE$
CART-F1	Ridge($\alpha = 5.0$)	0.762	0.156
MLP-F1	Ridge($\alpha = 5.0$)	0.749	0.159
CART-Acc	Ridge($\alpha = 1.0$)	0.661	0.146
MLP-Acc	Ridge($\alpha = 5.0$)	0.468	0.176

結果は回帰予測器の決定係数は概ね 0.5 を上回る良好な結果となったが，予測対象である各指標に値域 $[0.0, 1.0]$ に対し，平均二乗平方根誤差（Root Mean Squared Error，RMSE）が 0.15 程度となっていた．RMSE は予測における各事例で正解からの誤差がどの程度生じているかを示していると考えられるため，値域に対する割合として 15% 程度の誤差である．

ここで，生じた 15% 程度の誤差について考える．本研究で目的としているあるデータセットで分類タスクを行った際にどの程度の性能が期待できるかという指標として，15% 程度の誤差を持った実数による予測値はわかりにくい指標である．例えば，あるデータセットに対し 0.60 ± 0.15 という予測結果が得られた場合，そのデータにどの程度の分類タスクに対する期待度を持っているかという直接的な指標とすることは難しい．

また，この論文における実験では CART と MLP の Accuracy および F1 Value を個々のモデルで予測している．この方法をあるデータセットの分類タスクへの期待度を求めるために，ある分類モデルに対するある指標を求めるものとして拡張した場合，その目標に応じて

異なるモデルを構築する必要がある．そのため，個々のモデルでは，あるデータの分類タスクへの期待度ではなく，あるデータが特定のアルゴリズムで特定の指標においてどの程度の評価値が得られるかを予測するものとなる．

以上により，本研究の目的に対してあるデータのある識別器構築アルゴリズムによる分類器評価指標を実数値として求める方法は相応しくないと考えられる．それに対し，分類可能性を用いることにより上述の問題を解決できると考えられる．

まず，得られた予測値の利用方法という観点で考える．得られる予測値は複数のモデルのうち最大となる各評価値が全て閾値 θ を超えているか否かの二値である．この閾値は第 3.2.3 節で述べた通り，どの程度の性能を見込みたいか，具体的には正解と不正解の比率をどの程度に設定するかというものから決定する．これをシステムの提供者が予め設定しておくことで予測値としてのわかりやすさに繋がるのではないかと考えられる．

次に，特定の指標やモデルに対する適正ではなく分類タスクに対する期待度を求めるという観点で考える．分類可能性は複数のモデルと複数の評価指標を用いている．これによりいずれかのモデルで高い分類性能を示すか否かを示すことに加えて，正例負例問わず高い性能を示せるかということを示すことが可能になると考えられる．

本論文における分類可能性の定義は，目標である未知のデータセットがどの程度分類性能を示すことが可能かを示す指標として，第 2.3 節で述べた通り複数の学習アルゴリズムにより構築された分類器の性能を用いることから適しているものであると考えられる．また，第 2.4 節で述べた計算コストの低減についても，各データセットに対し分類可能性定義に用いている全ての学習アルゴリズムを用いて分類器構築を行うのに比べて，第 4 章で述べるメタ特徴を用いた予測によって低減が可能となる．

第 4 章

分類可能性予測

4.1　分類可能性予測の概観

分類可能性予測は，第 3 章で定義した分類可能性をデータセットの性質から予測するものである．データセットの性質はメタ特徴と表記し，その詳細は後述する．

分類可能性予測システムの概観は図 4.1 の通りである．

分類可能性予測には以下の 3 つのコンポーネントが必要である．

1. メタ特徴抽出器
2. 分類可能性予測器
3. 予測解釈器

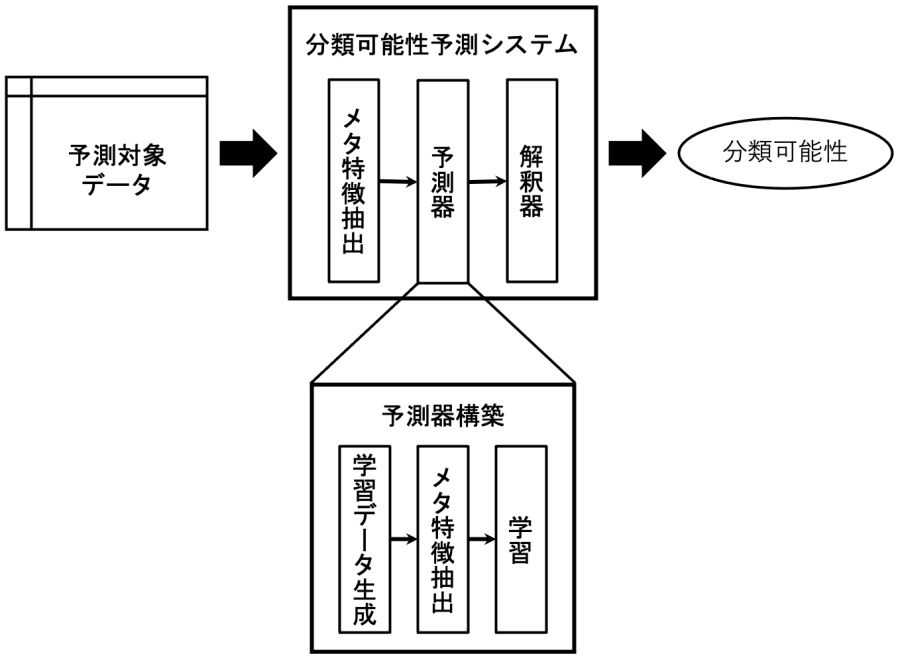

図 4.1　分類可能性予測システム概観

まず，メタ特徴抽出器は分類可能性を予測対象のデータセットのメタ特徴を抽出する．抽出するメタ特徴は第 4.2 節で述べる．

次に，分類可能性予測器は第 4.3 節で述べる方法により入力されたメタ特徴から分類可能性を予測する．

そして，予測解釈器は分類可能性予測器より得られた結果をもとに予測対象のデータセットの分類可能性を算出する．これは分類可能性予測器がデータセットに対し単一の結果を出力する場合は分類可能性予測器から得られた結果をそのまま出力とする．

本論文で提案する分類可能性予測システムは分類可能性予測器は単一の結果を出力するが，分類可能性予測システムの拡張において予測解釈器の動作は変化することが考えられる．予測解釈器の動作が変化するケースについては第 6.2 節で述べる．

分類可能性予測システムにおいて，分類可能性予測器は事前構築しておく必要がある．分類可能性予測器の構築は第 4.3 で述べるが，その構築には訓練データセット群が必要となる．訓練データセット群の持つべき性質とその訓練データセット群により予測できる対象との関係性に関する考察は第 5 章で行う．

また，分類可能性予測システムを本研究のユースケースに合わせて提供することを考えた場合，2 つの理由により構築済みのモデルを提供する必要がある．

1 つ目の理由は，訓練データセット群の収集が容易ではないためである．これは後述する訓練データセット群の持つべき性質に関連しており，本研究のユースケースであるデータサイエンスの非専門家が訓練データセット群を収集するのは困難であると考えられる．

2 つ目の理由は，分類可能性予測システムの学習にかかる計算コストの高いためである．訓練データセット群から分類可能性予測システムを構築する際に，分類可能性を抽出する必要がある．分類可能性の定義は後述の通りであるが，抽出の際に各分類器構築アルゴリズムによる学習と評価が必要である．

4.2 分類可能性予測器に用いるメタ特徴

メタ特徴はデータセットの性質を示す特徴量である．分類可能性予測をデータセットのメタ特徴を用いて行うことにより，分類器構築およびその評価を行うことなくそのデータセットが分類タスクを実行するにあたって必要な情報を持っているかどうかを簡易的に判断できるのではないかと考えられる．メタ特徴を用いた予測に関する関連研究は第 2.4 節で述べた．

本研究で提案する分類可能性予測に用いるメタ特徴は，主に Hutter らの書籍 [9] を参考に選出した．以下にそのメタ特徴の一覧を示す．

1. データセットのインスタンス数
2. データセットの次元性 [20][21]
   データセットのインスタンス数を n，データセットの属性数を m とした時，n/m で表される．

