Vol.2019-CE-152 No.2
Vol.2019-CLE-29 No.2
2019/11/15

情報基礎科目における毎週の出席およびテスト結果を用いた履修初期での合否予測手法の比較検討

森 美樹子[1,a)]　久保田 真一郎[2,b)]　杉谷 賢一[2,c)]　中野 裕司[2,d)]

概要：週ごとに学習活動の成果となる出席や確認テストのデータを使い，授業を不合格になりそうな学生を早期に見つけ出すことができれば，早期に学習指導を行うことで，不合格者を減らすことができると考えられる．本報告では，全学的に行われる情報基礎科目を対象に，週ごとの出席状況と確認テストの結果を入力値として，データを整理した上で linearSVC による不合格者の予測とその精度を確認した．総合評価点と確認テストが関係することから，総合評価点を目的変数とし，出席および確認テストを説明変数とする重回帰分析を行い，LinearSVC の結果と比較した．比較した結果，LinearSVC による予測の正答率は早い授業回で，重回帰分析の結果より微小に高い精度となっていた．また，不可となる人を不可と判定できた割合を示す再現率は，早い授業回から重回帰分析が高い精度を示していた．さらに検討を進め，精度の向上を目指す予定である．

Review on Comparing between Prediction Methods of At-risk Students to Use Weekly Data of Attendance and Quiz in the First-year University-wide Information Basic Course

1. はじめに

現在，学力や学習意欲などあらゆる面で学生が多様化している反面，学士としての質保証も求められている．そのため，データをもとに教学改善を支援する機能として，教学 IR の必要性が認識されつつある．教学 IR の役割の一つとして，学生の個に応じた学習支援を行うために，学士課程における学修状況の把握，分析がある．その中には，ドロップアウトする，あるいは高い成績を上げる学生のパターンを機械学習により発見する研究などが行われている．

例えば，学士課程におけるデータから学習状態を予測する研究がある [1]．雨森らは，入学前後から 1 年次の修学状態のデータから，3 年次前期までの単位取得状況を推定する決定木分析を行っている．Djulovic ら [2] は，決定木やニューラルネットワークなどの機械学習手法により，初年次の学生データから 1 年後の在学状況の予測を行った．また，大規模な学習ライフログを用いた学修モデル化によって学修支援への活用を考えた研究 [3] がある．近藤らは，入学前に得られる入試種別や入学前課題提出度，1 年次春学期中に得られる出席率や GPA といったデータを用いて，3 年次における在籍状況の予測を行なっている．複数の学習器を用いて分類を行い，その精度を比較した．結果として，ほとんどの学習器で再現率 50~60%程度が予測可能であり，春学期第 5 週の終了時点では最も良い学習器で再現率 40%が予測可能であった．以前行った研究 [4] では，LinearSVC を用いて，学生の合否予測について検討した．結果として，不合格になりそうな学生を全 15 回中 4 回から 5 回の段階で再現率約 70%程度で発見することができた．しかし，この研究で使用したデータは合格と不可の割合が大きく異なっていたため，うまく分析することができていなかった．

本研究では，履修途中の早い時期に高い精度の予測結果を得ることで，不合格となる学生を減らすのに役立てると

[1] 熊本大学大学院自然科学教育部
Kumamoto University Graduate School of Science and Technology
[2] 熊本大学総合情報統括センター
Kumamoto University Center for Management of Information Technologies
[a)] mkk@st.cs.kumamoto-u.ac.jp
[b)] kubota@cc.kumamoto-u.ac.jp
[c)] sugitani@cc.kumamoto-u.ac.jp
[d)] nakano@cc.kumamoto-u.ac.jp

