Vol.2020-DBS-171 No.3
Vol.2020-IFAT-140 No.3
2020/9/5

データの性質を用いた分類性能予測に関する検討

鳴海 雄登[1,a)]　新美 礼彦[2,b)]

概要：近年，様々な分野でデータマイニングが活用されている．しかし，分析を行おうと考えている全ての人がデータマイニングの知識を持っているわけではないため，分析を行うためにコストが生じることが考えられる．さらに，データに有用なパタンが含まれているとは限らないため，データ自体に分析する価値があるかどうか見定めるのは難しい．そこで我々は，データの性質に着目し，本格的なデータマイニングを行う前にデータを分析した結果，期待される結果が得られるかどうかを判定するシステムの構築を目指している．本稿では，データを分類した際の性能を回帰分析によって求められる可能性を検証した結果，7 つのデータの性質を用いた Ridge 回帰で決定係数が 0.762 となった．今後は，同様の方法を用いてより複雑なアルゴリズムによる分類性能を予測することで，期待される結果が得られるか事前に判断できるか検討する．

Prediction of Classification Performance Using Data Properties

1. はじめに

近年，様々な領域でデータマイニングが応用されている．例えば，土木工学分野では道路舗装管理における地理空間情報とデータマイニングの適用 [1] や，海洋工学分野では造船工場内のヒヤリハット報告のデータ分析 [2] が行われている．

1.1 問題点

しかし，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ある．これはデータマイニングを行いたいと考える主体が，必ずしもデータマイニングの専門家であるとは限らないことに起因していると考えられる．データマイニングに対する専門知識を持たない結果，データマイニングを専門とする外部機関や企業への委託，あるいはデータマイニングを行うことのできる人材の育成が必要になるためである．

また，データマイニングを行うためにコストを投じた際に，実際に有益な知識を抽出できるとは限らないという問題もある．これは，分析対象のデータというものは環境から抽出したサンプルであり，そのデータ自体に有用なパタンが含まれているとは限らないことに起因していると考えられる．従来は，知識発見に有用なパタンがデータに含まれているかということを，データマイニングによって判断していたが，それには上述の通りコストがかかることとなる．

1.2 目的

これらの問題に対し，我々は低コストでデータから有用な知識が抽出できるかを判断することによる解決を検討している．それは，データ自体から有用なパタンが抽出できるかを簡易的に判断するシステムの構築を目標とし，その結果を用いてその後の意思決定が行えるのではないかと考えている．しかし，現段階ではこのシステムを構築するための構成要素が不明となっている．

そこで，検討しているシステムを構築するための構成要素として，1 つの方法を提案する．その方法とはデータの性質に着目し，複数のデータセットから抽出したデータの性質を用いて，それらのデータセットを用いた際の分類性能を目的変数とした回帰分析を行うというものである．

本稿では，その方法を用いた実験結果から 2 つの目的に対して検討を行った．1 つ目の目的は，そのデータセット

[1] 公立はこだて未来大学大学院 システム情報科学研究科
Future University Hakodate,
Graduate School of Systems Information Science
[2] 公立はこだて未来大学 システム情報科学部
Future University Hakodate,
Faculty of Systems Information Science
[a)] g2120036@fun.ac.jp
[b)] niimi@fun.ac.jp

を用いた分類を行った際の性能を回帰分析により求められないかということである．2 つ目の目的は，その際に用いるデータの性質も未知であるため，どのような性質から予測が可能であるかということである．

1.3 データマイニングフローにおける本研究の位置付け

通常のデータマイニングフローでは，大別して “データウェアハウシング”，“前処理”，“パタン発見”，“解釈” とされているが，本研究では “データウェアハウシング” と “前処理” の間に該当していると考えられる．

この研究や今後提案するシステムでは，データ自体にデータマイニングを行う価値があるかを判断することを目的としている．それによって，データ自体に有用なパタンが含まれているかを，パタン発見を行う前に判断することができると考えられる．その結果，計算コストの低減を狙えるのではないかと考える．

