理解度確認テストの不合格者早期発見に向けた合格週予測モデルの構築

芳野 洸太 *　　竹川 佳成 *　　平田 圭二 *

概要. 本学の必修科目であるプログラミング演習の講義では，毎週実施されている理解度確認テストに合格することが単位認定の条件となっている．しかし，理解度確認テストに合格することができない学生が，例年 2 割存在する．毎年 2 割の受講者が増えると，全員が教室に収まらない，教師への負担が大きくなるという問題が生じる．そこで，本研究では，学生の理解状況を早期に把握し，学習を支援するアプリケーションの開発を目指し，学生が合格する週を予測するモデルを構築する．具体的には，学生の受験データを利用し，学生をクラスタリングする．その後，クラスタごとに重回帰分析により回帰モデルを構築し，学生が合格する週を予測する．提案モデルを実装し，leave-one-out 交差検証によりモデルの有用性を評価した．

1　はじめに

近年，学習管理システムや e ポートフォリオなどを利用し，学習履歴をデータマイニングし分析することで，学習者の達成度の評価や将来的な能力の予測などを行う Learning Analytics(以下 LA とする) という分野が盛り上がりを見せている [1]．LA とは，ネットワークにつながったコンピュータシステムを利用した授業において，学習者のシステム利用履歴を分析し，達成度の評価や能力の予測などを行う分野のことである．海外では，いち早く教育ビッグデータに対する LA の研究と実践が行われ，国内においてもその必要性が急速に認知されつつある [2]．

LA を実践する目的の 1 つとして，単位不認定者の削減が挙げられる．単位不認定者が現れた場合，特に必修科目においては，次年度に再履修するため，受講者が増えてしまう．受講者が増加すると，受講者が教室に収まらない，教育者の負担が大きくなるといった問題が生じる．

本研究では，本学のプログラミング演習の講義を対象に LA を実践し，学生の理解度の推移を予測する．また，対象講義における単位不認定者の特徴として，単位取得条件である理解度確認テストに合格できないことが挙げられる．そこで本稿では，理解度確認テストのデータを利用し，個々の学生が合格する週を予測するモデルを提案する．

2　関連研究

Itoh ら [6] は，ベイジアンネットワークを用いて，大学 1 年次の成績データから 2 年次の成績を予測した．Itoh ら [6] の研究における利用データは，大学 1 年次の成績データのみであり，1 年次の成績が確定するまで予測することはできない．本研究では，講義中に行われる理解度確認テストのデータを利用することで，従来研究よりも早期の予測を可能にする．また，Itoh ら [6] は，年度単位の成績を予測対象としているが，本研究では 1 つの講義における成績を対象とする．

3　対象講義について

本研究では，本学で実施されている Processing を学習するプログラミング演習の講義を対象とする．本講義は，1 回 90 分で半期に 14 回開講し，多くの受講生は初めてプログラミングを経験する 1 年生である．第 2 週以降には，単元ごとに 4 つのカテゴリに分かれている，理解度確認テストを実施している．設問数は 3，または 4 問となっており，各カテゴリの内容は表 1 の通りである．各カテゴリの実施回数は，カテゴリ 1〜3 は 11 回，カテゴリ 4 は 10 回となっている．対象講義における単位取得条件は，理解度確認テストのうち，カテゴリ 1 から 4 をそれぞれ 2 回全問正解することである．以降，全問正解することを完答とする．

表 1. 各カテゴリのテスト内容

カテゴリ	テスト内容
カテゴリ 1	矩形表示，計算，変数定義
カテゴリ 2	条件式，if 文
カテゴリ 3	for 文，while 文
カテゴリ 4	配列，関数の定義

4　提案モデル

4.1　使用したデータ

使用したデータは，2014 年度の対象講義で実施された，全 262 名の理解度確認テストデータである．説明変数は受験回数，欠席回数，週ごとの得点を用いて，目的変数は合格週とした．受験回数と欠席回数は，対象カテゴリを合格するまでに受験した回数，

* 公立はこだて未来大学

欠席した回数をそれぞれ示している．そのため，合格後の受験や欠席は数えない．週ごとの得点は，正答数を設問数で割ることで算出し，欠席の場合は 0 点とした．合格週は対象カテゴリを合格した週であるため，2 度目の完答時の週となる．なお，各データは使用する前に，正規化処理を行った．

