情報処理学会論文誌　教育とコンピュータ　Vol.6　No.2　52–60 (June 2020)

論　文

小テストの点数パターンによる学習者のクラスタリングとその推定

古川 雅子[1,2,a)]　逸村 裕[2]　山地 一禎[1]

受付日 2019年8月7日，再受付日 2019年12月25日，
採録日 2020年2月29日

概要：大規模公開オンライン講座 MOOC は，対象者を限定せず誰でも受講可能なオンラインコースとして，背景やニーズの異なる様々な学習者が学ぶことができる反面，ドロップアウトする学習者が多く，その修了率は 10%程度と低い．学習者のドロップアウトを防ぎ学習効果を高めるためには，より早い時期に学習者の特徴を把握し，それぞれの学習者に応じたサポートを行う必要がある．本稿では，ドロップアウト軽減のための学習行動の分析手法として，小テストの点数パターンに基づき学習者をクラスタリングし，特徴を明らかにした．さらに，それらの特徴をもとに学習者クラスタを推定することを試みた．具体的には，国立情報学研究所が提供した MOOC の講座を対象に，4 回の小テストの点数を並べた 4 次元データを個人ごとの特徴量とし，k 平均法により 4 つのクラスタに分割した．そしてクラスタごとの学習行動を比較することで，小テストの合計点が高いクラスタは映像の視聴回数が多く，短い間隔で繰り返し映像を視聴しているといった特徴を明らかにした．また，学習開始時からの学習履歴をもとに，ランダムフォレストを用いたクラスタの推定を行い，3 週目以降では，おおむね 7 割を超える正答率で，クラスタの推定ができることを明らかにした．これらの成果は，より学習者ごとに適応したサポートを実現するうえでの基礎の 1 つとなることが期待される．

キーワード：MOOC，ラーニングアナリティクス，k 平均法，クラスタリング，ランダムフォレスト

Analysis and Estimation of Learning Behaviors based on the Score Pattern of Quizzes

Masako Furukawa[1,2,a)]　Hiroshi Itsumura[2]　Kazutsuna Yamaji[1]

Received: August 7, 2019, Revised: December 25, 2019,
Accepted: February 29, 2020

Abstract: MOOCs are online courses that can be learned by many learners with different backgrounds and needs. But in general, many learners drop out and their completion rate tends to be as low as 10%. In order to prevent learners from dropping out and enhance the learning effect, it is necessary to grasp what kind of features the learners have and to support them according to the features of each learner. In this paper, we propose a method to analyze such features of the learners. That is, the learners are clustered based on the score pattern of quizzes, the characteristics of each cluster are clarified, and the cluster to which each learner belongs is estimated based on the result. The programming course developed at the National Institute of Informatics was used to evaluate the proposed method. Four-dimensional data of four quiz scores were taken as features of learners, and divided into four clusters by the k-means method. As a result, it was revealed that the cluster with a high total score of quizzes views movies many times and repeatedly at short intervals. Cluster estimation using random forest was performed, and it was clarified that the cluster can be estimated with about 70% of the accuracy rate after 3 weeks from the start of learning. These results are expected to be one of the bases to realize more adaptable support for each learner.

Keywords: MOOC, learning analytics, k-means method, clustering, random forest

1. はじめに

教育の情報化が加速する中，大規模公開オンライン講座（以下，MOOC）は，対象者を限定せず誰でも大学レベルの教育を受講可能なオンラインコースとして，世界各地

[1] 国立情報学研究所
National Institute for Informatics, Chiyoda, Tokyo 101–8430, Japan
[2] 筑波大学
University of Tsukuba, Tsukuba, Ibaraki 305–8577, Japan
[a)] furukawa@nii.ac.jp

でサービスが提供され，広く社会から支持を得ている．国内においても 2013 年に JMOOC がサービスを開始し [1]，2019 年現在では累計 340 講座が提供され，学習者数は 100 万人を超えている [2]．各大学や研究所が特徴的な講座を提供する中，国立情報学研究所でもプログラミング入門講座「はじめての P」を開講した [3]．

MOOC のようなオンラインコースでは，学習者の膨大な操作ログがシステム上に蓄積される．この学習行動に関連するログを収集・分析することで，学習における問題点を解決し教育改善につなげようとするラーニングアナリティクス（以下，LA）が，現在，脚光を浴びている [4]．LA の主な研究テーマには，学習者の行動推定 [5], [6] や，学習者の興味や能力に応じた教材を提供しようとする試み [7], [8] 等があり，そのほかにも教育改善に関する様々な研究が行われている．MOOC の講座では，大規模な受講者数が期待できる反面，講座を最後まで修了できずに途中でドロップアウトする学習者が多い．一般的な修了率は 10%程度と低い傾向にあり，MOOC が解決すべき重要な課題として取り上げられている [9], [10]．教師と学習者が対面できる大学の講義とは異なり，MOOC では学習者とのつながりが比較的希薄となるため，問題解決の手段としての ICT への期待も自然と高まる．MOOC の有効性をさらに高めるためにも，ドロップアウトの問題を LA により改善することには大きな意義がある．