2. 関連研究

本研究に関連のある研究は多岐に渡る．その中でも，非専門家でも扱えるツールという観点のものと，メタ学習に関するもの，データに関する説明を行うものを紹介する．

2.1 非専門家でも扱えるツール

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高レベル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 [3]．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，特定の質問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ことがある．さらに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．

本研究の目標はデータマイニングの自動化ではなく，あくまでデータウェアハウシングにおいてユーザの負荷を軽減することであるため，データマイニングにおける Tea とは言えない．しかし，Tea の論文では Tea による判断と専門家の選択との一致についての議論や，専門家以外のユーザが犯したミスの影響のシミュレートが行われているため，本研究でも参考にできるのではないかと考えている．

2.2 メタ学習

本研究はデータ自体に予測を行うための有用なパタンが含まれているかを判断することを目標としているが，これはデータの性質を用いてモデルの性能を予測するということに関連していると考えられる．メタ学習には，データの特徴を用いて，ハイパーパラメタの自動チューニングや機械学習モデルの自動構築などでモデルの性能予測や評価を行っている研究がある．

南保らはデータから抽出した 5 種類のメタ特徴（データの性質）を用いて，5 種類の識別機構築アルゴリズムの中から最適なアルゴリズムを選択する識別機構築アルゴリズム自動選択システムの提案を行った [4]．この論文で提案されたシステムでは，先行研究である Nakamura らのシステム [5] においては 26 種類のメタ特徴を用いていたのに対し，特徴選択を複数回適用することでより少ないメタ特徴を用いて正解率を向上させた．

また，Auto-Sklearn[6] では，メタ学習だけでなくメタ学習により選択された候補となる前処理や学習アルゴリズム，ハイパーパラメータなどのセットをベイズ的最適化で選択し，選択されたセットを用いたアンサンブル学習によって自動的な機械学習を行っている．これにより従来の自動機械学習（AutoML）手法と比較して，性能の向上と同時に効率性とロバスト性を向上させた．

上述の通り，本研究の目標はデータマイニングの自動化を行うことではない．しかし，南保らの用いた特徴選択手法や，Auto-Sklearn においてメタ特徴が類似するデータセットと同様の性質を持っていると考える方法が本研究でも参考にできるのではないかと考えられる．

2.3 データに関する説明

従来のデータに関する説明は，統計的分析やデータマイニングの結果によって行われていた．

近年の動向として，深層学習などのブラックボックスモデルが様々な領域で用いられている．ブラックボックスモデルは明瞭なモデルに比べて予測性能が高くなるなど様々な利点があると考えられているが，研究対象のデータに含まれるバイアスを特定できないというケースが多々見られる．この問題に対して，説明可能 AI（XAI）を活用してブラックボックスモデルに生じる学習データに依存するバイアスを特定する試みがいくつか行われている．

Angelino らは彼らの提案した解釈可能モデルである CORELS と COMPAS データセットを用いたブラックボックスモデルを比較し，解釈可能モデルはブラックボックスモデルに対してモデルの公平性について議論することが遥かに簡単であると示した [7]．

また，Lakkaraju らはブラックボックスモデルの事後説明手法において，忠実度の高い説明がブラックボックスモデルにおけるバイアスを正確に反映しない場合に，人間が誤解し信頼することの危険性を述べ，誤解を招く可能性のある説明を生成するアプローチを提案することでユーザの信頼にどの程度影響を与えるかを調べた [8]．

本研究では，与えられたデータ自体からそのデータを用いた分類モデルの性能を予測するため，データの属性に関するサブセットを用いた場合の分類モデルの性能の比較を行うことにより，そのデータの持つバイアスを予測できると考えられる．