4.2 予測手法

提案手法による合格週の予測は，4 段階に分けることができる．まず，学生を回帰モデルにフィッティングさせるため，Ward 法を用いて学生のクラスタリングを行う．単純に学生全体を回帰モデルで表現しても，個人差による影響を受け，一部の学生にしか当てはまらないモデルになってしまう．この個人差による影響を緩和するため，類似した傾向を持つ学生を同じクラスタとする．そして，クラスタごとの回帰モデルを構築することで，クラスタ内の学生に適したモデルを作ることができる．また，クラスタリングの際に，学生のデータと分類されるクラスタを SVM に学習させ，識別関数を導出する．この識別関数を利用することで，未知データがどのクラスタに分類されるのか判断することができる．最後に，図 1 に示すように，未知データを識別関数により分類されるクラスタを識別し，識別されたクラスタの回帰モデルを利用し合格週を予測する．

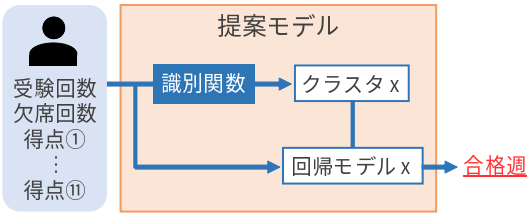

図 1. 提案モデルの全体図

5 評価

予測結果を四捨五入した値が，正解データと一致しているものを正答とし，合格週の予測結果の正答率を調べた．また，ここで予測する合格週というものは，最も早く合格可能な第 5 週（カテゴリ 4 のみ第 6 週）から，最後に受験可能な第 14 週までの値，5 から 14 をとる．予測する合格週の最大値は 14 となっているため，不合格の場合は 15 とした．この時，予測結果が 15 以上の場合は不合格と捉えられるため，予測結果が 15 以上の場合はすべて 15 とし，正答率を計算した．同様に，予測結果が 5 以下の値の場合も 5 に変更した．

評価の際には，A から L クラスの 262 人のデータから，データがまったく同じものを取り除き，評価した．データがまったく同じものが複数あると，leave-one-out 交差検証で取り出したデータがトレーニングデータにも含まれる場合がある．この方法では，

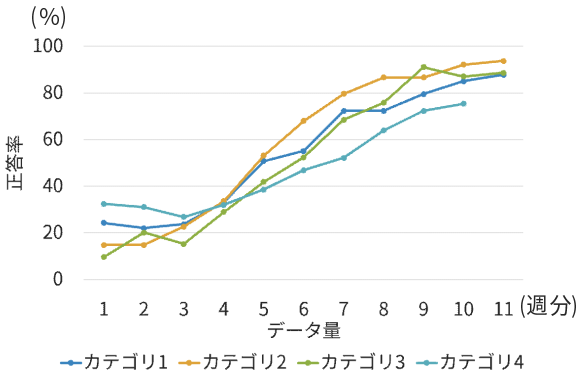

図 2. 合格週の予測正答率

正当に予測精度を評価できないと判断し，まったく同じデータは除外した．

また，本研究では理解度確認テスト不合格者の早期発見を目的としている．そのため，実施された 14 週のデータすべてを使わず，第 13 週までのデータや第 12 週までのデータのように，少ないデータ量で予測をする．提案モデルの評価の際には，データ量に応じた正答率を評価した．

合格週の予測結果を図 2 に示す．全 11 週分のデータを使った場合を見ると，すべてのカテゴリで 7 割以上の正答率となっている．カテゴリごとの正答率を見ると，カテゴリ 4 の正答率が他のカテゴリと比べて低くなっている．最も高い正答率となったカテゴリは，カテゴリ 2 であった．

6 まとめ

本研究では，本学のプログラミング演習の講義における理解度確認テストの不合格者早期発見を目的とし，学生の合格週予測モデルを構築した．提案手法では，学生を ward 法を用いてクラスタリングした後に，クラスタごとの回帰モデルを構築し，学生の合格週を予測した．提案モデルの予測精度を評価した結果，全 11 週分のデータを使用した場合，実施しているテストのすべてのカテゴリで，7 割以上の正答率が得られた．

参考文献

[1] 山川修. Learning Analytics とは. 情報処理, Vol.55, No.5, pp. 495-503, 2014.

[2] 近藤伸彦. 大学におけるビッグデータ・アナリティクスと教学 IR. 大手前大学 CELL 教育論集, No.6, pp. 11-18, 2016.

[3] H. Itoh, K. Itoh and K. Funahashi. Forecasting Students' Grades Using Bayesian Network Models and an Evaluation of Their Usefulness. The Journal of Information and Systems in Education, Vol.11, No.1, pp. 32-41, 2012.