

情報科目における成績予測の精度を向上させる特徴量の検討

駒谷優斗[†] 望月久稔[†]

[†]大阪教育大学

1 はじめに

授業の成績を予測できれば教員が成績不良の学生を早期発見し、指導することが可能となる [1]。本研究では成績予測の精度を向上させる特徴量の提案を目的とし、Moodle に蓄積された受講生のログデータから授業の成績予測に有益な特徴量を検証する [2]。情報科目の学習ログから複数の特徴量を抽出し、Random Forest [3] による授業の成績予測に用いる。各特徴量と組み合わせを予測結果における評価指標の値を用いて検証する。

2 成績予測に用いる特徴量の検討

受講生の学習ログから小テストの得点や受験回数など、授業の成績予測に用いる特徴量を抽出する。抽出した特徴量の数を N として $\sum_{r=1}^N n_r C_r$ 通りの組み合わせそれぞれに対してグリッドサーチ [4] を行い、最適な Random Forest のモデルを構築し回帰分析する。

Moodle に蓄積されているログデータから受講生に関する情報を抽出し、以下の 6 つの特徴量を定義する。

1. 小テストの受験回数の合計
2. 点数の上がり幅を加算した小テストの得点の合計
3. 小テストの平均点 (初回受験点数)
4. 小テストの平均点 (最高点数)
5. 小テストの所要時間の平均
6. 学習時間

学習時間は Moodle に蓄積された各学生のログデータから抽出する。

小テストに関する特徴量はオンデマンド型の授業で実施した小テストのデータから抽出する。小テストは何度でも受験可能であるため、学生によっては何度も受験し点数が上がる場合がある。そこで式 (1) により小テストの得点に点数の上がり幅を加算した特徴量を定義する。式 (1) の具体的な例を挙げる。

$$\sum \{(\text{平均点}) + (\text{最高点}) - (\text{最低点})\} \quad (1)$$

Examination of Features for Improving Accuracy of Grade Prediction in the Information Subjects

Yuto KOMATANI[†] and Hisatoshi MOCHIZUKI[†]

[†]Osaka Kyoiku University

例 点数の上がり幅を加算した小テストの得点の合計
受講生がある小テストにおいて最低点, 最高点, 平均点がそれぞれ 5 点, 15 点, 10 点であり, 別の小テストでは 7 点, 11 点, 8 点であったとする。この場合の特徴量の値は $(10+15-5)+(8+11-7) = 32$ となる。

各受講生の授業における評点を目的変数とする。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 6 割を学習用データ, 2 割をグリッドサーチ用データ, 残りの 2 割を検証用データに用いる。scikit-learn [5] にある GridSearchCV を用いて最適なパラメータで Random Forest のモデルを構築し, 検証用データを用いて授業成績を予測する。

3 授業の成績予測を用いた特徴量の評価

2020 年度から 2022 年度に実施した学部一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。予測結果の評価指標として, 実際の成績と予測した成績の決定係数 R^2 , 二乗平均平方根誤差 RMSE, 平均絶対誤差 MAE を用いる。

3.1 点数の上がり幅を加算した小テストの得点

2020 年度実施分に関して各特徴量単体で授業成績を予測すると, 点数の上がり幅を加算した小テストの得点を特徴量に用いた場合が最もよかった。2021 年度実施分に関して同様に予測すると, 小テストの平均点 (初回受験点数) を特徴量に用いた場合が最もよく, 点数の上がり幅を加算した小テストの得点を特徴量に用いた場合は 2 番目によかった。また, 2022 年度実施分に関して同様に予測すると, 小テストの平均点 (最高点数) を特徴量に用いた場合が最もよく, 点数の上がり幅を加算した小テストの得点を特徴量に用いた場合は 3 番目によかった。

年度によって差がでた理由として, 点数の上がり幅が 0 の受講生の割合が異なることが挙げられる。最も精度のよかった 2020 年度では, 点数の上がり幅が 0 の学生の割合が約 36.4% であったのに対し, 他の特徴量と比べて精度の悪かった 2021 年度と 2022 年度それぞ