3. 提案手法

本稿で提案する手法は，データの性質を用いた分類性能予測である．この手法により，分類性能の予測が可能であれば，我々が構築を検討しているシステムの構成要素として用いることが可能であると考えられる．この手法には 3 つの段階が含まれている．

3.1 前処理

この手法では複数のデータセットからデータの性質を抽出する必要があるため，データフォーマットを揃えるために前処理方針の固定化が必要である．

しかし，この研究が目標とするシステムのユースケースは，データマイニングに対する知識を持たないユーザによる利用を考えているため，前処理も簡易的に行えるものとする必要がある．そのため，以下の方針とした．

- 欠損値
  これはカテゴリ属性と数値属性で大別して考える．
  - カテゴリ属性
    欠損カテゴリとして扱う．
  - 数値属性
    Global Common Value として平均値を代入する．
- カテゴリ属性
  基本的にダミーコーディングとし，2 値属性に対しては {0,1} とする．
- 属性選択
  固有な値，例えば事例固有の ID や事例の名前など，を削除する．

固有な値であるかの判定において，判断の難しいカテゴリ属性がいくつか存在していた．そこで，ユニークな値の種類数と全事例数との比率をユニーク比として考え，それが 50%以上のものを固有な値として判断した．

3.2 データの性質の抽出

前処理済みの複数のデータに対し，実際に分類器の構築・データの性質の抽出を行い，学習データを作成する．

抽出するデータは以下のとおりである．

- 分類器の Accuracy と F1 値
  この手法で予測する分類性能である．
  多値分類問題においては F1 値のマクロ平均を使用する．
- 目的属性のクラス数
- 目的属性の情報エントロピー
- 目的属性と各説明属性との相関
  同時にこの相関から相関の二乗を求める．
- 目的属性に対する各説明属性の情報利得比
  ここではユニーク比が 50%未満の属性について求める．
- 決定木のクラス決定に対する各説明属性の寄与度

また，相関，相関の二乗，情報利得比，寄与度については，以下の値を計算する．

- 平均値
- 標準偏差
- 中央値
- 第一四分位
- 第三四分位
- 四分位範囲
- 最小値
- 最大値
- 値域
- グラフ上の面積類
  - Over-Mean
  - Over-Median
  - Under-Mean
  - Under-Median
- 各グラフ上の面積同士の差
- 各グラフ上の面積同士の商

ここで，グラフ上の面積類は値を昇順でグラフにプロットした際に，中央値，平均値をからの絶対差の総和を，中央値，平均値を上回る値，下回る値でそれぞれ算出したものである．

3.3 回帰による分類性能予測

この段階は，大まかに 3 つに分けて考えられる．

(a) ElasticNet を用いた属性選択

(b) 選択された属性間の相関による属性選択

(c) 回帰モデルの構築，評価

データの性質の抽出によって 78 属性が得られるが，スパース性向上のために ElasticNet によって属性選択を行っている．また，その属性選択では属性間の相関が考えられていないため，選択された属性間の相関が高いものから選択する．これは，以下の手順で属性選択を行う．

(1) 相関の二乗からどの程度相関しているかを考える

(2) 値の分布とデータの生成背景を考える

(3) ElasticNet で利用された回数を比較する

(4) 回帰係数を検定により比較する

これらについて詳細に述べる．(1) では，主に相関の

二乗が 0.5 を超えているものに関して属性選択を行うか否かを判断する．(2) では，現時点では手動による判断となるが，実験においてこの手法の有効性が確認できた際には，“データの性質の抽出” において抽出する方法について検討し自動化を図る．(3) では，実験において $\alpha = \{0.1, 0.5, 1.0, 5.0, 10.0\}$ で比較した結果，最も R^2 が高くなった α を用いた ElasticNet で利用された回数の比較を行う．(4) では，回帰アルゴリズムに Lasso 回帰，Ridge 回帰，ElasticNet を用いた結果 R^2 が正となったケースにおける回帰係数を用いる．これらのハイパーパラメタは α のみを調整し，(3) と同様に $\alpha = \{0.1, 0.5, 1.0, 5.0, 10.0\}$ とする．また，検定には正規分布であるかを Shapiro-Wilk 検定により確認し，双方が正規分布に則していると判断できた場合は t 検定，少なくとも一方が正規分布に則していると判断できない場合は U 検定を用いた．検定の結果，有意に回帰係数が高い属性を用いることとした．

