2ZH-09

情報科目での成績分析に用いる学習ログと特徴量の検討

西岡克記† 望月久稔†
†大阪教育大学

1 はじめに

教育現場への Moodle や Google Forms などの導入により大量に蓄積されたログを用いることで授業の成績を分析する [1][2]．本論は情報科目の学習ログから得られる特徴を成績分析に用いて検証する．学部 1 回生対象の情報リテラシーに関する授業に対して学習ログから特徴を抽出し，Random Forest による授業の成績予測に用いる．予測に用いる特徴を予測結果に対する評価指標から検証する．

2 学習ログを用いた特徴量の検討

受講生の学習傾向や小テストでの得点状況として，学習ログから不要な部分を取り除いて授業成績の予測に用いる特徴量を抽出する．授業成績の予測は，抽出した特徴量の数を N として $\sum_{r=1}^{N} {}_n\mathrm{C}_r$ 通りの組み合わせそれぞれに対してパラメータチューニングによって得られる最適なモデルで Random Forest により回帰分析する [3]．

Moodle のコースに記録されているログから受講生に関する記録を抽出して，以下の特徴量を定義する．

- 小テストの平均値
- 小テストの偏差値
- 課題の提出回数
- 時間帯による学習の傾向

小テストの平均値と偏差値は授業内で実施した小テストを通してのものとする．時間帯による学習の傾向を定義の例として挙げる．

例 時間帯による学習の傾向

Moodle 内のコンテンツへアクセスする時間帯は受講生によって異なる．そこで 1 日を 6 時間ごとに分割し，朝 (6:00〜11:59)，昼 (12:00〜17:59)，夜 (18:00〜23:59)，深夜 (0:00〜5:59) とみなす．受講生の Moodle 内のコンテンツへの総アクセス数と朝，昼，夜，深夜におけるアクセス数を集計し，総アクセス数を分母にして朝，昼，夜，深夜におけるアクセスの割合を算出することで受講生がどの時間帯によく学習する傾向にあるかを設定する．

表 1: チューニングする Random Forest のパラメータ

n_estimators	決定木の数
max_depth	決定木の最大深さ
min_samples_split	サンプルを分割する条件
min_samples_leaf	葉に存在するサンプルの最小数

各受講生の授業における評点を目的変数とし，定義した特徴量を用いてデータセットを作成し，そのうち 7 割を学習データに，3 割を検証データに用いる．データの分割には scikit-learn [4] の train_test_split 関数を用いてランダムに分割する．scikit-learn にある GridSearchCV を利用して Random Forest のパラメータ (表 1) をチューニングするために，学習データを用いる．最適な Random Forest のパラメータとモデルを取得し，検証データを用いて授業成績を予測する．

3 授業成績の予測

2020 年度と 2021 年度に実施した学部 1 回生対象の情報リテラシーに関する授業における成績を予測する．予測結果に対する評価指標として，予測した成績と実際の成績の偏差平方和の比，決定係数 R^2，二乗平均平方根誤差 RMSE，平均絶対誤差 MAE を用いる．

まず，2020 年度実施分に関して授業成績を予測したところ，全組み合わせ中最も良かったのは朝に学習している割合，小テストの平均値を特徴量に用いた場合であった．

次に，2021 年度実施分に関して授業成績を予測したところ，全組み合わせ中最も良かったのは朝に学習している割合，夜に学習している割合，小テストの偏差値を特徴量に用いた場合であった．

Examination of learning log and features analyzing grade in the information subject
Katsunori NISHIOKA†and Hisatoshi MOCHIZUKI†
†Osaka Kyoiku University

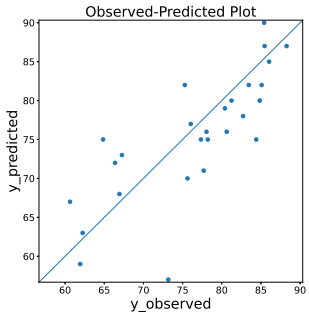

図 1: 2020 年度 予測-実際

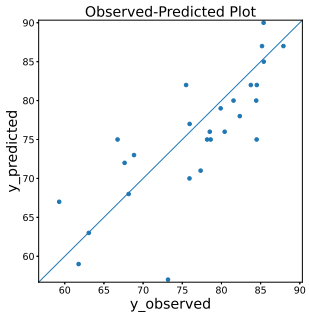

図 2: 2021 年度 予測-実際

表 2: 2020 年度，2021 年度それぞれの評価指標

年度	偏差平方和の比	R^2	RMSE	MAE
2020	0.82	0.65	4.70	3.68
2021	0.75	0.55	5.33	4.07

2020 年度，2021 年度それぞれの授業成績の予測で最も良かった組み合わせについて，実際の成績を横軸に，予測した成績を縦軸にした散布図を図 1，2 に示す．図中の対角線は $y = x$ の直線であり，その近傍に存在していることから実際の成績に近い値を予測結果として出していることがわかる．しかし，図中の対角線から大きく離れたものもあり，成績予測時に実際の成績に対して過大評価または過小評価していることがわかる．

評価指標の値を表 2 に示す．R^2 の値は，モデルが悪い場合に負の値をとり，1 に近づくほど予測精度が高いことを表す．また，予測と実際の誤差を見積もる RMSE と MAE は，最良の値 0 に近いほど予測精度が高いことを表すので実際の成績に近い値を予測できていることがわかる．偏差平方和の比は実際の偏差平方和を分母とし，値が 1 以下であるため，実際より予測のほうが散らばりが大きい．よって R^2，RMSE，MAE の値に影響を与えている．特に，RMSE は二乗誤差を用いるため，予測と実際の差が大きいものが存在するほどに大きな値になることから，予測した授業成績の中には実際の成績と大きく異なる値も存在していることがわかる．

4 おわりに

学習ログから定義した特徴量を用いて授業成績を予測し，予測結果に対する評価指標から予測に用いる特徴量を検証した．実験より，実際の成績に近い値を予測できているが，大きく異なる値を予測しているものも存在した．また，朝に学習している割合を特徴量に含む場合において，実際の成績と近い値を予測する傾向にあり，学習を行う時間に関する学習ログからの特徴量を予測に用いることは予測結果に対して良い影響を与えると考えられる．

今後の課題として，Random Forest 以外のアルゴリズムを用いた授業成績の予測における特徴量の検証，異なる方法によって定義した特徴量による予測結果の差異が挙げられる．

参考文献

[1] 緒方広明，殷成久，大井京，大久保文哉，島田敬，小島健太郎，山田政寛，大学におけるラーニングアナリティクスに基づく授業改善と教育革新，電子情報通信学会総合大会講演論文集，1 号，pp.92-93，2016/3.

[2] 松尾龍磨，伊藤恵，LMS を用いたプログラミング授業における機械学習による得点率予測，日本ソフトウェア科学会第 37 回大会講演論文集，2020/9.

[3] sklearn.ensemble.RandomForestRegressor, `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html`，2021/12/04.

[4] scikit-learn Machine Learning in Python，`https://scikit-learn.org/stable/`，2021/12/04.