Vol.2019-CE-152 No.2
Vol.2019-CLE-29 No.2
2019/11/15

表 1　受講者の構成と合否

	新入生	再履修生	合計
合格	1627	76	1703
不可	92	44	136
合計	1719	120	1839

いう目標のもと，LinearSVC と重回帰分析による精度比較を行った．

2. 使用データ

本研究では，2017 年度情報基礎科目のデータを使用した．本科目は，全学部生 1 年次の必修科目の一つである．この科目は，情報セキュリティ・法律・情報倫理を学んだり，Web サイト作成能力を身につけることを目的としており，授業内に提出する課題，授業後に行う確認テスト，授業後半で提出する作品課題によって理解度の確認を行う．

評定は毎週の課題 (30%) および確認テスト (30%) と，学期最後に提出する作品課題 (40%) によって計算される．また，毎回の授業で出席登録と課題の提出が課されており，出席登録と課題の提出を共に行った「出席回数」が全講義回数の 2/3(10 回) であり，かつ確認テストも同回以上であることが単位取得の必要条件である．出席登録は授業開始から 20 分の間のみ可能である．課題は授業終了の 30 分前から提出が可能になり，授業終了まで受け付けられる．確認テストは授業終了後 2，3 週間の期間が設けられており，期間内の最高点がその回の最終的な点数として扱われる．

学生が不合格になる原因は，成績が 60 点未満というのは殆どなく，出席が 2/3 未満であること，確認テストの合格が 2/3 であること，作品課題が未提出等であることが多い．作品課題は配点が大きいが，学期末に提出するものであり，履修途中で合否予測をするという目的に反するため今回の分析には使用しない．

今回の分析では，情報基礎科目から得られる課題提出状況，確認テスト得点の 2 種類を入力パラメータとして扱う．

次に，今回分析に使用した受講者の構成と合否に関するデータを**表 1** に示す．

表 1 より，新入生と比較すると再履修生の不合格の割合が高いことがわかる．データの性格がかなり異なることから，新入生と再履修生を分けて扱うこととし，本報告では，新入生に関してのみ報告する．再履修生に関しても新入生と同様の分析を行ったが，後述する新入生に対するような結果は得られておらず，原因も十分把握できていないため，今後の課題とする．

また，表 1 を見ると，新入生の中でも合格と不可の人数比が大きく異なることが分かる．そのため，今回の分析では合格と不可の割合が 1:1 になるようにデータ数を調整した．また，情報量の多いデータを扱うために，不可の学生のうち出席回数が 0 の学生は除外した．合格者のデータは，確認テストの分散が多い学生を選出した．予測に使用されるデータ数は合格，不可共に 88 であり，合計 176 のデータ数で分析を行う．

3. 分析手法

本研究では，機械学習を利用するにあたり，Python の機械学習ライブラリとしてオープンソースで公開されている scikit-learn[5] を使用する．学習器については，scikit-learn が提供する，チートシート [6] という学習器決定のためのフローチャートがあり，そのフローチャートに則り学習器を決定した LinearSVC と，精度比較のために重回帰分析を使用する．**図 1** に [6] より scikit-learn のチートシートを引用する．

START から始まり，データ数や分析方法などを元に，学習器を決定する．今回の分析では，ノード「START」から進み，データ数が 50 より多いためノード「category」に進む．分類を目的とするためノード「labeled data」に進む．使用データはラベル付けされたものであるのでノード「<100K」に進む．データ数が 10 万より少ないためチートシートから学習器として LinearSVC が先ずは選定される．LinearSVC とは，サポートベクトルマシン (SVM) を用いて，線形な識別表面を作成するものである [6]．SVM とは，教師あり学習を用いるパターン認識モデルの一つであり，分類や回帰といった分析に適応できるものである．訓練データから，判別する境界とデータとの距離 (マージン) が最大になるような，マージン最大化という考えを使って，分類を行うことができる．境界の近くにあるデータはサポートベクトルと呼ばれ，SVM ではサポートベクトルのみを用いて分類が行われる．