ドロップアウトの問題に取り組んだ LA の研究としては，これまでにもいくつかの報告がある．Lee ら [11] は，過去 10 年間に発表された中等教育後のオンラインコースのドロップアウトに関する研究をレビューし，ドロップアウトに影響を与える要因としては，学生の要因，コース/プログラムの要因，環境の要因があることを示した．Tan ら [12] は，学生の年齢，専攻，学習した科目数，テストの平均点等を特徴として利用し，90%ほどの正解率でドロップアウトの予測が行えることを示した．Manrique ら [13] は，学生の特徴として，登録した科目数や成績の平均といったグローバルな特徴を利用した場合と，登録した科目の成績を並べたローカルな特徴を利用した場合と，成績の時間変化を利用した場合の比較を行い，ドロップアウトの推定には，ローカルな特徴が有効であることを示した．Gitinabard ら [14] は，教材の視聴等の学習行動とともに，ディスカッションの中心になっているか等の指標も利用してドロップアウトの推定を行った．Park ら [15] は，学習管理システム（以下，LMS）に蓄積された，講義ノートの閲覧，課題の提出，掲示板への書き込みにともなうクリック数の推移から学習行動の転換点を検出した．

こうした従来の研究では，ドロップアウトを予測するための特徴量を抽出しようとする目的は一致するものの，それぞれの提案の内容は大きく異なり，共通の手段があるわけではない．その理由としては，対象とする学習者群の特徴や，LMS の機能や使われ方の違いがあげられる [16]．利用するシステムが異なれば，ログとして取得できるユーザの行動内容も異なることになる．また，学生が履修する授業全体のなかでの LMS の使われ方や LMS 以外のシステムから得られる情報も，LA を実施する際の制約条件になりうる．こうした状況を考慮すると，MOOC のドロップアウトを軽減させるための LA は，MOOC に関する特徴的な情報を加味した方法で実施する必要がある．

MOOC の受講者の類型として，藤本ら [17] は，課題に取り組まずに講義視聴のみを行う受講者や，少し試しただけで脱落する受講者等，参加動機の異なる学習者がいることを指摘した研究があることを述べている．また，ほとんど学習を行わない「様子見型」や講義ビデオの閲覧のみを行う「知識獲得型」等，4 つのタイプに学習者を分類できることを指摘した研究があることを述べている．Kahan ら [18] は，MOOC の参加者 21,889 名を，ビデオ講義，ディスカッションフォーラム，評価というコースの主な学習リソースにおける活動に基づいて 7 つのクラスタに分類した．その結果，学習活動を表すすべての特徴量の平均値が非常に低く，全主要学習資源の活動も非常に低いという特徴をもつ Tasters と呼ぶユーザ群が全体の 65%程度存在していたことを示した．この Tasters は，前述の少し試しただけで脱落する受講者や，ほとんど学習を行わない「様子見型」に対応する学習者群と考えられる．MOOC におけるドロップアウト率を軽減するための LA では，Tasters と呼ばれるような受講者群が存在することをふまえたうえで，どのクラスタの受講者をどのような特徴量を指標としてすくっていくかを考慮しながら，ドロップアウト率を軽減する方法を講ずる必要がある．入試等でフィルタリングされている大学の学生に対する支援とは異なり，MOOC の受講者のすべてを修了に導くことを前提とするのは非現実的である．従来の研究では MOOC を対象としてドロップアウトを推定するものはあるが [19], [20]，MOOC 受講者のよりきめ細かいサポートを考えた場合は，これらの特徴を考慮しつつも単純にドロップアウトを推定するだけではなく，MOOC 受講者がどのような学習者クラスタに属するかの推定を行うことが必要となる．

そこで本研究では，MOOC におけるドロップアウト率の改善策を講じる一環として，学習支援対象を特定し，その対象群をできるだけ早期に推定するための方法を確立することを目的とする．具体的には，小テストの点数は修了条件に直接的に関係することから，開講したプログラミング MOOC 講座の学習者を対象として，小テスト得点を用いたクラスタリングにより受講者を細分類し，各クラスタの特徴を明らかにする．そして，それぞれの学習者がどのクラスタに属するかを推定することで，学習支援対象となる学習者を推定することを試みる．

2. プログラミング MOOC 講座の概要

評価実験に使用した MOOC 講座は，国立情報学研究所が JMOOC のプラットフォームの 1 つである gacco [21] で開講した「はじめての P」である．講座は 2016 年 8 月 9 日から 70 日間開講された．講義は，3 人の講師と 1 人のナビゲータが担当した．

講座は，全 4 回（以下，Unit とする）で構成され，各 Unit の講義タイトルと内容は**表 1** のとおりである．各 Unit は 15 分～30 分程度の動画 3～5 本による講義と，5 問程度の小テストで構成した．小テストの点数が，全体の平均で 70 点以上になる場合は，修了証が発行される．また，受講者の属性を把握するために，開始時および終了時に Web 上でアンケートを実施した．

「はじめての P」の受講者数は 6,859 名となり，直近 1 年の gacco における受講者数平均の 4,139 名を上回った．また，修了率は 18%であり，gacco の平均修了率が 15%，MOOC の世界的なレベルでの修了率も 10%程度という中で比較的高い修了率となった．

表 1　講座の内容
Table 1　Content of the lecture.