3. 目的属性のエントロピー
4. 平均正規化情報エントロピー [22]
   各属性のエントロピーを最大エントロピー ($\log_2 n$) で割り，それらの平均値を取ることにより求まる．
5. 平均相互情報量 [23]
   各属性の目的属性に対する相互情報量を求め，それらの平均値を取ることにより求まる．
6. 最大相互情報量 [23]
   各属性の目的属性に対する相互情報量を求め，それらの最大値を取ることにより求まる．
7. 等価特徴数 [23]
   クラスエントロピーと平均相互情報量の比により求まる．
8. 雑音信号比 [23]
   各属性のエントロピーから平均相互情報量を引き，平均相互情報量で割ることで求まる．
9. 最大 Fisher 判別比 [24]
   2 つのクラスをそれぞれ c_1, c_2 とした時，それぞれに関する平均値を μ_{c_1}, μ_{c_2}，分散を $\sigma_{c_1}, \sigma_{c_2}$ として，$(\mu_{c_1} - \mu_{c_2})^2/(\sigma_{c_1}^2 - \sigma_{c_2}^2)$ により計算される Fisher 判別比を各属性で求め，それらの最大値を取ることにより求まる．

選出したメタ特徴の多くは情報理論に基づいた指標となっている．その理由として，決定木に代表されるいくつかの分類器構築アルゴリズムでは情報利得に基づいて適切な分割を探しており，情報理論的にデータの分割が容易であるデータセットは分類問題として解くのが簡単であると考えられるためである．

Fisher 判別比は，各属性の平均値および分散を用いるため，Stevens の尺度水準 [25] における非数属性の中でも多次元名義尺度 [26] および順序尺度を持つものについては適切に求めることができない．しかし，一次元名義尺度においては，その値に該当するか否かの割合として平均および分散を用いることが可能であると考えられるため，One-hot Encoding[27] を行った後抽出する．

4.3 分類可能性予測器の構築

分類可能性予測器は二値分類器である．この予測器はあるデータセットから抽出されたメタ特徴を受け取り，そのデータセットの分類可能性を出力する．

分類可能性予測器の構築は以下の手順により行われる．

1. 学習用メタ特徴・分類可能性データセットを作成する
   （a）学習用データセット群を収集する

(b) データセット群からメタ特徴と分類可能性を抽出する
2. 分類器構築アルゴリズムにより学習する

分類可能性予測器で予測が可能であるためには，学習用データセット群に予測対象データセットとデータ性質が類似したデータセットが含まれている必要がある．その際に必要となるデータ性質の多様性は不明であるが，それに関する考察は第 5 章で述べる．

学習用メタ特徴・分類可能性データセットは第 4.2 節で述べたメタ特徴と第 3 章で述べた分類可能性のセットの集合である．このデータセットにおける 1 事例は多くの場合 1 データセットである．また同一のデータセット群に対し分類可能性を算出する際の閾値 θ が異なるバリエーションが考えられる．

分類可能性予測器を構築する際に用いる分類器構築アルゴリズムは本研究では決定木を用いたアンサンブル学習手法の一種である Random Forest を採用した．その理由は，学習の際に必要となるハイパーパラメタ調整が最小で良いということに加えて，各決定木におけるメタ特徴の寄与から各メタ特徴がどの程度分類に寄与するかを考えやすいこと，単一の決定木に比べて高い分類性能が得やすいことである．

4.4 分類可能性予測器の性能評価

ここでは提案した分類可能性予測器の性能を人工データを用いた試行と実データを用いた試行により検証する．各データセット群の内容は各小節で述べる．

4.4.1 人工データを用いた試行

検証に用いる人工データは以下の 3 種類の方法で生成する．

データ群 1 数値（$\mathcal{N}(0,1)$ から生成）のみからなる
データ群 2 バイナリ値（等確率の二項分布から生成）のみからなる
データ群 3 上記の数値とバイナリ値を比率 1 : 9 から 9 : 1 で混合する

これらのデータ群は全て以下のステップにより生成する．

1. 行数 n，列数 m として $n \times m$ 行列を生成する
2. $\mathcal{N}(0,1)$ から係数 $a_i (i = 1, \cdots, m)$ を生成する
3. 各列の値 x_i に対し $\sum_{i=1}^{m} a_i x_i$ を各行で求める
4. 3 で求めた値が正ならば 1，負ならば 0 をその行のクラスとする
5. $\mathcal{N}(0.5, 0.2)$ で得られた割合 r_n（ただし $0 < r_n < 1$）で行をランダムサンプリングする
6. $\mathcal{N}(0.5, 0.2)$ で得られた割合 r_m（ただし $0 < r_m < 1$）でクラス属性列を除く列をランダムサンプリングする

生成した各人工データセット群からメタ特徴・分類可能性を抽出した．その際分類可能性の閾値は $0.65, 0.7, 0.75, 0.8, 0.85$ の 5 種類を用いた．しかし，表 4.1 に示す通り，特に高閾値において著しく正例率が低下したため，Random Undersampling によりこの不均衡性を解消した上で分類可能性予測器の構築および評価を行った．

表 4.1　人工データセット群の分類可能性正例率

閾値	データ群 1			データ群 2			データ群 3		
	正例数	負例数	正例率	正例数	負例数	正例率	正例数	負例数	正例率
0.65	1431	8367	14.6%	1392	8588	13.9%	1798	8096	18.2%
0.70	515	9283	5.3%	761	9219	7.6%	911	8983	9.2%
0.75	181	9617	1.8%	389	9591	3.9%	385	9509	3.9%
0.80	61	9737	0.6%	185	9795	1.9%	151	9743	1.5%
0.85	11	9787	0.1%	69	9911	0.7%	50	9844	0.5%

分類可能性予測器はそれぞれ Random Forest により構築し，評価は 10–Fold 交差検証により行った．結果を表 4.2 に示す．

表 4.2　人工データの各閾値における予測性能 (対処後)

閾値	データ群 1			データ群 2			データ群 3		
	ACR	FM_+	FM_-	ACR	FM_+	FM_-	ACR	FM_+	FM_-
0.65	0.867	0.867	0.867	0.804	0.807	0.800	0.627	0.635	0.619
0.70	0.932	0.932	0.932	0.798	0.803	0.792	0.659	0.662	0.657
0.75	0.961	0.961	0.961	0.802	0.803	0.802	0.681	0.685	0.675
0.80	0.975	0.976	0.975	0.838	0.835	0.840	0.666	0.660	0.671
0.85	0.909	0.917	0.900	0.819	0.823	0.815	0.670	0.660	0.680

ここで，FM_+ および FM_- は第 3.1 節で述べた通り，それぞれ正例と負例に関する F1 値である．

結果的にデータ群 1 とデータ群 2 では ACR，FM_+ および FM_- について 8 割以上の分類性能が得られ，良好な分類が可能であると言える．

また，データ群 3 においてデータ群 1，2 と比べて高い分類性能が得られなかったことについては，Random Under Sampling によりデータバリエーションが減少したことに起因していると考えられる．データ群 3 は数値やバイナリデータを混合したものであるためデータ群 1 や 2 に比べて複雑なデータであることから，より複雑なデータを用いる際には学習データのバリエーション確保が必要であると考えられる．

