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

鴨志田 亮太[1,a)]

概要：近年，データサイエンティストという職種に注目が集まる一方で，その人材不足が指摘されている．需要増という背景から，未習熟者がデータ分析に従事するケースが増えているが，経験・知識が不足している分析者が分析を行った場合，分析手順の誤りなどにより適切な分析を行うことができないことがある．そこで報告者は，未習熟者のデータ分析を支援することを目的として，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ツール MALSS を開発した．MALSS を利用することで，既存の機械学習ツールを利用した場合よりも質の高いデータ分析を行いながら，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となる．

キーワード：機械学習，データサイエンティスト，学習支援，MALSS

1. はじめに

ビッグデータの収集・蓄積が容易となったことで，データからビジネスに有効な知見を発見し，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．このデータサイエンティストに求められるスキルの一つに，機械学習が挙げられる [1]．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 [*1]，Jubatus [*2] 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，専門家でなくても容易に機械学習技術を利用することが可能となった．需要増とツールの充実という背景から，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，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．

そこで報告者は，この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を解決するために，機械学習支援システム (MAchine Learning Support System: MALSS) を，汎用プログラミング言語 Python のオープンソースライブラリとして開発した [7]．MALSS が備える機能は，1）自動化による分析支援，2）分析レポート作成による知識習得支援，の 2 つである．

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．

さらに，分析レポートを作成することで，分析プロセスをブラックボックス化せずに，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．

[1] 日立製作所
[a)] ryota.kamoshida.vw@hitachi.com
[*1] http://scikit-learn.org/stable/
[*2] http://jubat.us/

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

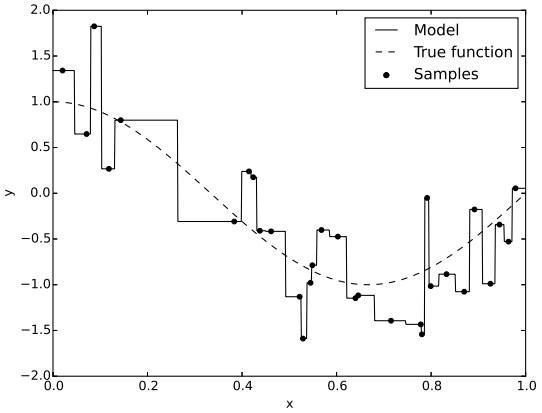

図 2 回帰分析結果（scikit-learn 利用）

2. 動機付け

本節では，動機付けの例として，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 1 は Python の機械学習ライブラリ scikit-learn 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 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 [5]．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．

3. MALSS

図 3 に，MALSS のユースケース図を示す．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 [*3]，および scikit-learn を利用している．

MALSS の主要な機能は，1）データの準備，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，の 5 つからなる．以下において，これらの機能について述べる．

3.1 データの準備

データの準備では，1）変数の種類に応じた欠損値の自動補間，2）ダミー変数を用いたカテゴリ変数の量的変数への変換 [2]，3）データの標準化処理 [5]，4）交差検証による汎化性能評価 [2] のためのデータシャッフル，を行う．