その後，それらのデータを用いて回帰モデルを構築し，評価を行う．

4. 実験

実験では，提案手法を用いて本稿の 2 つの目的のそれぞれを実験目的として検討を行う．

4.1 実験目的

1 つ目の目的は，多数のデータセットから抽出したデータの性質から，回帰分析によってそのデータセットを用いた分類を行った際の性能を求められるかを確かめるということである．

2 つ目の目的は，1 つ目の目的においてどのようなデータの性質を用いることで予測が可能であるかを確かめるということである．

4.2 実験手順

前述の提案手法に基づいて実験を行う．

具体的な手順は以下のとおりである．

(1) 前処理方針に従った前処理

(2) データの性質の抽出

(3) 回帰による分類性能予測

- (a) ElasticNet を用いた属性選択
- (b) 選択された属性間の相関による属性選択
- (c) 回帰モデルの構築，評価

データセットは UCI Machine Learning Repository[9] と UCR and UEA Time Series Classification[10] から 53 個収集した．

また，分類には決定木構築アルゴリズムである CART と多層パーセプトロン (MLP) を用い，分割数 10 の交叉検証の結果の Accuracy と F1 値を分類性能とした．

回帰分析の性能評価には決定係数 R^2 と二乗平均平方根誤差 ($RMSE$) を用い，Leave-One-Out 交叉検証によって評価した．

4.3 ElasticNet を用いた属性選択の結果

CART と MLP の Accuracy と F1 値のそれぞれで選択された属性集合の和集合として以下の 12 の属性が得られた．

- 目的属性のクラス数
- 目的属性の情報エントロピー
- 最大寄与度
- 最大寄与度と最小寄与度の差
- 寄与度の Over-Mean
- 寄与度の Under-Median
- 寄与度の Under-Mean
- 寄与度の Under-Mean と Under-Median の差
- 相関係数の中央値
- 相関係数の二乗の標準偏差
- 相関係数の二乗の中央値
- 相関係数の二乗の Over-Mean と Over-Median の商

4.4 属性間の相関による属性選択の結果

“ElasticNet を用いた属性選択” により選択された 12 の属性で，“選択された属性間の相関による属性選択” を行った結果，いくつかの属性間で 0.5 を超え強い相関を示し，以下の 6 属性と検定により有意差が認められなかった 2 属性が得られた．

- 目的属性のクラス数
- クラス数で正規化された目的属性の情報エントロピー
- 最大寄与度と最小寄与度の差
- 相関係数の中央値
- 相関係数の二乗の標準偏差
- 相関係数の二乗の Over-Mean と Over-Median の商
- 有意差が認められなかった 2 属性
  - 寄与度の Under-Mean
  - 寄与度の Under-Mean と Under-Median の差

“クラス数で正規化された目的属性の情報エントロピー” については，“目的属性の情報エントロピー” と “目的属性のクラス数” との間に強い相関が生じたが，値の分布とデータの生成背景を考えた結果，各クラスが均等にデータに含まれている場合，目的属性のクラス数に対して目的属性の情報エントロピーが単調増加するため，情報エントロピーをクラス数で正規化することにより相関を解消した．

4.5 回帰モデルの構築，評価の結果

回帰分析には，“属性間の相関による属性選択の結果” における “回帰係数を検定により比較した” と同様に，回帰アルゴリズムに Lasso 回帰, Ridge 回帰, ElasticNet, ハイパー