表 4.3 検証に用いた実データの一覧

データセット名	ターゲット数	合計クラス数	データセット名	ターゲット数	合計クラス数
Acute_Inflammations	2	4	MONK's_Problems_3	1	2
Adult	1	2	Mushroom	1	2
AI4I_2020_Predictive_Maintenance_Dataset	6	12	Musk_(Version_2)_clean1	1	2
Amphibians	7	14	Musk_(Version_2)_clean2	1	2
Anuran_Calls_(MFCCs)	3	22	Nursery	1	5
Arcene	1	2	Occupancy_Detection	1	2
Audit_Data	1	2	Online_Shoppers_Purchasing_Intention_Dataset	1	2
Autism_Screening_Adult	1	2	Page_Blocks_Classification	1	5
Balance	1	3	Parkinson_Speech_Dataset_with_Multiple_Types_of_Sound_Recordings	1	2
Bank_Marketing_Additional	1	2	Parkinson's_Disease_Classification	1	2
banknote_authentication	1	2	Pen-Based_Recognition_of_Handwritten_Digits	1	10
Blood_Transfusion_Service_Center	1	2	Pittsburgh_Bridges	1	6
Breast_Cancer	1	2	Poker_Hand	1	10
Breast_Cancer_Coimbra	1	2	Polish_companies_bankruptcy_data_1styear	1	2
Breast_Cancer_Wisconsin_Diagnostic	1	2	Polish_companies_bankruptcy_data_2ndyear	1	2
Breast_Cancer_Wisconsin_Original	1	2	Polish_companies_bankruptcy_data_3rdyear	1	2
Breast_Cancer_Wisconsin_Prognostic	1	2	Polish_companies_bankruptcy_data_4thyear	1	2
Car_Evaluation	1	4	Polish_companies_bankruptcy_data_5thyear	1	2
Census-Income_(KDD)	1	2	Post-Operative_Patient	1	2
Cervical_cancer_(Risk_Factors)	4	8	QSAR_androgen_receptor	1	2
Cervical_Cancer_Behavior_Risk	1	2	QSAR_Bioconcentration_classes_dataset	1	3
Chemical_Composition_of_Ceramic_Samples	1	2	QSAR_biodegradation	1	2
Chronic_Kidney_Disease	1	2	QSAR_oral_toxicity	1	2
Climate_Model_Simulation_Crashes	1	2	Qualitative_Bankruptcy	1	2
CNAE-9	1	9	SCADI	1	6
Congressional_Voting_Records	1	2	seismic-bumps	1	2
Connect-4	1	3	Shill_Bidding_Dataset	1	2
Contraceptive_Method_Choice	1	3	Shuttle_Landing_Control	1	2
COVID-19_Surveillance	1	2	Soybean_(Small)	1	7
Credit_Approval	1	2	Spambase	1	2
Crowdsourced_Mapping	1	6	SPECT_Heart	1	2
Cylinder_Bands	1	2	SPECTF_Heart	1	2
Dermatology	1	6	Statlog_(Australian_Credit_Approval)	1	2
Diabetic_Retinopathy_Debrecen_Data_Set	1	2	Statlog_(German_Credit_Data)	1	2
Divorce_Predictors_data_set	1	2	Statlog_(Heart)	1	2
Drug_consumption_(quantified)	19	133	Statlog_(Image_Segmentation)	1	7
Dry_Bean_Dataset	1	7	Statlog_(Landsat_Satellite)	1	6
Early_stage_diabetes_risk_prediction_dataset	1	2	Statlog_(Shuttle)	1	7
Ecoli	1	8	Student_Performance_on_an_entrance_examination	1	4
EEG_Eye_State	1	2	Teaching_Assistant_Evaluation	1	3
Electrical_Grid_Stability_Simulated_Data	1	2	Thoracic_Surgery_Data	1	2
Forest_type_mapping	1	4	Tic-Tac-Toe	1	2
Glass_Identification	1	6	Trains	1	2
Haberman's_Survival	1	2	Ultrasonic_flowmeter_diagnostics_MeterA	1	2
Hayes-Roth	1	3	Ultrasonic_flowmeter_diagnostics_MeterB	1	3
HCC_Survival	1	2	Ultrasonic_flowmeter_diagnostics_MeterC	1	4
HCV_data	1	5	Ultrasonic_flowmeter_diagnostics_MeterD	1	4
Heart_failure_clinical_records	1	2	Urban_Land_Cover	1	9
Hepatitis	1	2	Vertebral_Column	2	5
HTRU2	1	2	Wall-Following_Robot_Navigation_Data_2	1	4
ILPD_(Indian_Liver_Patient_Dataset)	1	2	Wall-Following_Robot_Navigation_Data_24	1	4
in-vehicle_coupon_recommendation	1	2	Wall-Following_Robot_Navigation_Data_4	1	4
Internet_Advertisements	1	2	Waveform_Database_Generator	1	3
Internet_Firewall_Data	1	4	Waveform_Database_Generator_withNoise	1	3
Ionosphere	1	2	Website_Phishing	1	3
Iris	1	3	Weight_Lifting_Exercises_monitored_with_Inertial_Measurement_Units	1	5
Lung_Cancer	1	3	Wilt	1	2
Madelon	1	2	Wine	1	3
MAGIC_Gamma_Telescope	1	2	Wine_Quality_Red	1	6
Mammographic_Mass	1	2	Wine_Quality_White	1	7
MONK's_Problems_1	1	2	Yeast	1	10
MONK's_Problems_2	1	2	Zoo	1	7

4.4.2 実データを用いた試行

検証に用いる実データは UCI Machine Learning Repository から取得した．取得したデータセットには多クラス分類問題や 1 つのデータに対し複数の分類問題が設定されているものがあったため，それらを全て二クラス分類問題に変換した上でデータセット毎に択一した．

表 4.3 に用いたデータセットの一覧を示す．表中のターゲット数は 1 つのデータセットにいくつの分類問題が設定されているかを表しており，合計クラス数は複数ターゲットのあるデータセットに含まれる全てのクラスの数を表している．

表 4.4 に 5 種類の閾値 $0.65, 0.7, 0.75, 0.8, 0.85$ における実データの分類可能性を予測した際の性能を Leave-One-Out 交差検証により調べたものを示す．

表 4.4 実データの各閾値における予測性能

閾値	ACR	FM_+	FM_-
0.65	0.814	0.880	0.596
0.70	0.782	0.844	0.640
0.75	0.806	0.859	0.692
0.80	0.806	0.850	0.727
0.85	0.782	0.809	0.748

表に示した通り，全ての閾値で ACR および FM_+ では高い性能を示した．FM_- は高閾値になるほど高くなっているが，その原因は表 4.5 に示す通り学習データのクラス不均衡によるものであると考えられる．

表 4.5 実データの各閾値における正例率

閾値	正例数	負例数	正例率
0.65	93	31	0.750
0.70	86	38	0.694
0.75	84	40	0.677
0.80	80	44	0.645
0.85	69	55	0.556

第 5 章

学習データに求められる性質

5.1　構築済みモデル提供のための学習データについて

第 4.1 節で述べた通り，本研究で想定されるユースケースにおいて構築済みモデルの提供が必要であると考えられる．構築済みモデルは学習データのデータ性質により予測できる対象が変わることが文献 [28] で示唆されている．しかし，そこでは分類可能性が良好に予測可能な学習データ群と予測対象データとの関係性は示されていない．

また，第 4.4 節では人工データおよび実データを用いた分類可能性予測の例を示したが，これらの結果からは学習データと予測対象データが互いにどのような性質を持っている必要があるかということは明らかではない．特に人工データ群 3 における 10–Fold 交差検証で得られた予測性能は良好であるとはいえず，第 4.4.1 節で述べた通り，学習データのバリエーション確保が必要であると考えられる．

分類可能性予測を行うために構築済みモデルを用意するためには，どのような学習データ群を用意することでどのような対象データの予測に使えるかということを考える必要がある．その試みとしてこの章では，いくつかのデータ群間での構築と予測を行う例を示し，その学習データ群を用いて対象データの分類可能性予測を行えるか否かの判断基準としてデータ群間距離の傾向を調べる方法について議論する．

5.2　人工データ対実データにおける試行

まず，第 4.4 節で用いた人工データを学習データ，実データを予測対象データとして予測を行った結果の分類評価を表 5.1 に示す．なお人工データ群は，前述の通りクラス不均衡を解消するために，Random Under Sampling を行った上で予測器を構築した．

表に示した通り，いずれの人工データ群により学習した予測器においても良好な予測ができているとは言えない結果となった．各人工データ群のみを用いた予測ではデータ群 1 および 2 について高い予測性能を示せたことから，これらの人工データ群から抽出したメタ特徴・分類可能性データセットは実データを予測するために相応しいものとは言えないのでは

表 5.1　人工データモデルで実データを予測した結果

閾値	データ群 1			データ群 2			データ群 3		
	ACR	FM_+	FM_-	ACR	FM_+	FM_-	ACR	FM_+	FM_-
0.65	0.653	0.779	0.189	0.661	0.764	0.400	0.581	0.687	0.366
0.70	0.355	0.310	0.394	0.669	0.760	0.468	0.694	0.771	0.537
0.75	0.347	0.243	0.426	0.677	0.759	0.512	0.726	0.793	0.595
0.80	0.411	0.291	0.497	0.685	0.769	0.506	0.710	0.788	0.538
0.85	0.484	0.347	0.573	0.694	0.750	0.604	0.556	0.689	0.225