[*3] http://pandas.pydata.org/

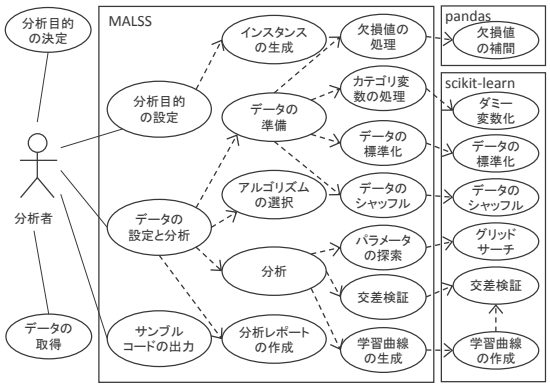

図 3 ユースケース図

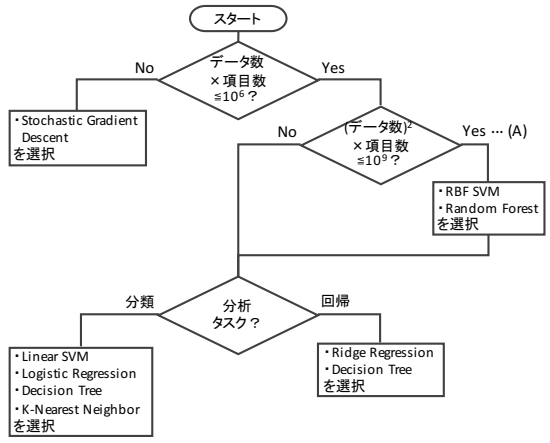

図 4 アルゴリズム選択ルール

3.2 機械学習アルゴリズムの選択

MALSS では，scikit-learn のチュートリアル [*4] のアルゴリズム選択チャートを参考に，独自のアルゴリズム選択ルールを作成した (図 4)．アルゴリズム選択は，処理時間が最も長くなる図 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ことを一つの目安としている．

3.3 分析

分析では，1）グリッドサーチ [6] によるハイパーパラメータ探索，2）汎化性能評価のための交差検証 [2]，3）分析指針を得るための学習曲線 [5] の生成，を行う．

[*4] http://scikit-learn.org/stable/tutorial/

3.4 分析レポートの生成

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や専門用語の解説なども併記することで，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，1）複数アルゴリズムの性能比較，2）データ概要およびデータ準備処理結果，3）各アルゴリズムの分析結果詳細（ハイパーパラメータ探索結果，モデルの性能，学習曲線など），から構成される．

3.5 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．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．