パラメタは α のみを調整し，$\alpha = \{0.1, 0.5, 1.0, 5.0, 10.0\}$ とした．

回帰分析の評価結果を**表 1** に示す．データ名列の書式は “分類アルゴリズム名-評価値-{1,2}” とし，有意差が認められなかった 2 属性のうち “寄与度の Under-Mean” を用いたものを “**-**-1”，“寄与度の Under-Mean と Under-Median の差” を用いたものを “**-**-2” とした．

表 1 回帰分析の結果
Table 1 Result of Regression

データ名	アルゴリズム	R^2	$RMSE$
CART-F1-1	Ridge($\alpha = 5.0$)	0.762	0.156
CART-F1-2	Ridge($\alpha = 5.0$)	0.761	0.157
MLP-F1-1	Ridge($\alpha = 5.0$)	0.749	0.159
MLP-F1-2	Ridge($\alpha = 1.0$)	0.756	0.162
CART-Acc-1	Ridge($\alpha = 1.0$)	0.661	0.146
CART-Acc-2	Ridge($\alpha = 1.0$)	0.658	0.146
MLP-Acc-1	Ridge($\alpha = 5.0$)	0.468	0.176
MLP-Acc-2	Ridge($\alpha = 5.0$)	0.459	0.176

また，実験結果が今回用いたデータに依存しているかを考察するために，「データセット毎の CART と MLP の F1 値の絶対差」と，CART と MLP の “**-**-1” と “**-**-2” それぞれの方法における「データセット毎の実際の F1 値と予測結果の絶対誤差」，またそれぞれの方法における「絶対誤差の降順順位」間の相関を求めた．

表 2 に「データセット毎の CART と MLP の F1 値の絶対差」と「データセット毎の実際の F1 値と予測結果の絶対誤差」，「絶対誤差の降順順位」を示す．

また，**表 3** に「データセット毎の CART と MLP の F1 値の絶対差」と，それぞれの方法における「データセット毎の実際の F1 値と予測結果の絶対誤差」，「絶対誤差の降順順位」との相関の二乗を示す．

5. 考察

本稿の目的は 2 つあり，それらを実験により検討した．

5.1 1 つ目の目的

1 つ目の目的は，多数のデータセットから抽出したデータの性質による回帰分析から，そのデータセットを用いた分類を行った際の性能を求められるかということである．この目的に対して，実験では，分類を行った際の性能として CART と MLP の Accuracy と F1 値を用いて，Ridge 回帰等の回帰アルゴリズムによって予測を行った．

実験結果の決定係数からは，CART と MLP の F1 値，CART の Accuracy において 0.5 を上回ったため，概ね性能が求められていると考えることができる．

しかし，RMSE においては決定係数で良好な結果が得られたデータであっても，0.146 ∼ 0.162 となった．これは，元々のデータが Accuracy と F1 値であるため 0.0 ∼ 1.0 を取るのに対し，事例毎に平均 ±0.146 ∼ 0.162 の誤差が生まれていることとなるため，無視できない大きさであると考えられる．この RMSE を最小限に抑えることが今後の課題であると考えられる．

また，表 3 に示した「データセット毎の CART と MLP の F1 値の絶対差」と，それぞれの方法における「データセット毎の実際の F1 値と予測結果の絶対誤差」，「絶対誤差の降順順位」との相関の二乗から，実験に用いた「データセット毎の CART と MLP の F1 値の絶対差」は提案手法に影響を与えていないと考えられる．これにより，CART と MLP の分類アルゴリズムの違いに起因する予測性能の差に影響されていないと考えることができるため，他のアルゴリズムに対しても同じように分類性能の予測が可能なのではないかと考えることができる．

5.2 2 つ目の目的

2 つ目の目的は，1 つ目の目的において用いるデータの性質を特定することである．この目的に対して，提案手法において述べたデータの性質の抽出を行った上で，ElasticNet を用いた L1/L2 正則化と相関を考えた手動による属性選択を行うことによって，分類性能予測に有効なデータの性質を考えた．