Unit	タイトル	内容
1	プログラマになる ―プログラミングの 魅力を学ぼう！―	講師の体験をもとにプログラミングの魅力や学習方法等について紹介する．Unit2 の演習を行うための基礎的なプログラミングの知識（変数と代入，四則演算，値の種類，配列）を学ぶ．
2	プログラミングのいろは ―ビットくんのツイートをいじり倒そう！―	Web ブラウザ上で簡単な JavaScript プログラムを入力することで，ビットくんのツイート表示を改造する．プログラミングの基礎（ステートメント・ループ・条件分岐・関数）について学ぶ．
3	プログラミング入門 ―ビットくんのゲームを完成させよう！―	ビットくんのゲーム (車にぶつからないように家にたどり着く) を通じて，自分にもプログラミングが出来そうという感覚を持ってもらう．ゲームプログラムの変更を通じて case 文，関数を学ぶ．
4	アルゴリズム入門 ―プログラミングの理論を体験で学ぼう！―	身近な題材を使って，コンピュータの背後にある数理のエッセンスを学ぶ．選択ソートとマージソート，右手法と幅優先探索，二進法と XOR．

3. 小テストの点数パターンによる学習者のクラスタリングとその推定手法

MOOC における Tasters のような受講者群の存在を加味した場合，ドロップアウトからすくう受講者群を単純なコースの修了者と非修了者から分けるのでは不十分である．

非修了者となりうる群にも，特徴的なサブクラスタが存在し，それぞれにおいて適切な学習支援を提供できる可能性がある．そこで本研究では，学習行動の分析手法として，小テストの点数パターンに基づき学習者を複数のクラスタに分割し，その特徴を明らかにするとともに，それらをもとに学習者クラスタを推定する手法の提案を行う．

具体的には，

step.1 コース終了時の小テストの点数パターンに基づき学習者をクラスタリングし，学習者にどのようなクラスタがあるかを明らかにする（4.1 節にまとめる）

step.2 それぞれの学習者クラスタの分析を行い，その特徴を明らかにする（4.2 節にまとめる）

step.3 受講開始から N 週目までに得られる特徴量をもとに学習者のクラスタを推定する（5 章にまとめる）

という手順で分析を行う．

小テストの点数は修了条件に直接的に関係するため，たとえば，ある Unit 以降，小テストを受けていなければ，その時点でドロップアウトをしたと考えられる．また，Unit ごとで小テストの点数の差があれば，各 Unit の難易度を推定するための根拠として利用できる．このように，小テストの点数は，学習者の学習行動や習得度の違いを把握するために利用できるとともに，複数回の小テストがあれば，より細かい単位で学習者のクラスタ分けを行えると考えられる．

そこで step.1 では，クラスタリングに小テストを利用することとした．小テストが全部で N 回あれば，その点数を並べた N 次元のベクトルがそれぞれの学習者の特徴となる．クラスタリングには，一般に用いられる k 平均法を利用する．これにより，学習者にどのようなクラスタがあるかを明らかにすることができる．このクラスタが，そのコースに特有のものであることを仮定すれば，任意の学生について，コース終了時の小テスト点数と，各クラスタ重心との距離を求め，最も距離が近いクラスタを求めることで，その学生がどのクラスタに属するかを決めることができる．

step.2 では，step.1 で分割されたクラスタの分析を行い，それぞれのクラスタの特徴を明らかにする．特徴としては，たとえば，教材のアクセス数等が考えられるが，特に，MOOC のような自由な時間でアクセスできるコースについては，受講を開始した時間や，繰り返して学習を行う間隔といった特徴が学習行動を反映していると考えられる．また，開始アンケートの結果も，学習者の属性情報を分析

するうえでの有用な情報源として活用する．

step.3 では，受講開始から N 週目までに得られる step.2 で述べた特徴量をもとに，step.1 で述べたコース終了時の小テスト点数によって決まる学習者のクラスタを推定することを行う．クラスタを推定するための手法としては，線形判別分析，サポートベクトルマシン等，様々な手法があるが，ランダムフォレストの場合，どの特徴量がクラスタの推定に寄与しているかを知ることができることから，ランダムフォレストを用いて，学習者クラスタの推定を行う．ランダムフォレストは，機械学習に用いられる手法の 1 つであり，複数の決定木の多数決により分類を行う手法である．説明変数の重要度を計算できるという特徴がある．

以下では，実際の MOOC 講座の学習履歴をもとに，提案手法を用いて，学習者のクラスタリングとその推定精度の評価を行う．

4. 学習者のクラスタリングと特徴の分析

4.1 小テスト点数による学習者のクラスタリング

学生の属性情報を合わせて分析を行うため，開始アンケートに回答した 2,415 名を対象として，3 章で提案した step.1 に基づき，学習者のクラスタリングを行った．

対象としたコースの各 Unit には，合計 100 点満点となる小テストが用意されている．この 4 つの小テストの点数を並べた 4 次元データが個人ごとの特徴量になる．そして，この 4 次元データを k 平均法によりクラスタリングすることで，学習者をいくつかの特徴的なクラスタへと分割する．

分割するクラスタ数については，エルボー法によって決定した．エルボー法では，クラスタ数を徐々に小さくしながら，それぞれのデータとそのデータが属するクラスタ重心との自乗誤差の和を計算し，その値が大きく増加する直前のクラスタ数を最適なクラスタ数とする．分割するクラスタ数を変化させた場合の自乗誤差の和を**図 1** に示す．図 1 を見た場合，クラスタ数を 4 から 3 にしたときに，自乗誤差の増加が大きくなっていることから，クラスタ数を 4 とした．

各小テスト得点を 4 次元データとして，k 平均法により，4 つのクラスタにクラスタリングした結果を**図 2** に示す．図 2 では，4 つのクラスタ，それぞれの重心の点数を示している．点数が高い順に A，B，C，D のラベル付けを行っており，それぞれのクラスタの人数は，**表 2** のようになる．

クラスタ A は，すべての小テストにおいて，ほぼ満点をとっている．クラスタ B は，他のクラスタに比べて人数は少ないものの，Unit1 から Unit3 まではほぼ 100 点，Unit4 は，ほとんど 0 点という特徴的な点数分布になっている．対象とした MOOC 講座では，小テスト点数が，全体で 70 点以上になる場合に修了証が発行されることから，クラスタ B は，修了証が発行された時点で学習を終了した学習者が属するクラスタと考えられる．クラスタ C は，Unit1