3.6 MALSS の使用方法

2 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 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，12 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

```
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 5 MALSS を用いて回帰分析する Python コード例

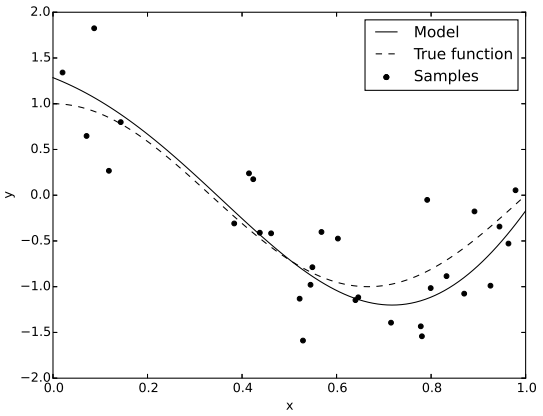

図 6 真の値と目的変数とモデルの予測値（MALSS 利用）

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 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．

4. 評価

1）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？ 2）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？という研究課題に対する 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．

4.1 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 のみ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や参考資料などを利用して分析を行う．

分析データは UCI Machine Learning Repository [*5] 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 7 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．

4.2 学習効果の評価

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った．

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．模擬データ分析実験で提示する資料に記

[*5] http://archive.ics.uci.edu/ml

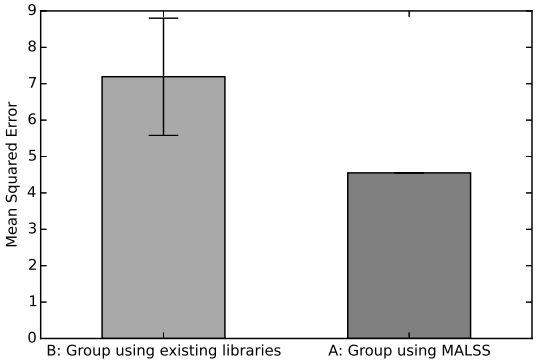

図 7 模擬データ分析実験結果

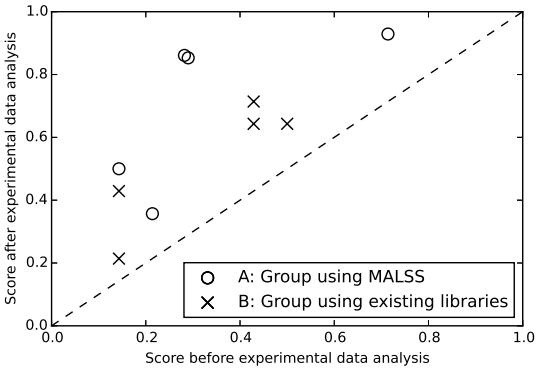

図 8 知識確認テスト結果

載されている情報をもとに正答できるように設計した．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．

知識確認テストの結果を図 8，および図 9 に示す．図 8 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．図 9 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．

4.3 研究課題に対する考察

模擬データ分析実験の結果，A グループの平均二乗誤差は全員が同一で，かつ，B グループの平均二乗誤差最小値よりも小さいことから，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ことを確認できた．

また，知識確認テストの結果，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たことから，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ことを確認できた．

5. 関連手法

scikit-learn には充実したチュートリアルページ

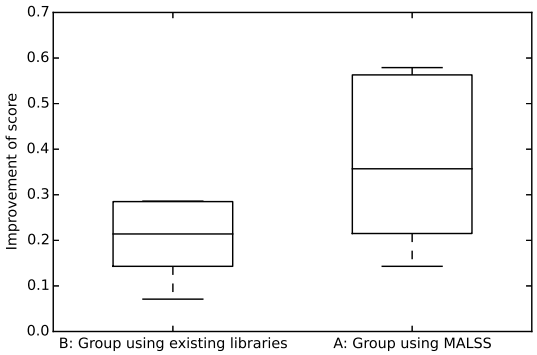

図 9 分析実験前後の知識確認テスト正答率差分

が用意されている．

統計解析向けプログラミング言語 R 向けには James らの書籍 [3] が無料で利用可能であり，関連した講義動画やスライドも充実している．

教育の分野では，Amershi らにより人工知能分野の指導・学習ツール AIspace が提案されている [4]．AIspace は人工知能分野に関連する技術領域を扱っており，機械学習技術もその中に含まれる．

このように，機械学習によるデータ分析を行うためのツール，および分析方法を学習するための資料は充実しており，分析者は独学でも十分必要な知識を身につけることができる．しかし，現実問題として，すべての分析者が十分な知識を習得した後に分析に従事するわけではない．報告者が開発した機械学習支援システム MALSS は，このような分析者を対象として，分析の自動化と分析レポートの作成を行うことにより，質の高い分析を行いながら，必要な知識を習得していくことを

可能とする．

6. おわりに

本報告では，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 を開発した．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レポート作成などが挙げられる．

MALSS はオープンソースライブラリとして公開しており，多くのユーザに利用してもらえるよう改良を続けていきたい．

参考文献

[1] Harris, H., Murphy, S. and Vaisman, M.: *Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work*, O'Reilly Media, Inc. (2013).

[2] Hastie, T., Tibshirani, R. and Friedman, J.: *The Elements of Statistical Learning*, Springer New York Inc., New York, NY, USA (2001).

[3] James, G., Witten, D., Hastie, T. and Tibshirani, R.: *An Introduction to Statistical Learning: with Applications in R*, Springer (2013).

[4] Knoll, B., Kisyski, J., Conati, C., Mackworth, A. and Poole, D.: AIspace: Interactive tools for learning artificial intelligence, *In Proceedings of the AAAI 2008 AI Education Workshop*, p. 3 (2008).

[5] Murphy, K. P.: *Machine learning: a probabilistic perspective*, The MIT Press, Cambridge, MA (2012).

[6] Richert, W. and Coelho, L. P.: 実践 機械学習システム，O'Reilly Japan, Inc. (2014).

[7] 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，Vol. J99-D, No. 4, pp. 428–438 (2016).

質疑・応答

伊知地　B グループの経験のある方の実験結果は？

鴨志田　他の実験協力者と同程度の成績であり，ノイズになるようなことはなかった．

美馬　既存のライブラリなどとの違いは？

鴨志田　機械学習のプロセスを自動化だけでなく，分析レポートを出力し，分析者の知識習得を支援するところ．

美馬　学習者のレベルに合わせる仕組みはあるのか？

鴨志田　今後の課題．現状はレポートの詳細説明を省略できる程度．

中山　アルゴリズムの選択理由が分かると良いのではないか．

鴨志田　確かに．是非加えたい．

角田　MALSS は誰でも自由に使えるのか？

鴨志田　YES．オープンソースとして公開している．

八木原　知識確認テストはどのようなものか？

鴨志田　（実物を見てもらう）分析プロセスの知識の有無を問うもの．

八木原　中級者以上が使うためにはどんな機能があれば良いか．

鴨志田　現状でもアルゴリズムの追加やパラメータの範囲設定が可能になっている．

八木原　分析を自動化してしまうことで，失敗から学ぶ機会を奪ってしまうのでは？

鴨志田　その懸念はある．考えていきたい．

伊知地　知識確認テストもシステムに組み込んではどうか？

鴨志田　良案と思う．そうすると利用者の知識に応じたレポートが作成できる．