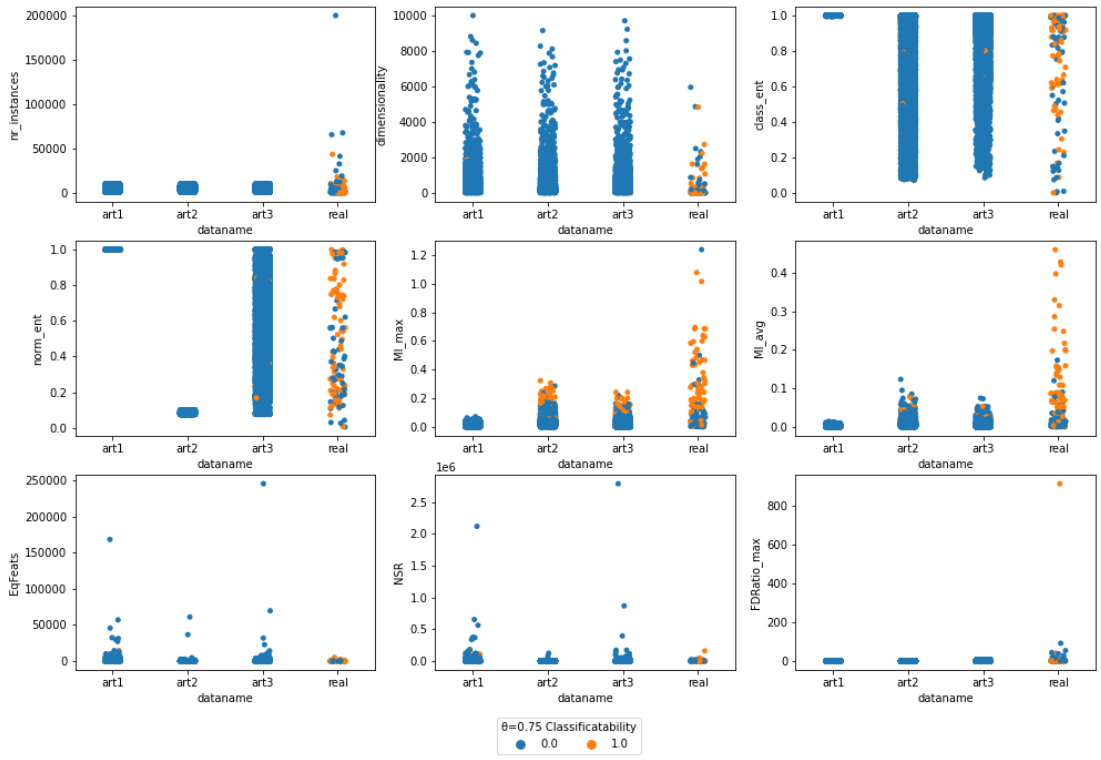

図 5.1　人工データ群のメタ特徴散布図

ないかと考えられる．

図 5.1 に各人工データ群および実データ群のメタ特徴ごとの散布図を示す．図中の “art1”，“art2” および “art3” はそれぞれ人工データ群 1 から 3 を示しており，“real” は実データ群を示している．また，図中の各点は $\theta = 0.75$ の分類可能性毎に色分けされており，1.0 が θ を超えているもの，0.0 が超えていないものを指している．

この図からわかることは，各データ群には他のデータ群とは異なる様々な特色があるということである．この節では各人工データ群で学習したモデルから実データ群の分類可能性を予測することを考えているため，ここでは各人工データ群と実データ群との対比で特色を述

べる．

“nr_instances” は各事例（各データセット）のインスタンス数である．各人工データ群はインスタンス数は最小 1000，最大 10000 として生成したが，実データ群では最小 8，最大 199523 である．これは実際のユースケースを考えた際にも起こりうることであるが，学習モデル側でこのメタ特徴がどのように扱われているかにより問題となるか否かが分かれると考えられる．

“dimentionality” は各事例の次元性，つまりインスタンス数と属性数の比である．各人工データ群では 10 程度から 10000 程度まで分布しているが，これは属性数を 10 から 110 として生成した後に各属性をランダムな確率によりランダムに削除しているためである．それに対し，実データ群においては 0.01 程度から 6000 程度まで分布しており，インスタンス数に対し属性数が 100 倍程度あるデータセットも含まれていることがわかる．このように実データには様々なデータ形状が考えられる．

“class_ent” はクラス属性のエントロピーである．人工データ群 2 や 3，実データ群はまばらに分布しているが，人工データ群 1 ではほぼ 1.0 付近に集中している．1.0 付近に集中している原因は，多くのデータセットでクラス属性が均衡分布となっていることが考えられる．人工データ群 1 は $\mathcal{N}(0,1)$ から生成された数値で構成されたベクトルに，$\mathcal{N}(0,1)$ から生成された係数により線型結合をすることで計算された数値の正負でクラス属性を生成しているが，結果的に $\mathcal{N}(0,1)$ が保持された数値の正負となっているためであると考えられる．

“norm_ent” は各属性のエントロピーをインスタンス数に対する最大エントロピーで正規化したものを属性について平均値を取ったものである．人工データ群 3 や実データ群では 0.0 から 1.0 にまばらに分布しているが，人工データ群 1 では 1.0 付近，人工データ群 2 では 0.0 付近に集中している．1.0 付近に集中する理由は，$\mathcal{N}(0,1)$ から得られた数値データのみで人工データ群 1 において，全ての事例で値が重複するとは考え難いため，全てのデータが異なっている状態である各属性が最大エントロピーと一致することによると考えられる．0.0 付近に集中する理由は，2 値データのみで構成される人工データ群 2 において，最大エントロピーで正規化していることによると考えられる．

“MI_max” および “MI_avg” はそれぞれ説明属性対クラス属性の相互情報量の最大値および平均値であり，相互情報量は各説明属性によりどの程度クラス属性のエントロピーを減少させられるかを示している．実データ群のこれらの値は各人工データ群の値に対して平均値，中央値，最大値の全てにおいて高い値を示しているが，その原因は各人工データ群のクラス属性の生成後にいくつかの属性を無作為に削除していることであると考えられる．この無作為な属性の削除は，人工データセットにおいてクラス属性の算出に寄与する観測可能な値とそうでない値を再現するために行ったが，実データ群に含まれる多くのデータセットはそのような性質になっていなかったのではないかと考えられる．

“EqFeats” は等価特徴数であり，“class_ent” と “MI_avg” の比により計算される．各人工データ群におけるこの値は実データ群のそれに対して非常に大きな値となっている．その原

因として “class_ent” が非常に大きいこと，“MI_avg” が非常に小さいことが考えられるが，前述の通り各人工データ群の “MI_avg” は実データ群のそれに対し非常に小さいものとなっているため後者が原因であると考えられる．

“NSR” は雑音信号比であり，各属性のエントロピーの平均値から “MI_avg” を引き “MI_avg” で割ったものである．人工データ群 1 と人工データ群 3，人工データ群 2 と実データ群がそれぞれ似た傾向を示している．人工データ群 2 と実データ群がそれ以外の人工データ群と比較して似た傾向を示した原因は，人工データ群 2 は 2 値属性で構成されているのに対し人工データ群 1 および 3 では数値属性を含んでいることから，実データと乖離するエントロピーを示しているのではないかと考えられる．なお “NSR” を算出する際の各属性のエントロピーは “norm_ent” とは異なり最大エントロピーで正規化していない．

“FDRatio_max” は最大 Fisher 判別比であり，各属性でクラス間の平均値の差と分散の差を用いて算出したものの最大値である．実データ群は各人工データ群に対し最小値，平均値，中央値および最大値で大きな値を示した．Fisher 判別比はクラスごとの事例群が各属性でどの程度離れているかを示しているが，各人工データ群は実データ群に比べクラスごとの事例群の距離が狭いと考えられる．この理由として，人工データ生成は無作為なものであり，実データ群に存在していると考えられる説明属性間の相関が存在しないことであると考えられる．

このように，第 4.4 節で生成した人工データは実データを予測するためには性質が乖離しているものであると考えられる．そのため，様々な実データを予測するために必要な学習データセット群はどのようなものを用意する必要があるかを考える必要性がある．

5.3 半データ学習モデルによる試行

“半データ学習モデル” は分類可能性予測の学習データについて考察した文献 [28] で提案した方法である．これは分類可能性予測において予測対象データの持つデータ性質を学習データ群に含めることを目的として提案した．

一般に，様々なデータセットはある母集団からサンプリングされたものであると考えられる．そのようなデータセットのサブサンプルも同様にその母集団からサンプリングされたものであると考えられる．そこで半データ学習モデルでは，あるデータセットの事例を 2 分割することにより，同母集団からサンプリングされた 2 つのデータセットを得る．従って，この方法は分析対象データセットを分割して分析・検証する方法であるが，交差検証などの分析対象データセットのサブセットをそのまま用いて学習器を構築する諸手法とは異なる．

