N-023

学生の成績予測における精度向上のための特徴量とデータ加工手法の検証

Examination of Features and Data Processing Methods to Improve Accuracy in Predicting Student Grades

髙橋 龍人[1）]　駒谷 優斗[1）]　望月 久稔[1）]
Ryuto TAKAHASHI　Yuto KOMATANI　Hisatoshi MOCHIZUKI

1　はじめに

教育の情報化が伸展し，蓄積された大量のログデータの分析に基づいてより効果的な教育や学習の実現を目的としたラーニングアナリティクスの研究が盛んになった [1]．その中の成績予測では，成績不振の学生を早期発見することで，対処が可能になる．高い精度で成績を予測できれば，より適切に対処できる．しかし，成績予測に用いるデータは，授業によって偏りやデータ数の不足が生じ，そのまま成績予測に用いても高い精度にならないことがある．成績予測の精度向上を目的とし，データ加工によりデータの偏りを軽減し，データ数を増やすことで精度が向上すると考え，成績予測に有益なデータ加工手法を検証する．また，成績予測に有益な特徴量も検証する．

2　授業成績の予測に用いる特徴量とデータ加工手法

成績予測に用いる特徴量とデータの加工手法を提案し，予測の流れを述べる．

2.1　特徴量の定義

受講生の学習ログから特徴量を定義する．先行研究 [2] で使用された小テストの平均点や学習時間，点数の上がり幅を加算した小テストの得点の合計などの既存の特徴量に加えて，提出率を新たな特徴量として定義する．提出率は Moodle に蓄積された授業のログデータから抽出する．成績を決定する際に判断材料に用いる Moodle 上の小テストや課題などのコンテンツのうち，受講者が解答や提出した割合を提出率と定義する．

2.2　データの加工

授業成績のデータは授業によって高成績の受講者が多いなどの偏りが生じた不均衡データになる．また，データ数は，受講者の人数分しか収集できないため，機械学習に用いるには不足する可能性がある．よって，精度向上のために以下の 4 つのデータ加工手法により，データの偏りを軽減し，データ数を増やす．

1. SMOGN[3] によるデータ加工
2. 一般の正規分布に従う乱数によるデータ生成
3. 手法 1 後に手法 2
4. 手法 2 後に手法 1

手法 1 で用いる SMOGN は，不均衡データの対処に用いる多数派のデータを減少させるアンダーサンプリングと少数派のデータを増加させるオーバーサンプリングを組み合わせ，回帰データに対応した手法 [3] である．成績の値を基準に用いると，成績データの存在する割合が小さい範囲ではデータが増加，大きい範囲ではデータが減少し，データの偏りが軽減する．加工の強さをハイパーパラメータとして設定でき，本稿では異なる加工の強さで 3 通り使用する．手法 2 は，各特徴量の平均値と標準偏差から一般の正規分布に従う乱数を生成し，データ数を増やす．データを生成する際に使用する乱数は，データごとに一つの値を用いるため，例えば成績の低いデータを生成したときは，他の特徴量の値も小さくなる．本稿では，データの増加数を 20 から 300 で 10 通り用意する．手法 3,4 は，手法 1,2 を組み合わせることで，データの偏りが軽減し，データ数が増加する．手法 1,2 によりパラメータの組み合わせは $3 * 10 = 30$ 通りある．

2.3　授業成績の予測

各受講生の授業における評定を目的変数として成績予測する．全体図を図 1 に示す．Moodle のログデータから 2.1 節で定義した提出率を抽出し，既存の特徴量からなるデータセットに加える．作成したデータセットを学習用データ，グリッドサーチ用データ，検証用データに分割する．分割した学習データを 2.2 節で述べた手法を用いて加工し，新たな学習用データを作成する．グリッドサーチ用データに対して，最適なパラメータで学習用データからランダムフォレスト（以下，RF）または勾配ブースティング（以下，GB）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し，評価する．

3　特徴量とデータ加工による予測精度への影響の評価

実験データには 2021 年度から 2022 年度に実施した学部 1 年生対象の情報リテラシーに関する授業の Moodle のログと，先行研究 [2] によりログから抽出したデータを用い，2 章で述べた特徴量とデータ加工による予測精度への影響を評価する．

提出率とデータ加工による予測精度を評価するため，特徴量の組み合わせごとに成績データの決定係数 R^2 を比較し，全ての組み合わせのうち，R^2 が向上した組み合わせの割合で評価する．R^2 が向上した組み合わせの割合が 1 に近いほど，提出率導入またはデータ加工したことで，多くの組み合わせで R^2 が向上したことを表す．反対に R^2 が低下した組み合わせの割合が 1 に近いほど，多くの組み合わせで R^2 が下がったことを表す．

3.1　提出率のみによる成績予測

提出率のみを用いて成績を予測し，各年度とモデルの予測結果を R^2 で評価する．

2022 年度は R^2 が 0.673 と 0.676 で実際の成績に近い値を提出率のみで予測できたが，2021 年度は-0.081 と 0.243 で提出率のみでは予測できなかった．この理由として，2021 年度の提出率が，学生ごとで値に大きな差がないことが挙げられる．提出率の標準偏差は，2021 年度が 0.125，2022 年度が 0.172 で，2021 年度の提出率の