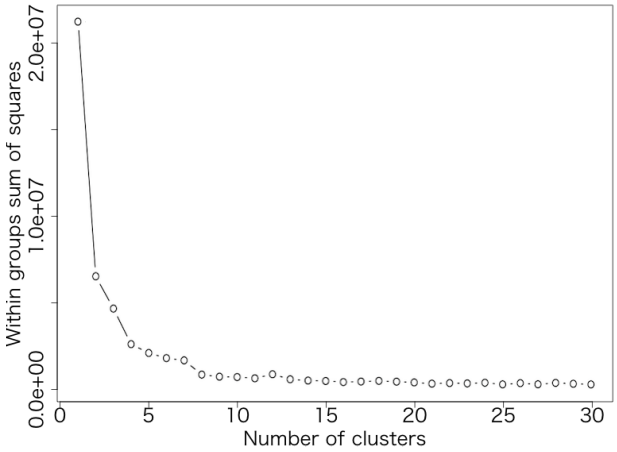

図 1　クラスタ数の決定
Fig. 1　Decision of the number of clusters.

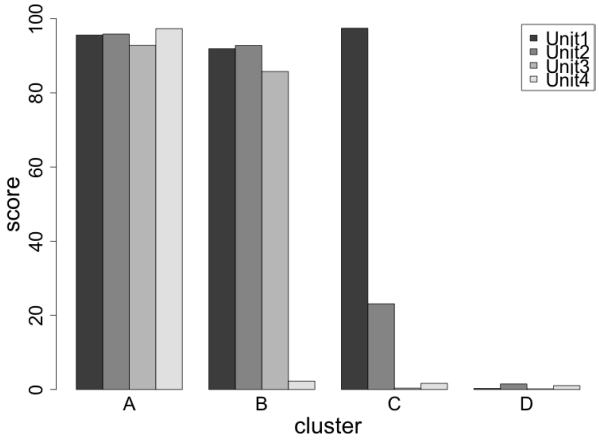

図 2　小テスト成績のクラスタリング結果
Fig. 2　Clustering results of quiz scores.

表 2　各クラスタの人数
Table 2　Number of learners in each cluster.

クラスタ	A	B	C	D
人数	793	80	606	936

はほぼ 100 点であるものの，Unit2 は，ほぼ 20 点，それ以降は，ほぼ 0 点となっており，最初の 1 回は学んだものの，それ以降は，学習をやめてしまったクラスタと考えられる．クラスタ D は，Tasters のように，すべての Unit について，小テストの点数がほぼ 0 点であり，学習をほとんど行っていないクラスタと考えられる．

4.2 学習者クラスタの分析

以下では，3 章で提案した step.2 に基づき，各学習者クラスタの特徴を解析した結果について説明する．具体的には，開始アンケートから得られた受講者の年齢や，プログラミングのスキルレベル，また受講を開始した時間や，繰り返し学習を行う間隔といった学習行動に関する情報を分析した．

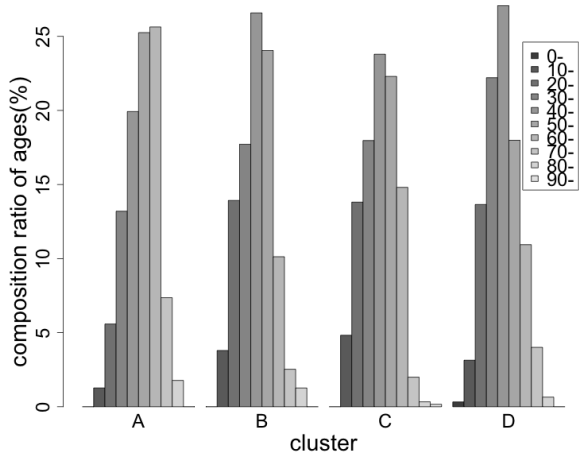

図 3　年齢の分布
Fig. 3　Distribution of ages.

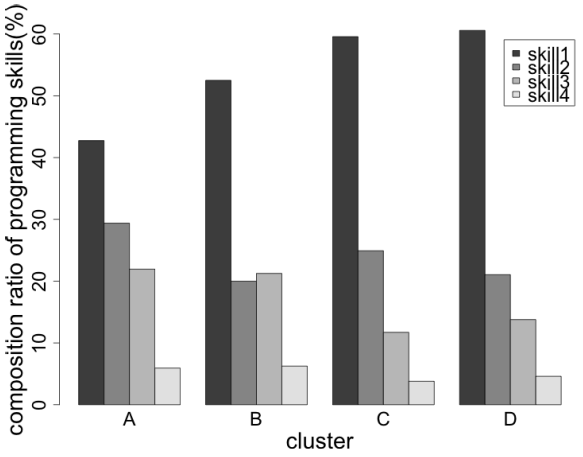

図 4　プログラミングスキルの分布
Fig. 4　Distribution of programming skills.