また，同じデータセットから抽出したメタ特徴を 2 度持ちいることとの違いは，分割後のデータセットのメタ特徴は必ずしも元のデータセットおよびもう一方の分割後のデータセットのメタ特徴とは一致しないと考えられることである．その理由は，分割後のデータセットは同一母集団から得られたサンプルであるが，データセットに含まれる事例は一致しないためである．しかし，全く異なるデータを用いることと比べると母集団が同一であると考えら

れるため，類似した性質を持つデータとなることが期待される．

半データ学習モデルの構築および評価は以下の手順で行う．

1. データセット群に含まれる全てのデータセットについて，以下の手順を行う．
   i データセットの事例をクラス属性の分布が同等となるように 2 等分し，それぞれ半データ 1，半データ 2 とする．
   ii 半データ 1 および半データ 2 について，それぞれメタ特徴と分類可能性を抽出し，その組をメタ特徴・分類可能性データセットの 1 事例とする．
2. 半データ 1 のメタ特徴・分類可能性データセットで分類可能性予測器を構築する．
3. 分類可能性予測器に半データ 2 のメタ特徴を入力し，分類可能性を予測する．
4. 予測結果と半データ 2 の分類可能性を比較し，評価する．

表 5.2 に第 4.4 節で用いた実データ群で半データ学習モデルの構築および評価を行った結果を示す．なお，“HCC_Survival” および “Trains” はデータ分割を行った結果事例数が著しく減少したためこの実験には用いなかった．

表 5.2 半データ学習モデルの各閾値における予測性能

閾値	ACR	FM_+	FM_-
0.65	0.828	0.884	0.667
0.70	0.836	0.884	0.722
0.75	0.861	0.889	0.813
0.80	0.861	0.884	0.825
0.85	0.852	0.871	0.827

結果的に低閾値において FM_- が高い値を示さなかったが，閾値が低いほど正例率が高くなることによるものであると考えられる．しかし，同閾値で同程度の正例率となった第 4.4 節で行った実データの Leave-One-Out 交差検証の結果と比較して高い値を示していることから，半データ学習モデルによる試行は実データに対する試行と比較して性能が高くなると考えられる．

5.4 多クラス・多ターゲットにおける試行

多クラス・多ターゲットにおける試行は，あるデータセットにおいて 1 組の説明属性に対し複数のクラスがあるものおよび複数のクラス属性があるものを用いたものを示す．前者は多値分類問題を 1-vs-rest で分割したものを指し，後者は各クラス属性が独立したものを指している．これらの違いは前者はある事例が 1 つのクラスを満たしているとき他のクラスを満たさないのに対し，後者はそのような制約がないという点である．

この試行では，あるデータセットにおいて説明属性と分離した各クラスおよび各クラス属

性とを組にしたものをデータセットとみなし，それぞれのデータセットから抽出したメタ特徴および分類可能性をメタ特徴・分類可能性データセットの 1 事例として考え，Leave-One-Out 交差検証を行うことで分類可能性予測モデルの構築と評価を行う．この試行の目的は，いずれも説明属性が同じことから，それぞれのデータセットの性質が類似していると考えられるためである．

また，分類可能性の抽出は $\theta = 0.75$ で行い，分類可能性を抽出した際に正例と負例が少なくとも 2 つ以上含まれるデータセットのみを，第 4.4 節で用いた実データ群のうちから選び対象とした．ただし，第 4.4 節では各データセットについてメタ特徴・分類可能性データセットにおける 1 事例となるように用いたが，ここでは全てのデータセットのうち対象となるものを選んだ．

対象としたデータセットは表 5.3 に示す．表中で，多ターゲットとなったデータセットはターゲット数，多クラスとなったデータセットはクラス数を記述している．また，“Drug_consumption_(quantified)” は多ターゲットかつ多クラスとなったが，このデータセットのそれぞれのクラス属性はある薬物に対して使用したことがあるかないかの 2 値分類問題であるとも考えられるため，各クラス属性で 2 値分類問題として変形し多ターゲットなデータセットとして扱った．

表 5.3 多クラス・多ターゲットにおける試行で用いたデータセット

データセット名	ターゲット数	クラス数	正例数	負例数	正例率
AI4I_2020_Predictive_Maintenance_Dataset	6	-	4	2	0.667
Drug_consumption_(quantified)	19	-	4	15	0.211
Ecoli	-	8	5	3	0.625
Glass_Identification	-	6	3	3	0.500
HCV_data	-	5	3	2	0.600
Pittsburgh_Bridges	-	6	4	2	0.667
SCADI	-	6	3	3	0.500
Soybean_(Small)	-	7	2	5	0.286
Urban_Land_Cover	-	9	6	3	0.667
Yeast	-	10	5	5	0.500

これらのデータセットごとに分類可能性予測を行った結果を表 5.4 に示す．表に示す通り，いくつかのデータセットでは高い予測性能が得られなかった．

その理由として考えられるのは，同じ説明属性を持っていたとしてもクラス属性が異なることからほとんどのメタ特徴において似た傾向を持つとは限らないということである．本研究において分類可能性予測に用いているメタ特徴の多くは各説明属性とクラス属性との間の関係を表しているものである．それらのメタ特徴において説明属性が同一である状態で似た性質を示すということは，各クラス属性間で高い相関を示しているということになる．しかし，多ターゲットなデータセットにおいては高い相関があるかどうかはデータセットに依るため明らかではなく，多クラスなデータセットにおいてはそのクラス属性に含まれるクラス

表 5.4 多クラス・多ターゲットにおける試行における予測性能

データセット名	ACR	FM_+	FM_-
AI4I_2020_Predictive_Maintenance_Dataset	0.833	0.889	0.667
Drug_consumption_(quantified)	0.895	0.750	0.933
Ecoli	0.625	0.667	0.571
Glass_Identification	0.500	0.571	0.400
HCV_data	0.800	0.800	0.800
Pittsburgh_Bridges	0.333	0.500	0.000
SCADI	0.500	0.400	0.571
Soybean_(Small)	0.286	0.000	0.444
Urban_Land_Cover	0.667	0.727	0.571
Yeast	0.400	0.400	0.400

が多くなるほど相関が低下する．

従って，多クラス・多ターゲットにおける試行では説明属性が同じであるが，本研究で用いるメタ特徴および本研究の目的である分類可能性の予測において用いるデータの性質は必ずしも類似しているとは考えられない．

5.5 各試行結果に対する考察

ここまでは分類可能性予測に用いる予測器の学習データ群と予測対象データを変化させることで，どの程度の予測性能が得られるかを試行した．それらは全て第 4.4 節で示した実データ群やその一部を予測対象としている．それらの予測性能をまとめて表 5.5 に示す．なお，多クラス・多ターゲットにおける試行では ACR がそれぞれ最小値，中央値，最大値となっている例を示す．また，全て $\theta = 0.75$ で求めた分類可能性を予測対象としたもののみを示す．

表 5.5 各試行 $\theta = 0.75$ における予測性能

試行	ACR	FM_+	FM_-
実データ Leave-One-Out	0.806	0.859	0.692
人工データ対実データ データ群 1	0.347	0.243	0.426
人工データ対実データ データ群 2	0.677	0.759	0.512
人工データ対実データ データ群 3	0.726	0.793	0.595
半データ学習モデル	0.861	0.889	0.813
多クラス・多ターゲット 最小値（Soybean_(Small)）	0.286	0.000	0.444
多クラス・多ターゲット 中央値（Glass_Identification）	0.500	0.571	0.400
多クラス・多ターゲット 最大値（Drug_consumption_(quantified)）	0.895	0.750	0.933

この表から，今回用いた実データ群における予測性能は以下の順となっている．

1. 半データ学習モデル
2. 実データの Leave-One-Out
3. 人工データ対実データ
4. 多クラス・多ターゲット

まず，半データ学習モデルは各予測対象事例（データセット）は必ず 1 つの同一母集団から抽出されたデータセットが学習データ群に持っていることとなる．その結果これらの試行において最も高い性能を示したと考えられる．

次に，実データの Leave-One-Out は本研究で収集したデータセット群（表 4.3 に示したもの）のうち，各データセットごとに 1 事例となるようにメタ特徴・分類可能性データセットを対象として Leave-One-Out 交差検証により予測器構築および評価を行ったものである．そのため各事例を予測する際には同一母集団から抽出されたデータセットが学習データ群にないと考えられるが，人工データ対実データによる試行および多クラス・多ターゲットにおける試行に比べて高い予測性能を示した．従って，半データ学習モデルのように意図的に学習データ群と予測対象データとの関係性を作らなくても分類可能性予測が可能であるということが考えられる．