れの点数の上がり幅が0の学生の割合は約40.6%、と約45.5%であり、2020年度と比べて多かった。

各年度において点数の上がり幅を加算した小テストの得点のみを特徴量に用いた場合の評価指標の値を表1に示す。

R^2 の値は1に近いほど予測精度が高いことを表し、RMSEとMAEは0に近いほど予測した成績と実際の成績の誤差が小さいことを表すため、実際の成績に近い値を予測できていることが分かる。小テストの得点に点数の上がり幅を加算することで、他の特徴量と比べて精度のよい特徴を定義できた。よくなった理由として、小テストの得点のみに比べて成績との相関が強くなったことが挙げられる。例えば2020年度において、小テストの平均点(初回受験点数)と実際の成績との相関係数は約0.70であったのに対し、点数の上がり幅を加算した小テストの得点と実際の成績との相関係数は約0.79であった。

表 1: 各年度における特徴量単体での評価指標

年度	R^2	RMSE	MAE
2020	0.670467624	6.006216895	5.101251002
2021	0.778870531	4.097549676	3.332095069
2022	0.878662289	7.364769439	5.719836845

3.2 特徴量の組み合わせ

2020年度、2021年度、2022年度それぞれにおいて成績を予測した際の最も精度のよかった組み合わせを2章に示した特徴量の番号を用いて表2に示す。

表 2: 各年度における特徴量の組み合わせ

年度	最も精度の高い組み合わせ
2020	1, 3, 6
2021	1, 2
2022	2, 4, 5

2021年度と2022年度における授業の成績予測で最もよ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。また、各年度それぞれにおける評価指標を表3に示す。

表 3: 各年度における評価指標

年度	R^2	RMSE	MAE
2020	0.752467861	5.205563689	4.546159077
2021	0.826596600	3.628521667	2.983208791
2022	0.948783586	4.784822570	4.115788771

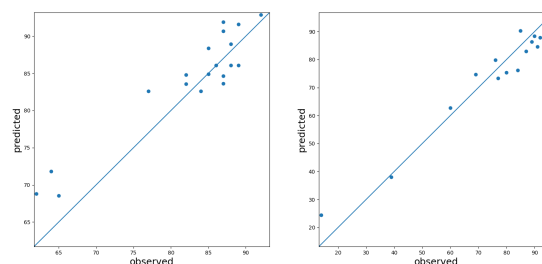


図 1: 2021 予測-実際 図 2: 2022 予測-実際

各評価指標の値より、全ての年度で実際の成績に近い値を予測できていることが分かる。しかし、図1では図中の対角線より上に位置するものが多く、実際の成績に対して過大評価しているものが多かった。また、2021年度と2022年度の最もよい組み合わせに2の特徴量が含まれているが、単体での精度がよかった2020年度には含まれていない。含まれていない理由として、2020年度に相性のよい他の特徴量がないことが挙げられる。2の特徴量と組み合わせると精度がよくなる特徴量は小テストの平均点(初回)のみであった。

4 おわりに

Moodleに蓄積された受講生のログデータから授業の成績予測に有益な特徴量を検証した。実験より、実際の成績に近い値を予測でき、点数の上がり幅を加算した小テストの得点の合計は授業の成績予測に有益であることがわかった。

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、成績予測を用いたフィードバックが挙げられる。

参考文献

- [1] 山田政寛: ラーニング・アナリティクス研究の現状と今後の方向性, 日本教育工学会論文誌, Vol.41, No.3, pp.189-197(2018).
- [2] 西岡克記: 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, 情報処理学会第84回全国大会講演論文集, Vol.1, pp601-602(2022).
- [3] sklearn.ensemble.RandomForestRegressor, <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>, 2022/12/07.
- [4] sklearn.model_selection.GridSearchCV, https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html, 2022/12/07.
- [5] scikit-learn Machine Learning in Python, <https://scikit-learn.org/stable/>, 2022/12/07.