4.2.1　学習者クラスタと開始アンケート

ここでは，学習者クラスタと開始アンケートとの関係性について調べた．

図 3 に年齢の分布を示す．縦軸は，それぞれの年齢の構成割合を示している．開始アンケートでは，生年を答えるようになっていたため，開講年度である 2016 年から生年を引いた値を年齢とした．ただし，生年については，数値を入れる形式であったため，開講した 2016 年時点での年齢が 0 から 100 歳の範囲以外の不自然な値になっているものが 24 件あったため，これらは除外して分布を求めている．平均年齢は，A，B，C，D の順に，52.4，44.4，44.6，44.0 歳で，全体の平均年齢は，46.9 歳となった．年齢の分布を見ると，B，C，D では，40 代が最も多く，A では，50 代，60 代が多いことが分かる．

図 4 にプログラミングスキルの分布を示す．開始アンケートでは，プログラミング経験について，

skill1. 未経験，
skill2. プログラミング入門書や入門サイトで勉強したことがある，
skill3. 自分で考えたプログラムを作ったことがある，
skill4. 日常的にプログラミングをしている（専攻，職業），

の 4 つから選択するようになっている．縦軸は，それぞれの構成割合を示している．A から D に向かうに連れてプログラミングの未経験者が増えていくことが分かる．

4.2.2　学習者クラスタと学習行動

次に，学習者クラスタと学習行動との関係を見る．対象とした講座の学習は基本的に映像を視聴することにより行われるため，映像の視聴回数は，学習行動の特徴の 1 つとして利用することができる．

gacco のプラットフォームは，オープンソースの MOOC プラットフォームである Open edX [22] をベースに構築されており，学習ログから play_video イベントを抽出することで，視聴回数の情報を得ることができる．しかし，実際に，play_video イベントを抽出すると，play_video イベントが 1 秒間に数回発生し，それが数十秒にわたって続くといった場合があることが分かった．これは，視聴環境によっては，再生できなかった場合等，play_video イベントを発生し続ける場合があったためと考えられる．このような大きく外れた値を避けるため，視聴回数の分析については，各 Unit の映像を 1 日に 1 回以上見た場合に 1，それ以外を 0 として分析を行った．各 Unit には 3〜5 本の映像が含まれており，それぞれの映像ごとに視聴回数をカウントすることもできるが，クラスタリングに用いたテストは Unit 単位で行われるものである．また，各 Unit は，Unit ごとに映像の数が異なっていたとしても，同程度の学習時間を想定していることから，個々の映像の視聴回数ではなく，Unit ごとの視聴回数の方が，学習量の単位として，より適切であると考えられる．このため，Unit ごとの視聴回数をカウントすることとした．教材は 4 つの Unit で構成されるため，それぞれの学習者の映像の視聴回数は，1 日で最大 4 増えることになる．

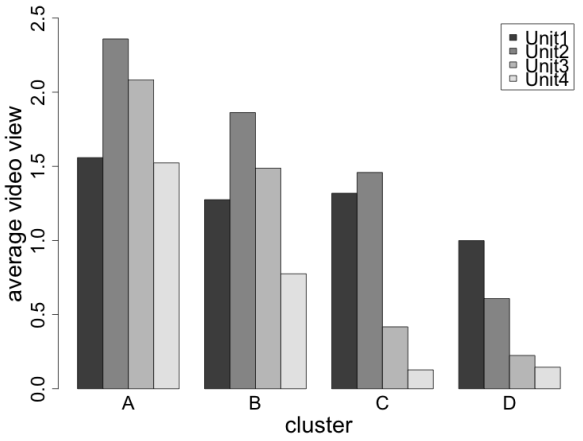

図 5　映像視聴数の分布
Fig. 5　Distribution of video views.

図 5 にクラスタごとの映像視聴数の分布を示す．縦軸

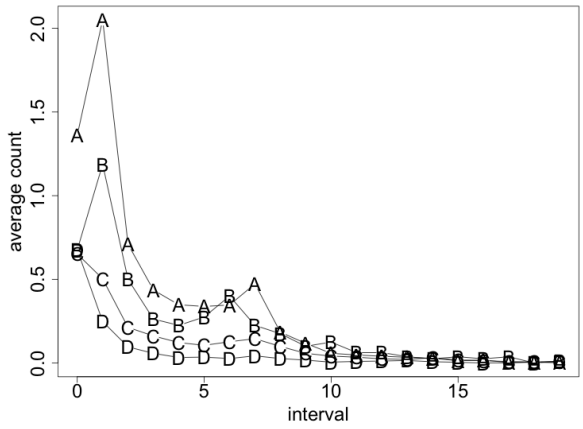

図 6　映像の視聴間隔の分布
Fig. 6　Distribution of intervals of video views.

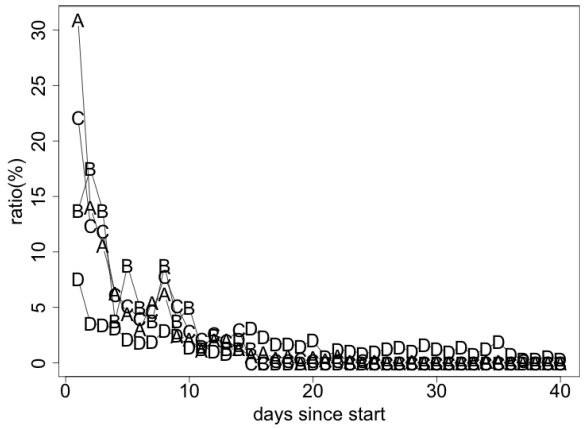

図 7　映像の視聴開始日の分布
Fig. 7　Distribution of first access to videos.