最後に，人工データ対実データおよび多クラス・多ターゲットでは同程度の予測性能を示したと考えられる．これらは高い予測性能を示しているとは言えず，本研究のユースケースに対して分類可能性予測器を事前構築する際に用いる学習データ群として相応しいものではないと考えられる．人工データ対実データにおいて高い予測性能を示せなかった理由は，第 5.2 節で述べた通り生成した人工データ群が実データ群を予測するためにはデータの性質が乖離しているものであると考えられる．多クラス・多ターゲットにおいて高い予測性能を示せなかった理由は，第 5.4 節で述べた通り本研究で用いるメタ特徴および本研究の目的である分類可能性の予測において用いるデータの性質は必ずしも類似しているとは言えないと考えられる．

これらの試行に対する考察から，特に実データの Leave-One-Out で高い予測性能の実現が可能であり人工データ対実データでそれが可能でなかった理由について，判断する指標を考えることで分類可能性予測モデルにおける学習データ群と予測対象データの関係性を示すことができると考えられる．その指標は用意した学習データ群に対してどの程度利用可能性があるかの判断基準や，その学習データ群により構築された予測器で対象データの分類可能性を予測した際にどの程度信憑性が高いかという判断基準になるのではないかと考えられる．

5.6 データ群間距離に基づく考察

第 5.5 節で言及した，分類可能性予測における学習データ群と予測対象データとのデータの関係性について示す指標として，データ群間距離を用いた判断が行えないかと考えた．

ここでは，各対象データの分類可能性予測が行われる際の学習データ群との群間距離を考え，以下の手順により求める．

1. 全ての学習データと予測対象データに対し，メタ特徴ごとに 0-1 スケーリングを行う
2. 予測対象データ群から 1 つの事例を取り出し，全ての学習データとの最小距離を計算する
3. 計算した距離をその予測対象データの学習データとの群間距離とする

メタ特徴ごとに 0-1 スケーリングを行う理由は，メタ特徴ごとのスケールの違いにより生じる距離計算における重みの差異を解消するためである．その際学習データと予測対象データ全てに対して最小値が 0 最大値が 1 となるようにスケーリングすることにより，データ間の相対的な大きさの比は変化しないものと考えられる．

また，予測対象データと学習データ群との距離計算を行う際に最小距離を計算する理由は，分類可能性予測が良好に行われるためには学習データ群の中に最も予測対象データと近いものが含まれている必要があると考えられるためである．

図 5.2 に各試行におけるデータ最小群間距離のヴァイオリンプロットを示し，表 5.6 に統計値を示す．ヴァイオリンプロットは各値がどの程度存在しているかを示しており，図中では横軸に示している各学習データ群がそれぞれ予測対象データ群とどの程度最小群間距離を持つかを表している．

表 5.6　データ最小群間距離の統計値

統計値	halfdata	realdata	mtarcls	real_art1	real_art2	real_art3
平均	0.1006	0.1477	0.4956	0.8644	0.3830	0.2022
標準偏差	0.1244	0.1278	0.3361	0.2772	0.2804	0.1868
最小値	0.0128	0.0175	0.0833	0.0727	0.0115	0.0194
中央値	0.0616	0.1316	0.3693	0.8900	0.2804	0.1312
最大値	0.9508	0.9426	1.3587	1.5572	1.0393	1.0332

図の横軸および表の列名では，各試行と以下のように対応している．

halfdata　半データ学習モデル
realdata　実データ Leave-One-Out
mtarcls　多クラス・多ターゲット
real_art1∼3　人工データ対実データ，1∼3 はそれぞれ人工データ群 1∼3 を指す

最小群間距離の傾向の類似は予測性能の傾向の類似と関係があると考えられる．高い予測性能を示した半データ学習モデルおよび実データの Leave-One-Out は他の試行と比べて最小群間距離が小さく，0 に近い値となった．次いで予測性能が高くなった人工データ群 2 および 3 を用いた人工データ対実データの試行では共に 2 つのコブを持った分布であると考

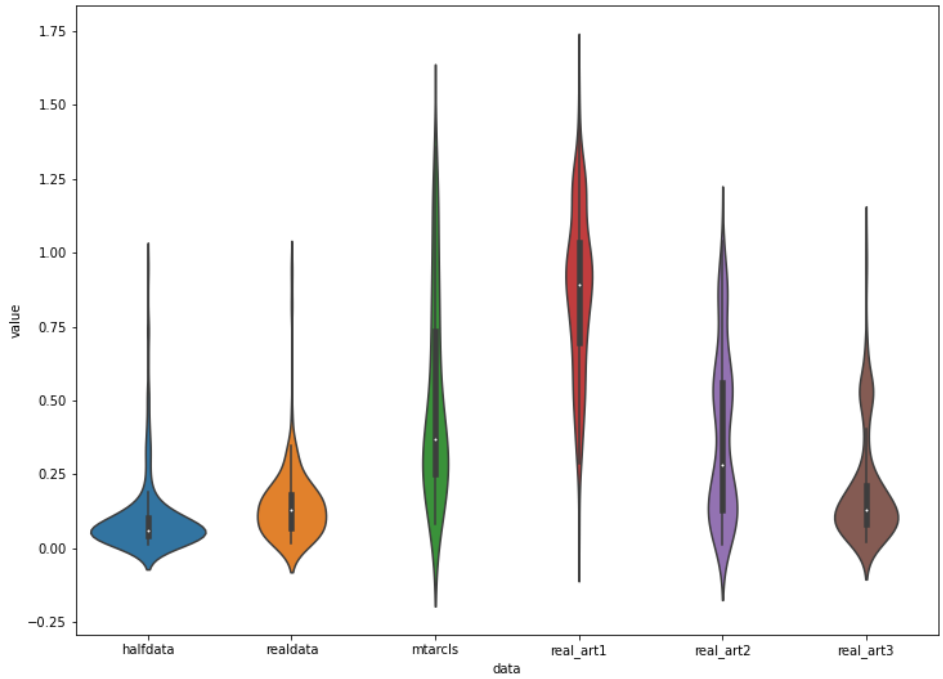

図 5.2　データ最小群間距離のヴァイオリンプロット

えられ，最小群間距離も人工データ群 1 を用いた人工データ対実データおよび多クラス・多ターゲットの試行と比べ小さく，半データ学習モデルおよび実データの Leave-One-Out の試行と比べ大きいという結果となった．従って，分類可能性予測で高い予測性能を達成させるならば，学習データ群と予測対象データとの最小群間距離が小さいほど望ましいと考えられる．

また，半データ学習モデルおよび実データの Leave-One-Out ではヴァイオリンプロット上の最小群間距離の分布がコブを 1 つ持つものとなっているため，これらの中央値に近い値であれば，これらの予測性能である ACR において 8 割程度の予測性能が達成可能であると考えられる．なお，人工データ群 3 においては 2 つのコブのうちより 0.0 に近いものでは実データの Leave-One-Out に近い分布であると考えられ，予測性能においても 3 つの人工データ群の中では最も高い値を達成したことから実データを予測するために必要な性質を部分的に持っていると顔が得られる．

従って，分類可能性予測において学習データと予測対象データとの関係は，最小群間距離が小さいものほど望ましく，最小群間距離が大きければ高い予測性能が期待できないのではないかと考えられる．

第 6 章

分類可能性予測に関する諸問題

6.1　分類可能性予測に関する諸問題

ここまでの章で述べた分類可能性システムに関する議論は，それぞれのトピックに関するものであった．この章では，それらが解決されていたとしてなお起こりうる問題について述べる．

まず，予測された分類可能性の利用方法について述べる．第 4.1 節に示した分類可能性予測システムを構成するためには，第 4 章で述べた分類可能性予測器だけでは不足している．概観は第 4.1 節に述べた通りであるが，予測器から得られた分類可能性をどのように利用するかということについては議論していない．ここではその利用方法について議論する．

次に分類可能性予測では対処できない問題について述べる．ここでは，あらゆるデータに対しても良好な分類可能性予測が行える学習データ群を用意したとしても対処できない問題について，形式の問題と計算コストの問題の 2 つの側面から議論する．