4. 入力データの前処理

学習器に入力するデータは，課題提出状況，確認テスト得点の 2 種類である．課題提出状況は，各授業回に提出があった場合を 1，なかった場合を 0 とした．たとえば，ある学習者が第 1 回から第 5 回までのうち，第 4 回以外のすべての回の課題を提出した場合，課題提出状況を示す要素は $(1, 1, 1, 0, 1)$ と表現される．確認テスト得点は，0~100 で表される得点を 100 で割り，0~1 の値として扱う．課題提出状況，確認テスト得点のどちらも授業回の数だけ要素を持つため，第 $i(1 \leq i \leq 15)$ 回授業までの結果を入力データとして扱う場合，入力データの要素数は $2i$ 個となる．第 i 回の授業での課題提出を $A_i(1 \leq i \leq 15)$，第 i 回の授業での確認テスト得点を $K_i(1 \leq i \leq 15)$ と表すと学習器に入力する特徴ベクトルは

$$(A_1, \cdots, A_i, K_1, \cdots, K_i),$$

と表される．

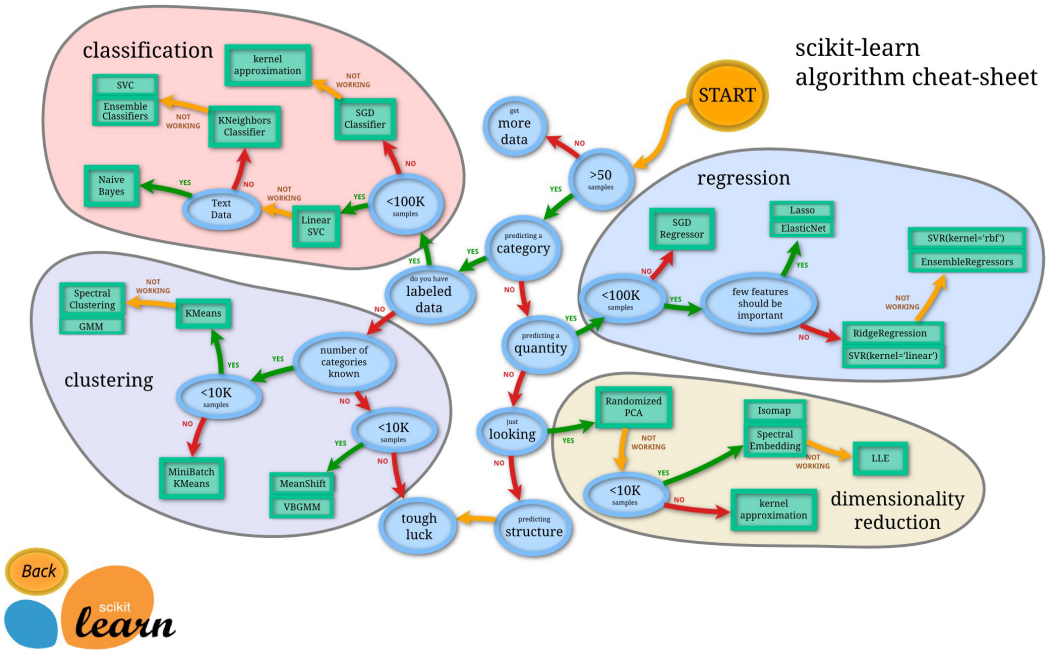

図 1　scikit-learn のチートシート ([5] より引用)

5. 分析結果

今回の分析で出力されるデータは，予測結果から計算された正答率，適合率，再現率の 3 つである．正答率 (accuracy) は予測に対し，答えがどれだけあっていたかを示す．適合率 (precision) は，正しいと予測したもののうち本当に正しいものの割合を示す．今回の分析においては，不可と予測した人のうち実際に不可だった人の割合である．再現率 (recall) は，見つけるべきもののうち正しく見つけることのできたものの割合を示す．今回の分析においては，実際に不可である人のうち不可と予測することができた人の割合である．

正答率を算出するために，本研究では K 分割交差検証を行った．今回はデータを 10 分割し，1 つをテストデータ，9 つを訓練データとして扱う．これを 10 回行い，結果を平均することで正答率が算出される．