は，各 Unit の映像の平均視聴数である．この図を見ると，成績が良い A，B，C，D の順に，映像の視聴数が多いことが分かる．また，クラスタ B は，Unit4 の小テストの点数がほぼ 0 であるものの，Unit4 の映像を，他の映像の半分程度であるが見ていることが分かる．また，クラスタ C は，小テストの点数が Unit2 以降低くなるものの，Unit2 の映像を Unit1 の映像と同程度見ていることが分かる．

このことから，学習をやめるのは，映像を視聴した後，小テストを行う直前であることが分かる．クラスタ D については，小テストの点数が全 4Unit でほとんど 0 であるものの，Unit1 の映像を，ある程度見ていることが分かる．

図 6 に，映像の視聴間隔の分布を示す．映像の視聴間隔は，短い期間で集中的に学習しているか，あるいは，長い時間をかけて学習しているかの目安として利用することができる．映像を見た同じ日に別の Unit の映像を見ていれば間隔は 0 日，前日に映像を見ていれば間隔は 1 日とカウントしている．もし，1 回以下しか映像を見ていないのであれば，間隔を求めることができないので，この図のカウントには含まれない．

図 6 を見ると，クラスタ A，B では，0 日や 1 日の間隔で映像を見ていることが分かる．これは，集中的に学習している学習者と考えられる．また，クラスタ A，B では，間隔が 6 日，7 日の部分にもピークが見られる．これは，1 週間に 1 度程度，定期的に学んでいる学習者と考えられる．クラスタ C や D では，1 度以下しか映像を見ない学習者が多く，平均のカウント数は小さくなっている．

図 7 に，最初に映像を視聴したのが，講座開講から何日目であるかの分布を示す．開講日を 1 日目としている．この図を見ると，いずれも 1 日目から受講している人数が多い．これは，講座がいつ始まるかをあらかじめ知っていて，開講初日からアクセスをした学習者と考えられる．

それぞれのクラスタごとに，映像を 1 回以上視聴した学習者の平均開始日を求めると，クラスタ A，B，C については，4.7，4.9，5.2 となり，開講から 1 週間以内にアクセスを開始する学習者が多いことが分かる．一方，クラスタ D については，平均で 17.1 となり，2 週目以降にアクセスする学習者の割合が多いことが分かる．これは，講座の開講後に講座の存在を知り，アクセスをした学習者の割合が多かったためと考えられる．

5. 学習者クラスタの推定

4 章では，学習者を 4 つのクラスタに分け，それぞれのクラスタごとに学習行動に違いがあることを明らかにした．以下では，3 章で提案した step.3 に基づき，それぞれの学習者の属するクラスタを推定する．

ドロップアウト率を軽減する策を講じるためには，学習者がどのクラスタに属するかを，学習を開始してからなるべく早い時期に推定する必要がある．これにより，学習者に応じた適切なサポートを行うことが可能になり，修了率や満足度の向上を図ることができると考えられる．このため，学習者から得られた様々な特徴から，学習者クラスタをどの程度推定できるか評価を行った．

クラスタの推定に利用した特徴量について，開始アンケートからは，以下の特徴量を利用した．

1) 年齢

年齢については，図 3 を求めた場合と同様の処理を施したが，年齢が 0〜100 の範囲以外の不自然な値になる 24 件については，全体の年齢の平均値の小数点以下を切り捨てた 46 歳として扱った．

2) プログラミングの経験

プログラミングの経験については，図 4 に示した skill1 から skill4 までの内容を，それぞれ，1 から 4 で数値化した．数値が大きいほど，プログラミングの経験があるという指標になる．

学習行動については，以下の特徴量を利用した．

1) ビデオの視聴数

各 Unit の映像を 1 日に 1 回以上見た場合に 1，それ以外を 0 として視聴数を求めた．教材は 4 つの Unit で構成されるため，映像の視聴回数は，4 次元のデータになる．

2) 視聴間隔のヒストグラム

これは，集中的に短い間隔で視聴するか，間を空けて見るかの特徴になる．20 日以上間を空けて視聴する数はほぼ 0 になるため，視聴間隔は 0 日から 19 日までの 20 次元のデータとした．

3) 開講から何日目に学習を開始したか

学習開始時点で，開講から何日目かが分かる．一度もアクセスがなければ閉講後の 71 日目を学習開始日とした．

以上の年齢 1 次元，プログラミングの経験 1 次元，ビデオの視聴数 4 次元，視聴間隔のヒストグラム 20 次元，開講から何日目に学習を開始したかの 1 次元を足し合わせた 27 次元のデータを各学習者の特徴とした．

学習者がどのクラスタに属するかの推定には，ランダムフォレストを用いた．ランダムフォレストの処理には，R 3.5.3 上の randomForest パッケージ（バージョン 4.6-14）を利用した．また，構成する木の数等，すべてのパラメータは，デフォルト値のまま利用した．

ランダムフォレストの場合，ジニ係数の減少量を見ることで，どの特徴量がクラスタの推定に寄与しているかを知ることができる．このジニ係数の減少量を求めるため，すべての学習が終わった 10 週後の，2,415 件のすべてのデータを学習データとし，1 つの学習モデルを作成した．その場合のジニ係数の減少量を**図 8** に示す．開始日と Unit4 のビデオの視聴数（V4）がクラスタの推定に特に重要であることが分かる．また，Unit2，Unit3 のビデオの視聴数（V2，V3）や年齢も上位に位置している．DN は，学習間隔が N 日の数を表し，D0，D1，D2 は短い間隔での繰り返しの学習，D6，D7 は，週 1 程度での繰り返しの学習に対応し，これらの数もクラスタの推定に有用であることが分かる．

クラスタの推定精度の評価については，学習データと評価データを分けて評価を行うため，10-fold クロスバリデーションを行った．すなわち，N 週目の精度を求めるために，N 週目までに取得できた視聴回数等のデータを 10 に分け，そのうち 9 のデータを使って，A，B，C，D のラベルを推定できるように学習を行った．そして，その学習結果をもとに，残りの 1 つのデータのラベルの推定を行うという処理を 10 回繰り返し，その正答率（推定したラベルのうち正しかった割合）の平均を求め N 週目の精度とした．