6.2　予測された分類可能性の利用方法について

分類可能性の利用方法として最初に考えられるのは，第 5.6 節で述べた最小群間距離と組み合わせて用いる方法である．分類可能性予測は，予測器構築に用いる学習データ群と予測対象データの性質の類似度により予測性能に差が生じることは第 5 章で述べた通りであるが，その類似度を示す指標として最小群間距離が利用可能であることが考えられる．そこで予測器により得られた分類可能性と予測器の学習データ群と予測対象データとの最小群間距離を合わせて提示することで，予測対象データがどの程度分類問題として期待できるかということと予測された分類可能性がどの程度尤もらしいかということを同時に考慮することができると考えられる．

しかし，最小群間距離と分類可能性の予測性能の関係性を確かめるためには，第 5.6 節の試行だけでは参考程度の値しか示せていない．この試行では，8 割程度の予測性能を示した実データの半データ学習モデルおよび Leave-One-Out 交差検証における最小群間距離の中

央値 0.0616 ～ 0.1316 程度を示せれば同程度の予測性能を示せるだろうということが示されたが，これは本研究で用いた実データ群（表 4.3 参照）における試行に過ぎないことから他のデータにおいてこの値がどの程度必要となるかがわからない．そのため，実際に分類可能性予測システムとして運用するためには一般にどの程度の最小群間距離を確保できれば十分であるかという議論が必要であると考えられる．

また，複数の分類可能性予測器を構築しておき，それぞれの予測器の構築に用いた学習データ群と予測対象データの群間最小距離を求めた上で，最も群間最小距離が小さい学習データ群を用いて構築された分類可能性予測器を用いて予測を行うことで，より良い分類可能性予測を行えるだろう．この方法を用いる場合，分類可能性予測器をメタ特徴空間上におけるいくつかのグループに事前に分けられた学習データ群をグループごとに用いて構築された分類可能性予測器を用いることで，予測対象データと類似した特徴を持つグループによる予測が行えると考えられる．

次に考えられるのは，分類可能性は様々な閾値により定義することが可能な値であることから，いくつかの閾値における分類可能性を予測し組み合わせる方法である．本論文では，分類可能性を計算する際の閾値の一例として 0.75 という値がそのデータの予測性能においてどのようなケースで考えられるかを第 3.2 節で紹介した上で，いくつかの試行では $0.65, 0.70, 0.75, 0.80, 0.85$ という 5 つの閾値を用いていた．第 3.2 節で述べた通り，どの程度の分類性能を示せれば良好な予測であるかということはその結果の利用方法や分析主体の考え方に左右されるため一概には言えない．しかし，本研究のユースケースが非専門家に対し対象データがどの程度分類問題における期待度を持っているかを提示することであるということを考えると，一つの閾値を決めてその閾値における分類可能性を提示するか，複数の閾値における分類可能性を提示することが考えられる．現段階では前者を分類可能性として提示することとなり，その場合は提示する際に利用した閾値における注釈を加える必要があると考えられるが，後者を採用した場合，各閾値における分類可能性をどのように組み合わせて提示するかということが次の課題となると考えられる．

6.3 分類可能性予測では対処できない問題

分類可能性予測システムは第 5 章で述べた通り，予測器の構築に用いる学習データの性質により予測対象データの分類可能性をどの程度予測できるかが変化する．しかし，仮にあらゆる予測対象データに対しても高い予測性能を達成できる学習データを用意できたとしても，本研究のユースケースである未知データの価値判断を行ううえでこのシステムで対処できない問題も存在している．ここではそのうちいくつかを取り上げてその対応策について議論する．

1 つ目は，分類可能性予測システムで想定されていない形式に関わる問題である．分類可能性予測システムは，説明属性とクラス属性を用意した上でそれらの説明属性でクラス属性を予測可能であるかを予測するものである．そのため，データ自体はあるが予測したい対象

が不明であるケースや，用意されているデータが説明属性ベクトル $\mathbb{X} = \{x_1, x_2, \cdots, x_n\}$ (n は属性数）およびクラス属性 y として任意のモデル $\mathcal{M}$ を用いて $y = \mathcal{M}(\mathbb{X})$ と表せないケースでは利用できない．それらのケースにおいては分類可能性予測システムの利用の前に，予測対象のクラス属性を決定し，予測に用いる一連の説明属性をベクトルとして用意する必要がある．

2 つ目は，分類可能性抽出におけるコストに関わる問題である．本論文で提案した分類可能性予測器の構築方法では学習データ群を拡張しようと考えた際に，分類可能性抽出における計算コストの高さが問題となる．分類可能性予測システムにおいて，分類可能性予測器は事前に構築したものを提供することを想定しているが，その構築に用いる学習データ群を作成する際に分類可能性を抽出する必要がある．その際，分類可能性は第 3.2 節で述べた通り複数の分類器構築アルゴリズムを用いて計算され，これらは計算コストが高い．そのため，特に大規模なデータセットについて分類可能性を計算しようと考えた場合は何らかの対策を行う必要がある．

3 つ目は，メタ特徴抽出におけるコストに関わる問題である．分類可能性予測器の構築に用いる学習データ群および予測対象データにおいてメタ特徴を抽出する際，特に相互情報量を用いたメタ特徴において計算コストの高さが問題となる．本論文で提案する分類可能性予測で用いるメタ特徴としての相互情報量は各属性でクラス属性との間で計算し，その平均や最大値，また相互情報量を用いたいくつかの値を計算する．またその際，全ての値を離散確率変数として扱っているため，特に実数における相互情報量の計算はコストが高いものとなっている．そのため，2 つ目の問題と同様に何らかの対策を行う必要があると考えられる．また，2 つ目の問題と比較して，こちらはユーザが利用する際の計算コストに関わるため，より対策すべきであると言える．

1 つ目の問題への対策は，システムからユーザに対し適切な補助を行うことである．その具体例として，システムへの入力時にデータの形状を特定し，どのように変形することでシステムで利用可能になるかという提示を行うことがある．この対策は，ユーザがどのようなデータを入力し得るかということを分析したうえで，どのように提示することで解決できるかを検証する必要があると考えられる．

2 つ目と 3 つ目の問題への対策は，計算コストを低減させることである．その具体例として，各データセットのサブセットを用いてメタ特徴および分類可能性を抽出することがある．この対策はそれぞれの計算コストに直結するデータセットのインスタンス数を低減することを目的としている．しかし，作成したサブセットと元のデータセットから抽出したそれぞれのメタ特徴および分類可能性が分類可能性予測に与える影響が許容できるものであるかを確かめる必要があると考えられる．

第 7 章

結論

7.1　まとめ

本研究は，未知のデータセットにどの程度の分類性能を見込めるかを簡易的に調べるという目標のため，分類可能性という指標を定義し，メタ特徴からそれを予測する方法を提案し議論した．

分類可能性の定義は，第 2.3 節および第 3.3 節で述べたように，複数の指標を複数のアルゴリズムから求めることでデータセットの分類タスクにおける期待度を求めることができると考えられる．

そして，メタ特徴から分類可能性を予測することで，第 2.4 節にて紹介した他の研究のようにデータセットから分類可能性を抽出するのに比べて計算コストを低減させることが可能であり，第 5.1 節で述べたように構築済みモデルを用意することで分類性能を簡易的に調べることに繋がると考えられる．

分類可能性予測器の性能は，第 5 章で述べた通り学習データ群と予測対象データとの関係性により変化すると考えられ，第 4.4.2 節や第 5.3 節で得られた 8 割程度の予測性能を示す関係性の目安は第 5.6 節で示した通り群間最小距離により議論できると考えられる．

分類可能性予測は，データサイエンスの知見を持たない実体があるデータを分析したいと考えた際にコストを投じる前にそのコストに見合った分析結果が期待できるかを簡易的に判断する指標として利用されることが期待されており，それによりデータマイニングの活用促進に繋がると考えられる．しかし，第 6 章で示したように，分類可能性予測器より得られた予測結果の利用方法は分類可能性予測システムを構築する際に検討する必要があると考えられることに加え，分類可能性予測では対処できない問題に関してはシステム利用者に留意させる，あるいはそれを補助する仕組みが必要であると考えられる．

結果的に，分類可能性はデータセットの分類タスクにおける期待度を求めることができる指標の一つであると考えられ，分類可能性予測は本論文で定義した分類可能性を予測することが可能な手法であると言える．

7.2 今後の展望

第 6 章で示したように，分類可能性予測システムを構築するためにはいくつかの問題が残されている．そのため今後システム構築を行う際にはそれらの問題の対策を行なう必要があると考えられる．