今回の分析では，より多くの不合格となりそうな学生を早期に予測することを目的としている．適合率が高い場合，不可と予測した人数が少なくとも，予測が合っていれば高い割合となる．しかしこれでは，合格と予測した学生の中に実際は不可の学生がいるというケースが発生してしまう．再現率が高い場合，実際は合格の学生を不可と予測してしまうケースが含まれる可能性があるが，実際に不可である学生をより多く予測できている事になる．そのため，本研究では主に再現率に注目する．

5.1 LinearSVC の場合

まずを LinearSVC を用いて分析を行なった．入力値は，

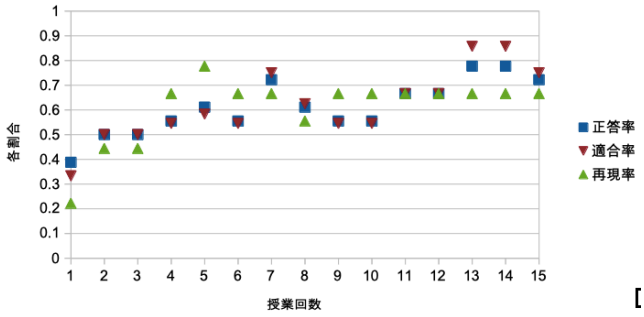

図 2　LinearSVC による予測精度

第 1 回 ～ 第 X 回までの各パラメータである．例えば第 3 回までの課題提出状況 (A_1, A_2, A_3) と確認テスト得点 (K_1, K_2, K_3) を入力とした場合，入力する特徴ベクトルは $(A_1, A_2, A_3, K_1, K_2, K_3)$ となる．分析結果を図 **2** に示す．

図 2 より，全体的に右肩上がりのグラフになっている．第 5 回の段階で，約 8 割の不可となる学生を予測できている．

5.2 重回帰分析の場合

次に，重回帰分析を用いて分析を行なった．重回帰分析は，統計的な予測手法として用いられており，計算しやすく誤差の少ない予測式を作成できる．そのため，比較手法として重回帰分析を用いることとした．重回帰分析で扱うデータは，LinearSVC で利用したデータと同様に各学生の課題提出状況，確認テスト得点の 2 種類であり，合格者 88，不可 88 の計 176 のデータで分析を行った．更に，K 分割交差検証により精度を確認している．説明変数は課題提出状況 $(A_1, \cdots, A_i)$，確認テスト得点 $(K_1, \cdots, K_i)$ であり，説明変

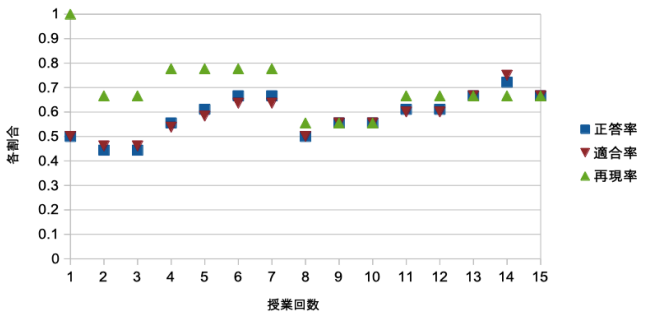

図 3 重回帰分析の予測精度

表 2 第 6 回の回帰係数

説明変数	係数
第 3 回課題提出	-0.149
第 1 回確認テスト得点	0.257
第 6 回課題提出	1.193
第 5 回確認テスト得点	1.732
第 2 回課題提出	4.719
第 3 回確認テスト得点	5.697
第 1 回課題提出	6.402
第 6 回確認テスト得点	7.396
第 5 回課題提出	7.829
第 2 回確認テスト得点	8.075
第 4 回課題提出	10.065
第 4 回確認テスト得点	13.803