受講開始から N 週後のクラスタ推定の正答率を**図 9** に示す．マーカが + のものがクラスタ A，B，C，D の 4 つに分類した場合を示しており，3 週目以降は，70%を超える正答率で学習者クラスタの推定ができている．また，クラスタ C については，クラスタ C と推定すべきもののうち，3 週目で平均 63.6%，10 週目で平均 68.2%を正しく推定することができた．

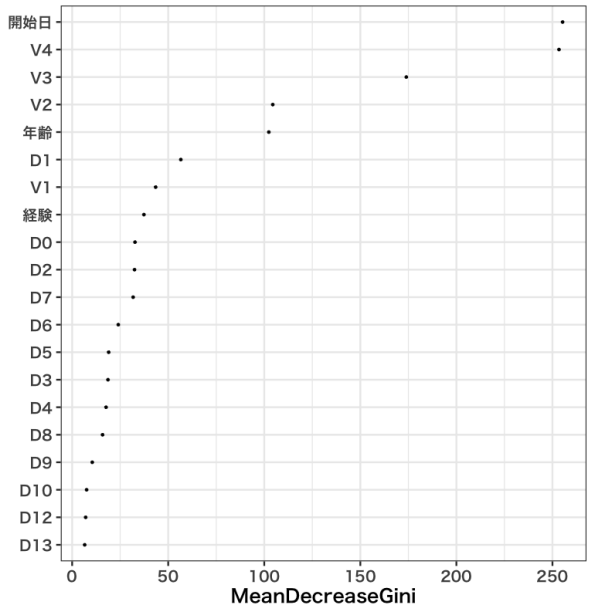

図 8　ジニ係数の減少量

Fig. 8　Mean decrease Gini.

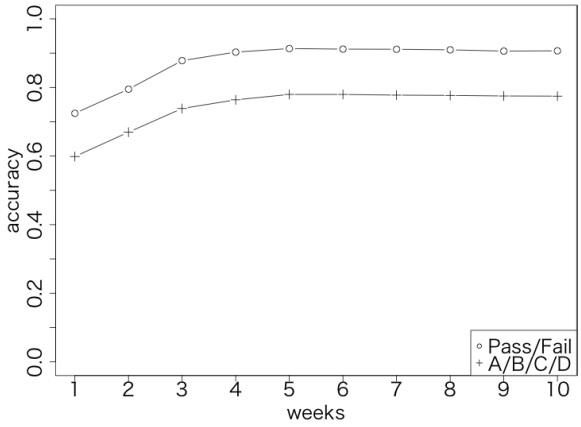

図 9　学習行動からのクラスタの推定

Fig. 9　Prediction of learners' cluster.

一方，小テスト全体の点数が 70 点以下でクラスタ A，B に分類されたものが 40 例あったが，それ以外については，クラスタ A，B に属していれば合格，クラスタ C，D に属していれば不合格であり，クラスタ A，B と，クラスタ C，D の分類は，ほぼ，講座の合格，不合格に対応する．

マーカが○のものが，10-fold クロスバリデーションを用いて，クラスタ A，B と，クラスタ C，D の分類を行った場合を示しており，これは，ほぼ，最終的な合格，不合格の分類に対応する．3 週目以降は，ほぼ 90%の正答率で学習者クラスタの推定ができていることが分かる．

6.　まとめと考察

本稿では，学習行動の分析手法として，小テストの点数パターンに基づき学習者をクラスタリングし，それぞれの学習者クラスタの特徴を明らかにするとともに，その結果

をもとに，それぞれの学習者の属するクラスタを推定する手法の提案を行った．また，実際に，国立情報学研究所で開発したプログラミング講座を対象に，提案手法を用いて，学習者のクラスタリングとその推定精度の評価を行った．

具体的には，全 4 回分の小テストの点数を並べた 4 次元データを個人ごとの特徴量とし，k 平均法により 4 つのクラスタに分割し，クラスタごとの学習行動を比較した結果，小テストの合計点が最も高いクラスタ A は映像の視聴回数が多く，短い間隔で繰り返し映像を視聴しているといった特徴があることが明らかになった．2 番目に小テストの点数が高かったクラスタ B は，最初の 3 回分は，小テストでほぼ満点をとっているものの，Unit4 の小テストの点数は 0 点に近かった．開講した講座では，テスト全体で 70 点以上をとれば修了証が出ることから，学習することよりも，修了証をとることが目的のクラスタと考えられる．クラスタ C は，最初の 1，2 回のみを学習し，それ以降の学習をやめてしまったクラスタと考えられる．クラスタ D は，小テストの点数がほとんどすべて 0 点であり，また，映像への最初のアクセスが他のクラスタに比べて遅いことから，学習するというよりも，講座の開講後，講座があることを知り，どのような講座かを見るためにアクセスしてきたクラスタと考えられる．

ここで，学習者の支援を考えた場合，クラスタ A とクラスタ B は，基本的に合格点に達していることから，合格点に達していないクラスタ C やクラスタ D が支援対象の候補になると考えられる．しかしながら，クラスタ D は，必ずしも学習することが目的ではないと考えられることから，クラスタ D を合格にまで導くことは必ずしも容易ではないと考えられる．一方，クラスタ C は，開講すぐにアクセスする人の割合が高く，あらかじめ開講を知っていたと考えられ，また，ユニット 1 の小テストでは，ほぼ満点をとっていることから，講座を終了するために必要な能力は基本的に持っていると考えられる．このため，より早い時期に，学習者がどのクラスタに属するか，特にクラスタ C に属する学習者を推定できれば，効率的に，学習の際にドロップアウトを防ぎ，修了率を高めることができると考えられる．そこで，学習開始時からの学習履歴をもとに，ランダムフォレストを用いたクラスタの推定を行い，3 週目という早い時期に，クラスタ C については平均 63.6%，全体では 70%を超える正答率で，クラスタの推定ができることを明らかにした．また，最終的な合格，不合格については，ほぼ 90%の正答率で推定ができることを示した．

