Vol.2019-CLE-27 No.12
2019/3/21

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

概要：合格水準に達せず不合格になる学生をできるだけ少なくすることを目的とし，全学情報基礎科目における学習データから受講している学生の合否を早期に予測することについて検討する．新入生と再履修生で，学生の合否の割合が大きく異なるため，今回は新入生のみを対象に検討を行った．授業各回の出席，課題の提出，確認テスト得点をパラメータとしたLinearSVCにより合否の予測を行った．結果，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．入力データの前処理，パラメータの追加などにより，より早期に予測精度を向上させることを検討する．

Investigation of pass / fail prediction by using LinearSVC for learning activities 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．

それに対して本研究では，単一の科目の範囲内ではあるが，入学生ほぼ全員が受講する科目の学習データを詳細に扱うことで，個々の学生の合否の予測を行い，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ことを目標とした．

2. 使用データ

本研究では，2017年度全学必修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．

本科目は毎週の課題(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%)によって評定が計算される．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

[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

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

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．

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．

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．

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[4] 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 [5] 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．**図 1** に [5] 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，出席状況，確認テスト得点の 3 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．出席状況は，課題提出と出席処理の両方が行われた場合を 1，それ以外を 0 とした．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，出席状況，確認テスト得点のいずれ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 $3i$ 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 回の授業での出席を $S_i(1 \leq i \leq 15)$，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

$$(A_1, \cdots, A_i, S_1, \cdots, S_i, K_1, \cdots, K_i),$$

と表される．

表 1 で示したように，新入生の中でも合格と不可の人数比が大きく異なる．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．予測に使用されるデータ数は合格，不可共に 92 であり，合計 184 のデータ数で分析を行う．合格者のデータは，全合格者からランダムに 92 個のデータを取り出す処理を行った [7].

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．

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

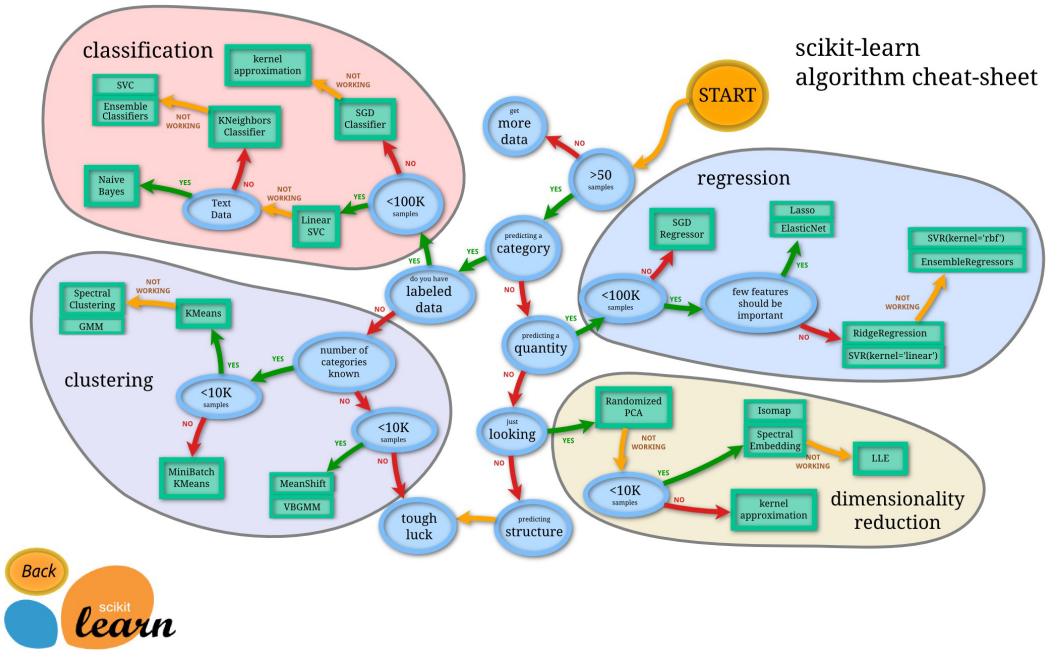

図 1　scikit-learn のチートシート ([2] より引用)