分類可能性予測システムを構築する際にまず問題となるのは，第 6.2 節で述べた分類可能性の提示および利用の方法についてである．前述の通り，本論文で提案した分類可能性予測器はある閾値における分類可能性を出力する．しかし，閾値と識別性能の対応を表 3.1 に示した通り，分類可能性の計算に用いる閾値は目的により変化することが考えられる．そのため複数閾値における分類可能性を予測し，それらを組み合わせた結果を提示することがユースケースに適合していることも考えられる．また，第 5.6 節で述べた分類可能性予測器を構築する際の学習データ群と予測対象データの群間最小距離は分類可能性の予測性能に影響すると考えられるため，分類可能性予測の正確さを表すために群間最小距離と分類可能性を組み合わせた結果を提示することも期待度の参考値として有用であると考えられる．ただし，群間最小距離が高い値を示した場合，すなわち学習データ群に対し予測対象データで予測された分類可能性が参考になりにくいものであると判断された場合であっても，事前構築した分類可能性予測器を用いているならばユーザはそれをどのように扱えば良いのかという部分で問題が残ることとなる．

次に問題となるのは第 6.3 節で述べた，現状の分類可能性予測だけでは対処できない問題についてである．本論文では 3 つの問題点を挙げたが，大別すると形式の不一致と計算コストの 2 点に要約される．形式の不一致に関してはどの程度ガイドするかを検討した後に分析および検証を行う必要があり，計算コストはどの程度低減すべきであるかを検討した後に計算コストの低減を行う必要がある．これらの問題は本論文の目的の次のステップで検討すべき課題であると考えられ，特にユーザ補助の観点で言えば前述の分類可能性の提示方法と共に考えるべき課題であると言える．

分類可能性がデータマイニングに関する意思決定ののための一因となり，データマイニングの更なる発展に繋がるだろう．そのためには前述の課題が解決される必要があると考えられる．

謝辞

本研究を進めるにあたり，研究内容やその方針に関するご指導をいただきました公立はこだて未来大学システム情報科学部情報アーキテクチャ学科の新美礼彦教授，ならびに研究の助けになる様々な助言をいただきました同大学同学部複雑系知能学科の高橋信行教授，情報アーキテクチャ学科の白石陽教授に心から感謝いたします．

The authors would like to thank Enago (www.enago.jp) for the English language review.

発表・採録実績

発表等

[1] 鳴海雄登, 新美礼彦, “データの性質を用いた分類性能予測に関する検討”, 情報処理学会研究報告, Vol. 2020-DBS-171, No. 3, pp. 1–7, 2020 年 9 月 5 日, オンライン口頭発表.

[2] 早川雄登, 新美礼彦, “メタ特徴を用いた分類可能性予測”, DEIM Forum 2021, E11-4, pp. 1–6, 2021 年 3 月 1 日, オンライン口頭発表.

[3] 鳴海雄登, 新美礼彦, “分類可能性予測における学習データに求める性質の検討”, 情報処理学会研究報告, Vol. 2021-DBS-173, No. 4, pp. 1–6, 2021 年 9 月 16 日, オンライン口頭発表.

参考文献

[1] 和泉潔, 坂地泰紀, “特集:「ファイナンスにおける人工知能応用」にあたって”, 人工知能, Vol. 36, No. 3, pp. 260–261, 2021.

[2] 清田陽司, 大向一輝, “特集:「図書館情報学と AI の新展開」にあたって”, 人工知能, Vol. 35, No. 6, pp. 744–747, 2020.

[3] L. Kotthoff, C. Thornton, H.H. Hoos, F. Hutter, K. Leyton-Brown, “Auto-WEKA 2.0: Automatic Model Selection and Hyperparameter Optimization in WEKA”, Journal of Machine Learning Research, Vol. 18, No. 1, pp. 1–5, 2017.

[4] M. Feurer, K. Eggensperger, S. Falkner, M. Lindauer, F. Hutter, “Auto-Sklearn 2.0: The Next Generation”, arXiv:2007.04074, 2020.

[5] “Auto ML Tables |Google Cloud”, https://cloud.google.com/automl-tables (Accessed Jan. 4, 2022).

[6] “IBM Watson Studio – AutoAI |IBM”, https://www.ibm.com/cloud/watson-studio/autoai (Accessed Jan. 4, 2022).

[7] E. Jun, M. Daum, J. Roesch, S. Chasins, E. Berger, R. Just, K. Reinecke, “Tea: A High-level Language and Runtime System for Automating Statistical Analysis”, Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology, pp. 591–603, 2019.

[8] “Amazon SageMaker Canvas を発表 – ビジネスアナリスト向けの視覚的でノーコードの機械学習機能 |Amazon Web Services ブログ”, https://aws.amazon.com/jp/blogs/news/announcing-amazon-sagemaker-canvas-a-visual-no-code-machine-learning-capability-for-business-analysts/ (Accessed Jan. 4, 2022).

[9] F. Hutter, L. Kotthoff, J. Vanschoren, “Automated Machine Learning”, Springer, 2019.

[10] L. Weng, “Meta-Learning: Learning to Learn Fast”, https://lilianweng.github.io/lil-log/2018/11/30/meta-learning.html (Accessed Jan. 4, 2022).

[11] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, T. Lillicrap, “Meta-Learning with Memory-Augmented Neural Networks.”, International Conference on Machine

Learning, pp. 1842–1850, 2016.

[12] G. Koch, R. Zemel, R. Salakhutdinov, “Siamese Neural Networks for One-Shot Image Recognition”, ICML Deep Learning Workshop, Vol. 2, 2015.

[13] C. Finn, P. Abbeel, S. Levine, “Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks”, International Conference on Machine Learning, pp. 1126–1135, 2017.

[14] A. Arjmand, R. Samizadeh, M.D. Saryazdi, “Meta-Learning in Multivariate Load Demand Forecasting with Exogenous Meta-Features”, Energy Efficiency, Vol. 13, No. 5, pp. 871–887, 2020.

[15] L.P. Garcia, F. Campelo, G.N. Ramos, A. Rivolli, A.C.D.L. Carvalho, “Evaluating Clustering Meta-Features for Classifier Recommendation”, Brazillian Conference on Intelligent Systems, pp. 453–467, 2021.

[16] N. Seliya, M. Taghi, and J.V. Hulse, “A Study on the Relationships of Classifier Performance Metrics”, 2009 21st IEEE International Conference on Tools with Artificial Intelligence, pp. 59–66, 2009.

[17] S. Tavasoli, “Top 10 Machine Learning Algorithms List [2021 Updated]”, `https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article` (Accessed Jan. 4, 2022).

[18] 早川雄登, 新美礼彦, “メタ特徴を用いた分類可能性予測”, DEIM Forum 2021, E11-4, pp. 1–6, 2021.

[19] 鳴海雄登, 新美礼彦, “データの性質を用いた分類性能予測に関する検討”, 情報処理学会研究報告, Vol. 2020-DBS-171, No. 3, pp. 1–7, 2020.

[20] A. Filchenkov, A. Pendryak, “Datasets Meta-Feature Description for Recommending Feature Selection Algorithm”, 2015 Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT), pp. 11–18, 2015.

[21] G. Wang, Q. Song, H. Sun, X. Zhang, B. Xu, Y. Zhou, “A Feature Subset Selection Algorithm Automatic Recommendation Method”, Journal of Artificial Intelligence Research, Vol. 47, pp. 1–34, 2013.

[22] C. Castiello, G. Castellano, A.M. Fanelli, “Meta-data: Characterization of Input Features for Meta-Learning”, International Conference on Modeling Decisions for Artificial Intelligence, pp. 457–468, 2005.

[23] D. Michie, D.J. Spiegelhalter, C.C. Taylor, J. Campbell, “Machine Learning, Neural and Statistical Classification”, Ellis Horwood, 1994.

[24] T.K. Ho, M. Basu, “Complexity Measures of Supervised Classification Problems”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 3,

pp. 289–300, 2002.

[25] S.S. Stevens, “On the Theory of Scales of Measurement.”, Science, Vol.103, No.2684, pp.677–680, AAAS, 1946.

[26] 足立浩平, “多変量カテゴリカルデータの数量化と主成分分析”, 心理学評論, Vol.43, No.4, pp.487-500, 心理学評論刊行会, 2000.

[27] D. Harris, S. Harris, “Digital design and computer architecture”, Morgan Kaufmann, 2012.

[28] 鳴海雄登, 新美礼彦, “分類可能性予測における学習データに求める性質の検討”, 情報処理学会研究報告, Vol. 2021-DBS-173, No. 4, pp. 1–6, 2021.

図目次

表目次