その結果，6 つの属性とどちらかを選択する必要のある 2 つの属性が選択された．そのうち，“目的属性のクラス数” と “目的属性の情報エントロピー” の間の強い相関については，値の分布とデータの生成背景を考えた結果，“クラス数で正規化された目的属性の情報エントロピー” とすることにより解消できたため，“データの性質の抽出” において予め考慮することを検討すべきである．

6 つの属性と選択する必要のある 2 つの属性のそれぞれを用いた回帰分析の，各データにおいて最も高い性能を示した結果を比較する．決定係数では，CART の F1 値と CART と MLP の Accuracy において “寄与度の Under-Mean” を用いた属性セット，MLP の F1 値においては “寄与度の Under-Mean と Under-Median の差” を用いた属性セットで他方に対し少し高い結果が得られたが，ほとんど変わらないとも言える．RMSE では，CART と MLP の F1 値において “寄与度の Under-Mean” を用いた属性セットが少し低く，CART と MLP の Accuracy においては両群間に違いが見られなかった．

そこで，2 つの属性のうちいずれかを用いたデータの回帰分析結果のうち，決定係数が正になっているもので，決定係数と RMSE に有意差があるかどうかを検定によって比較した．いずれの属性を用いた場合も決定係数が正になっている事例数は等しく，決定係数と RMSE のどちらについても有意差が見られず，片側検定の結果もどちらかが有意に高いという結果も得られなかった．

表 2 「F1 値の絶対差」と「絶対誤差」
Table 2 "Absolute Difference of F1 Value" and "Absolute Error"

データセット名	F1 値の絶対差	CART-F1-1		CART-F1-2		MLP-F1-1		MLP-F1-2	
		順位	絶対誤差	順位	絶対誤差	順位	絶対誤差	順位	絶対誤差
Annealing	0.3769	3	0.3467	3	0.3627	32	0.0528	33	0.0555
Heart Failure Clinical Records	0.3630	36	0.3839	43	0.3613	3	0.0398	3	0.0290
FordB TRAIN	0.2560	35	0.0476	46	0.0141	13	0.1506	13	0.1461
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Chess King Rook vs King	0.0002	51	0.0057	44	0.0280	49	0.0100	32	0.0555
Mushroom	0.0000	42	0.0278	34	0.0512	11	0.1212	16	0.1561
Monks-3	0.0000	16	0.1213	17	0.1158	7	0.1992	10	0.1852

表 3 「F1 値の絶対差」と「絶対誤差」の相関二乗
Table 3 Correlation Square of "Absolute Difference of F1 Value" and "Absolute Error"

CART-F1-1		CART-F1-2		MLP-F1-1		MLP-F1-2	
順位	絶対誤差	順位	絶対誤差	順位	絶対誤差	順位	絶対誤差
0.0134	0.0333	0.0013	0.0246	0.0072	0.0194	0.0156	0.0193

また，これらのデータの性質についても 1 つ目の目的の考察において述べた RMSE を最小限を抑えるという課題の解決方法次第では変化することが考えられる．

6. おわりに

ここでは，近年のデータマイニングにおいて生じている問題を背景に本稿で取り組んだ課題の解決案に関する検討についてのまとめと，今後の方針について述べる．

6.1 まとめ

近年，様々な領域でデータマイニングが応用されているが，データマイニングを行うために追加のコストが生じる場合があり，分析対象のデータから実際に有益な知識を抽出できるとは限らないという問題がある．そこで我々は，低コストでデータから有用な知識が抽出できるかを判断することによりこの問題を解決しようと考えているが，そのための判断基準が不明である．

そこで本稿では，データの性質に着目し，データの性質を用いて，分類性能を目的変数として回帰分析を行い，予測することが可能であるかということと，その際に有効なデータの性質を特定することを目的として，この手法が判断基準として用いることができるかを実験により確かめた．実験結果により，データから抽出した 7 つの性質を用いて行った回帰分析によって，CART と MLP の F1 値，CART の Accuracy に対して決定係数 R^2 が 0.5 を上回り，良好な予測ができたと考えられた．