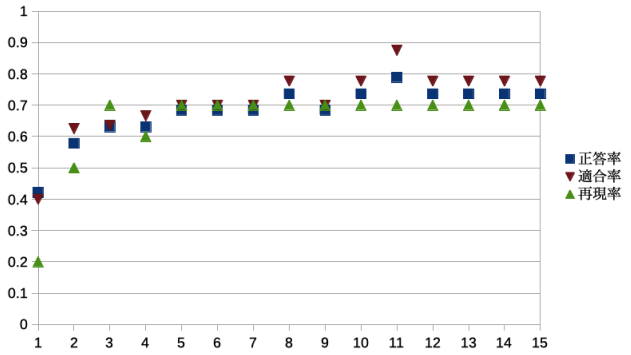

図 2　出席状況と合否の予測

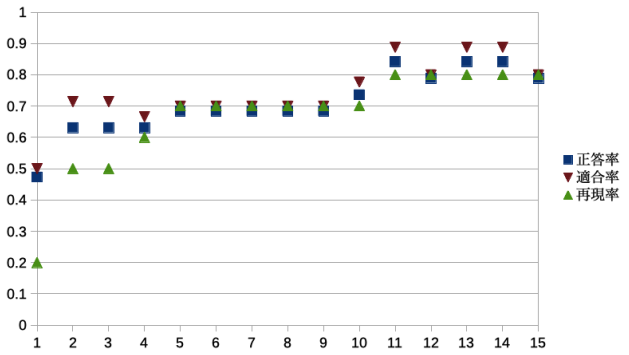

図 3　課題提出状況と合否の予測

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　入力パラメータが 1 種類の場合

まず，出席状況，課題提出状況，確認テスト得点の 3 つのパラメータのうち，1 種類のみを入力として予測を行なった．入力値は，第 1 回 ~ 第 X 回までの各パラメータである．入力が出席状況の場合の正答率，適合率，再現率を図 **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．再現率は第 5 回以降 0.7 で安定している．適合率に関しても，0.7~0.8 程度で安定している．

次に，入力パラメータが課題提出状況の場合の正答率，適合率，再現率を図 **3** に示す．

図 3 より，図 2 と同様に，右肩上がりのグラフになっている．適合率と再現率の差が少なく，第 5 回までのデータから約 7 割の不可となりそうな学生を予測できている．

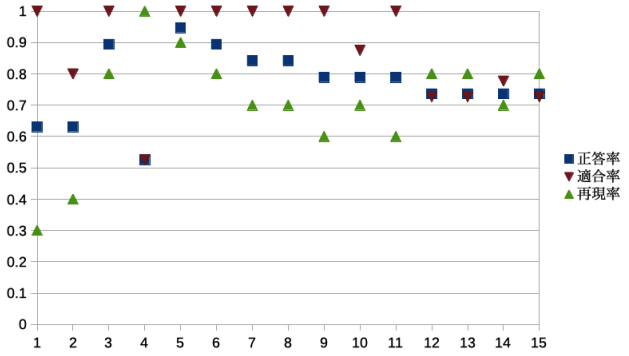

図 4　確認テスト得点と合否の予測

次に，入力パラメータが確認テスト得点の場合の正答率，適合率，再現率を図 **4** に示す．

図 4 より，第 11 回までは適合率と再現率ともに入力データが増えても，その確率は単一に上昇せず上下している．第 12 回以降では適合率と再現率の差が少なくなっている．第 4 回に時点で，不可となる学生を 100%予測することが