1）大阪教育大学　Osaka Kyoiku University

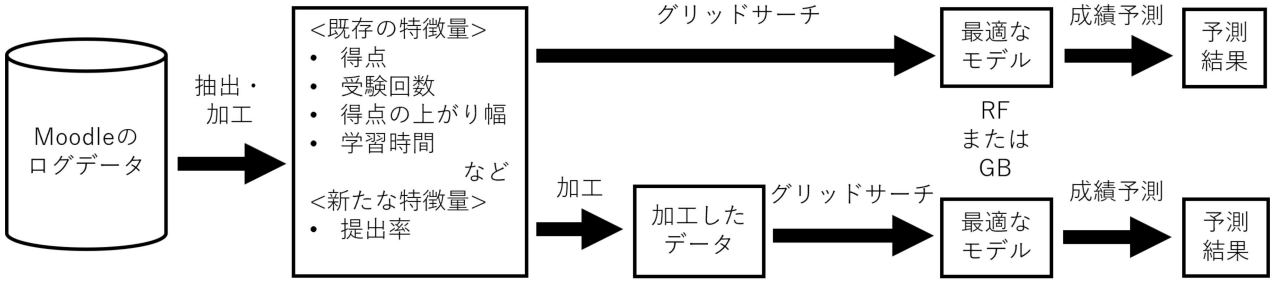

図 1　実験の全体図

表 1　提出率の追加による精度の変化の割合

年度	モデル	向上	変化なし	低下
2021	RF	0.19	0	0.81
	GB	0.51	0	0.49
2022	RF	0.65	0	0.35
	GB	0.86	0	0.14

表 2　各データ加工手法による精度が向上した割合

年度	モデル	手法 1	手法 2	手法 3	手法 4
2021	RF	0.82	0.98	0.87	1
	GB	0.29	0.96	0.91	1
2022	RF	0.64	0.36	0.87	0.87
	GB	0.43	0.81	0.77	0.86

方が散らばりがないとわかる．よって，2021 年度の提出率は学生ごとで値に大きな差がなく，似た成績を予測するため，精度が低いと考える．

3.2　提出率の有無による予測精度の評価

既存の特徴量の組み合わせに提出率を加えた場合の予測精度の変化を評価する．各年度における提出率の追加により精度が向上，変化なし，低下した特徴量の組み合わせの割合を表 1 に示す．

表 1 より，2021 年度よりも 2022 年度の方が精度が向上した組み合わせが多い．この理由として，2022 年度の提出率のみによる予測精度が高いことが挙げられる．2021 年度は提出率のみによる精度は低いが，GB では精度が向上した組み合わせが多く，提出率のみによる精度が低ければ組み合わせたときに精度が低下するものが多くなるとは限らないことがわかった．提出率の導入によって提出率のみによる予測精度が高い年度は，より多くの特徴量の組み合わせで予測精度が向上した．

3.3　データ加工による予測精度の評価

各データ加工手法を評価する．各年度における各データ加工手法による精度向上した割合を表 2 に示す．

表 2 より手法 1 は，両年度において RF を用いたときは精度向上の割合が多く，GB を用いたときは少ない．GB を用いたときに精度が低下した組み合わせが多い理由として，加工によるデータ数の減少が挙げられる．加工によるデータ数の変化は，加工の強さの 3 通りのうち 2021 年度では 2 通りで減少し，2022 年度では全て減少した．RF でも同じデータを用いるため加工によってデータ数が減少したが，精度が向上した組み合わせが多い．これは，RF がデータ数の減少による精度への影響が少ないためであると考える．

表 2 より手法 2,3,4 は，手法 2 の 2022 年度の RF を用いたときを除き，精度が向上した組み合わせが多い．この理由として，データの不均衡さの軽減が挙げられる．加工前の尖度は 2021，2022 年度それぞれ 26.87，9.71 で，手法 2 では，2021 年度は 5 から 24，2022 年度は 1 から 9 減少し，手法 3,4 では 2021 年度は 10 から 27，2022 年度は 5 から 11 減少した．よって，手法 2,3,4 によってデータの不均衡さが軽減し，精度が向上したと考える．

尖度の変化が似た手法 3,4 で，手法 4 がより多くの組み合わせで精度が向上した理由として，元のデータの特徴をある程度保ったことが挙げられる．手法 3 では，データ加工のパラメータの組み合わせ 30 通りのうち，平均値は 2021 年度は 18 通りで 0 から 12，2022 年度は 25 通りで 0 から 18 減少し，標準偏差は 2021 年度は 20 通り，2022 年度は 19 通りで 6 から 9 増加し，データの広がりが大きくなった．手法 4 ではデータ加工のパラメータの組み合わせ 30 通りのうち，平均値は 2021 年度は 24 通り，2022 年度は 23 通りで 0 から 9 増加し，標準偏差は 2021 年度は 21 通り，2022 年度は 17 通りで 3 から 6 増加し，データの広がりが大きくなったが，変化は手法 3 より小さく元のデータの特徴をある程度保ったと考える．手法 4 によって，元のデータの特徴をある程度保ちつつデータの不均衡さが軽減し，より多くの特徴量の組み合わせで予測精度が向上するとわかった．

4　おわりに

受講生の学習ログから抽出した提出率を既存の特徴量に加えた後に加工し，成績予測に用いた．その後，提出率の有無とデータ加工による予測精度の変化を評価した．実験より，提出率は成績予測に有益だとわかった．一般の正規分布に従う乱数によるデータ生成後，SMOGN による加工によって，より多くの特徴量の組み合わせで予測精度が向上した．今後の課題として，生成する値による精度の変化や，異なる成績分布での予測精度の検証などが挙げられる．

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（オンライン），DOI：10.15077/jjet.42024 (2018).

[2] 駒谷優斗，望月久稔：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，第 85 回全国大会講演論文集，Vol. 2023, No. 4, pp.929-930 (2023).

[3] Branco, P., Torgo, L., and Ribeiro, P.R.:SMOGN: a Pre-processing Approach for Imbalanced Regression, Proc. Machine Learning Research, pp.36-50 (2017).