6.2 今後の方針

提案手法により R^2 からは良好な結果が得られたと考えられるが，RMSE については目的属性である F1 値と Accuracy の値域に対して無視できない誤差が生じていると考えられた．そのため，この手法を判断基準として用いるには更なる RMSE の低減が必要であると考えられる．よって，考察において述べた通り，実験により求められた RMSE を低減させられるように対策を行うということが，今後第一に検討すべき課題である．

RMSE の低減について現時点で考えられる対策としては 3 つある．1 つ目は，回帰問題ではなく，何らかの基準により分類性能にラベル付けを行うことで分類問題とすること．2 つ目は，現状ほとんどの数値属性に対して情報利得比を算出していないため，適切な閾値を設けるなどの対処を行って情報利得比を導出し利用すること．3 つ目は，ElasticNet を用いた属性選択の前に予め属性選択を行うことである．

更に，実験に用いたデータセットの妥当性について，また実験では CART と MLP の分類結果を用いたが，更に複雑なアルゴリズムによる分類性能の予測にこの手法を用いることができるかということについても検討する必要があると考えられる．

実験に用いたデータセットの妥当性に関しては，データセットに対する分析を更に行い，データセットのメタ的性質に合わせた個別の分析を行えるように方針を変更することが考えられる．

今後の分類性能の予測で検討すべき複雑なアルゴリズムとしては，Random Forest や勾配ブースティングなどのアンサンブル学習や深層学習を用いた手法が上げられる．これは現状問題となっているブラックボックスモデルにおける説明可能性や解釈可能性の欠如に対する解決策となることも考えられる．考察において，今回実験で用いたアルゴリズムである CART と MLP による分類性能がどの程度異なるかによって予測性能が影響されないと考えられたため，他のアルゴリズムへの転用の可能性はあると考えられる．

参考文献

[1] 福士直子, 矢吹信喜, “道路舗装管理における地理空間情報とデータマイニングの適用に関する考察”, Proceedings of the symposium on civil engineering informatics, Vol. 41, pp.101–104, 2016.
[2] 篠田岳思, 田中太氏, 白神祐輔, “2016S-GS4-4 テキストマイニングよる造船工場内のヒヤリハット報告のデータ分析”, 日本船舶海洋工学会講演会論文集, Vol. 22, pp. 349–350, 2016.
[3] 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.
[4] 南保英孝, 大塚敦史, 木村春彦, 上田芳弘, “メタ特徴の最適化処理による識別器構築アルゴリズム自動選択システム”, 科学・技術研究, Vol. 5, No. 2, pp.179–184, 2016.
[5] M. Nakamura, A. Otsuka, H. Kimura, “Automatic Selection of Classiffication Algorithms for Non-Experts Using Meta-Features”, China-USA Business Review, Vol. 13, No. 3, pp. 199–205, 2014.
[6] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, F. Hutter, “Efficient and Robust Automated Machine Learning”, Advances in Neural Information Processing Systems, pp. 2962–2970, 2015.
[7] E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, C. Rudin, “Learning Certifiably Optimal Rule Lists for Categorical Data”, The Journal of Machine Learning Research, Vol.18, No.1, pp.8753-8830, 2017.
[8] H. Lakkaraju, O. Bastani, ““How do I fool you?”: Manipulating User Trust via Misleading Black Box Explanations”, Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp.79-85, 2020.
[9] D. Dua, C. Graff, “UCI Machine Learning Repository”, http://archive.ics.uci.edu/ml (accessed 2020-07-18), University of California, Irvine, School of Information and Computer Sciences, 2019.
[10] “UCR and UEA Time Series Classification Repository”, http://timeseriesclassification.com (accessed 2020-07-18).