数は成績 (0∼100) である．例えば第 3 回までの課題提出状況 (A_1, A_2, A_3) と確認テスト得点 (K_1, K_2, K_3) を入力とした場合，入力する特徴ベクトルは $(A_1, A_2, A_3, K_1, K_2, K_3)$ となる．0∼100 で算出される分析結果より，実際の評定と同じ様に 60 点以上を合格，60 点未満を不可として各出力を算出した．分析結果を**図 3** に示す．

図 3 より，LinearSVC 同様右肩上がりのグラフになっているが，第 8 回で大幅に値が低下していることがわかる．第 4 回の段階で，約 8 割の不可となる学生を予測できている．

また，図 3 より最も正答率の高い第 6 回における回帰係数を**表 ??**に示す．

表 **??**より，第 4 回のデータが分析に大きく影響していることが分かる．

5.3 分析結果考察

LinearSVC を利用した場合，最も良い部分で 7.7 割の不可となりそうな学生を予測できている．正答率を見ると，LinearSVC は第 6 回で落ち込むものの，第 7 回では 7 割を超えており，重回帰分析と比べて良い結果が得られている．また，5 章で述べたように，本研究では再現率に注目している．再現率を見ると，LinearSVC より重回帰分析の方がより早い段階で 6 割を超えている．重回帰分析の第 1 回における再現率が 100%になっているが，各授業回数における第 1 回の回帰係数を比較したところほとんどが低い値となっていること，正答率が 5 割であることから，正しく予測できているとは言えない．より早期に高い精度での分析を行う為にさらなる検討が必要である．

6. おわりに

本稿では，LinearSVC と重回帰分析を用いて，授業で得られる課題提出状況，確認テスト得点から学生の合否予測を行うことについて検討を行った．分析結果から，最も良い部分で再現率約 8 割程度，不可となりそうな学生を予測することができている．2 つの予測手法を比較すると似たような結果であり，本研究において適した学習器を定めることができなかった．そこで，新しい種類の入力パラメータを利用したり，別の学習器を使うことについて検討を行う必要がある．現在は，時系列データの予測に長けた RNN(Recurrent Neural Network) を用いた分析を進めている．RNN は入力値に対応する出力値が隠れ層を介して得ることができ，この隠れ層の情報がさらに次の入力に使用される．したがって，過去の状況を考慮した分析ができるという利点がある．RNN を用いることで，より良い精度の分析ができるのではないかと考えている．また，今回再履修生を除いた分析を行ったが，再履修生も含めた検討も必要である．

参考文献

[1] 教学 IR の–一方路島根大学の事例を用い–, 雨森聡, 松田岳士, 森 朋子, 京都大学高等教育研究, 第 18 号,pp.1–10 (2012)
[2] Towards Freshman Retention Prediction: A Comparative,Djulovic, A. and Li, D, International Journal of Information and Education Technology, Vol. 3, No. 5, pp.494–500(2013)
[3] 学士課程における大規模データに基づく学修状態のモデル化, 近藤 伸彦, 畠中 利治, 教育システム情報学会誌 vol.33, no.2, pp.94–103 (2016–05)
[4] 初年時全額情報基礎科目における学習データを用いた LinearSVC による履修途中での合否予測の検討, 森 美樹子, 久保田 真一郎, 杉谷 賢一, 中野 裕司, 研究報告教育学習支援情報システム（CLE）,2019-CLE-27(12),1-5 (2019-03-13)
[5] scikit-learn Machine Learning in Python, 入手先 ⟨https://scikit-learn.org⟩ (2019 年 10 月確認)
[6] Choosing the right estimator — scikit-learn 0.20.2 documentation, 入手先 ⟨https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html⟩ (2019 年 10 月確認)
[7] sklearn.svm.LinearSVC — scikit-learn 0.20.2 documentation, 入手先 ⟨https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html⟩ (2019 年 010 月確認)
[8] sklearn.model_selection.train_test_split — scikit-learn 0.20.2 documentation, 入手先 ⟨https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html⟩ (2019 年 10 月確認)