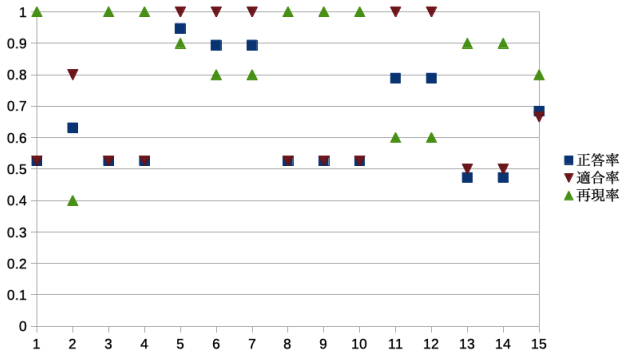

図 5　確認テスト得点，課題提出状況と合否の予測

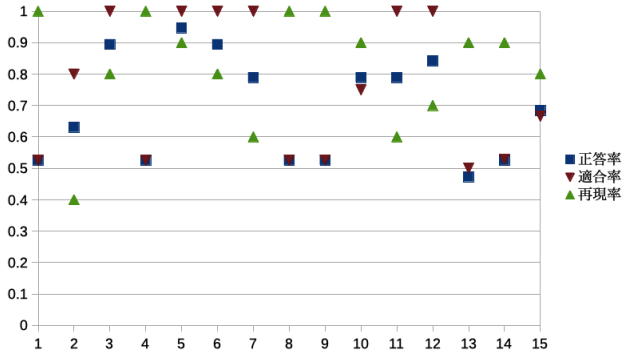

図 6　確認テスト得点，出席状況と合否の予測

できているが，同時に実際は合格である学生を不可と予測してしまっている．

5.2　入力パラメータが 2 種類の場合

次に，出席状況，課題提出状況，確認テスト得点の 3 つのパラメータのうち，2 種類を入力として予測を行なった．入力値は，第 1 回 ～ 第 X 回までの各パラメータである．例えば第 3 回までの出席状況 (A_1, A_2, A_3) と課題提出状況 (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．入力が課題提出状況，確認テスト得点の場合の正答率，適合率，再現率を**図 5** に示す．

図 5 より，通常，授業回が増えると入力データが増え，正答率などの割合は単一に上昇すると考えられるが，いずれの割合も授業回が増えても上下しており，うまく予測できていないことがわかる．第 5 回には，適合率，再現率が近い値で高い割合を示している．

次に，入力パラメータが出席状況，確認テスト得点の場合の正答率，適合率，再現率を**図 6** に示す．

図 6 より，授業回が増えても適合率，再現率ともに上下に変化し，第 5 回のみそれぞれ近い値で高い割合を示している．

次に，入力パラメータが出席状況，課題提出状況の場合の正答率，適合率，再現率を**図 7** に示す

図 7 より，全体的に右肩上がりのグラフになっている．

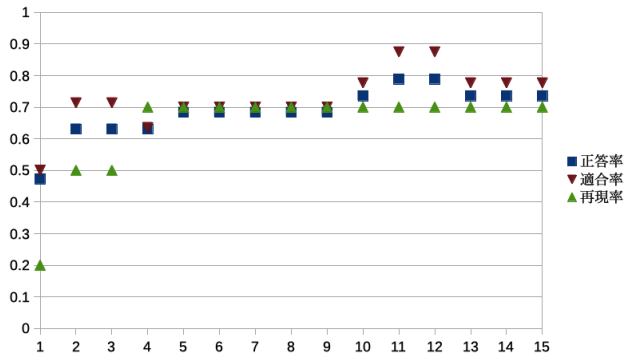

図 7　出席状況，課題提出状況と合否の予測

再現率は第 4 回以降 0.7 で安定している．適合率に関しても，0.7~0.9 程度で安定している．

5.3　分析結果まとめ

入力パラメータが 1 種類の場合，出席状況と課題提出状況が入力の時において，最も良い部分で約 7 割の不可となりそうな学生を予測できている．確認テスト得点が入力の時は，最終的には約 8 割の不可となりそうな学生を予測できているが，早期の段階では適合率と再現率共に値が不安定である．出席状況と課題提出状況の結果が似ているのは，この 2 種類のデータが似たデータであるからだと考えられる．入力パラメータが 2 種類の場合，確認テストが含まれるものは授業会が増えても適合率，再現率ともに値が上下に変化している．確認テストの結果を入力に含む場合，いずれも第 5 回において適合率，再現率が近い値で高い割合となっており，第 5 回までのデータによる予測の可能性があると考えられる．入力パラメータを増やした時，通常であれば分析結果の安定，適合率や再現率の値が向上すると考えられるが，今回の分析ではそのような結果が得られなかった．これは学習器のトレーニングに用いるデータ数が不足していると考えられる．

6.　おわりに

本稿では，LinearSVC を用いて，授業で得られる出席状況，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，最も良い部分で約 7 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．しかし，入力パラメータの増加に対して結果の改善を見ることができなかった．そこで，遅刻状況や確認テストの取り組み状況といったような新しい種類の入力パラメータを利用したり，再現率の低いデータは入力から省くといったことによる予測精度の検討を行う必要がある．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．今後，学習器の見直しなど含め検討を進めたい．

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] scikit-learn Machine Learning in Python, 入手先 ⟨https://scikit-learn.org⟩ (2019 年 02 月確認)
[5] 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 02 月確認)
[6] 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 02 月確認)
[7] pandas.DataFrame.sample, 入手先 ⟨https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.sample.html⟩ (2019 年 02 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 02 月確認)