以上のように，小テストの点数パターンに基づき学習者をクラスタリングすることで，より細かく学習者の特徴を明らかにすることができた．また，それらの特徴をもとに高い精度で学習者クラスタを推定できることを示した．

これらの成果は，MOOC 上で，より学習者に適応したサポートを実現するうえでの基礎の 1 つとなると考えられる．たとえば，ドロップアウトすると推定される時点の前に，特定の学習者クラスタに対してサポートのメールを送ることや，より理解を促進するようなコンテンツを推薦するといったことが考えられる．また，コース内容を十分に理解していると考えられるクラスタと，ドロップアウトする可能性の高いクラスタの学習者同士の交流をフォーラム上で促進するといったことも考えられる．

学習者のクラスタリング結果については，本研究では，1 つの特定の MOOC コースを対象として分析を行ったものであり，一般の MOOC コースを対象とした場合，必ずしも，本研究の結果のような 4 つのクラスタに分類できるとは限らない．たとえば，クラスタ B は，コースの修了条件に依存したクラスタと考えられ，一般には，学習者がどのようなクラスタに分割されるかは，コースに依存すると考えられる．一方，学習者を小テストの点数によってクラスタリングし，学習者が属するクラスタを推定するという枠組みについては，今回，クラスタ C のような特徴的なクラスタの存在を明らかにし，その推定もある程度の精度で実現できていることから，学習者に多様性があると考えられる他のコースを分析する際の手法の 1 つとして利用できると考えられ，今後，その検証を行っていく予定である．また，学習行動に基づく支援のあり方や教材改善等についても，さらに検討を行っていく予定である．

参考文献

[1] 福原美三：日本初 MOOC の可能性と課題，研究報告教育学習支援情報システム（CLE），Vol.2014-CLE-12, No.1, p.1 (2014).

[2] JMOOC, available from ⟨https://www.jmooc.jp/en/⟩ (accessed 2019-07-01).

[3] 国立情報学研究所：はじめての P，入手先 ⟨https://www.nii.ac.jp/service/jmooc/hajimete/⟩（参照 2019-07-01）.

[4] New Media Consortium: Learning Analytics and Adaptive Learning, *NMC Horizon Report 2016 Higher Education Edition*, pp.38–39 (2016).

[5] 緒方広明，殷　成久，毛利考佑，大井　京，島田敬士，大久保文哉，山田政寛，小島健太郎：教育ビッグデータの利活用に向けた学習ログの蓄積と分析，教育システム情報学会誌，Vol.33, No.2, pp.58–66 (2016).

[6] Käser, T., Hallinen, N.R. and Schwartz, D.L.: Modeling exploration strategies to predict student performance within a learning environment and beyond, *Proc. 7th International Learning Analytics & Knowledge Conference* (*LAK '17*), pp.31–40, DOI: 10.1145/3027385.3027422 (2017).

[7] Kandula, S., Curtis, D., Hill, B. and Zeng-Treitler, Q.: Use of topic modeling for recommending relevant education material to diabetic patients, *Proc. AMIA Annual Symposium*, pp.674–682 (2011).

[8] Zapata-Gonzalez, A., Menendez, V.H., Prieto-Meńdez, M.E. and Romero, C.: Using data mining in a recommender system to search for learning objects in repositories, *Proc. 4th International Conference on Educational Data Mining*, pp.321–322 (2011).

[9] Reich, J. and Ruipérez-Valiente, J.A.: Supplementary Material for The MOOC pivot, available from ⟨https://

science.sciencemag.org/content/sci/suppl/2019/01/09/363.6423.130.DC1/aav7958-Reich-SM.pdf⟩ (accessed 2019-07-01).
[10] Pursel, B.K., Zhang, L., Jablokow, K.W., Choi, G.W. and Velegol, D.: Understanding MOOC students: Motivations and behaviours indicative of MOOC completion, *Journal of Computer Assisted Learning*, Vol.32, No.3, pp.202–217, DOI: 10.1111/jcal.12131 (2016).
[11] Lee, Y. and Choi, J.: A review of online course dropout research: implications for practice and future research, *Educational Technology Research and Development*, Vol.59, No.5, pp.593–618, DOI: 10.1007/s11423-010-9177-y (2011).
[12] Tan, M. and Shao, P.: Prediction of Student Dropout in E-Learning Program Through the Use of Machine Learning Method, *International Journal of Emerging Technologies in Learing*, Vol.10, No.1, pp.11–17, DOI: 10.3991/ijet.v10i1.4189 (2015).
[13] Manrique, R., Nunes, B.P., Marino, O., Casanova, M.A. and Nurmikko-Fuller, T.: An Analysis of Student Representation, Representative Features and Classification Algorithms to Predict Degree Dropout, *Proc. 9th International Conference on Learning Analytics & Knowledge* (*LAK '19*), pp.401–410 (2019).
[14] Gitinabard, N., Khoshnevisan, F., Lynch, C.F. and Wang, E.Y.: Your Actions or Your Associates? Predicting Certification and Dropout in MOOCs with Behavioral and Social Features, *Proc. 11th International Conference on Educational Data Mining*, pp.404–410 (2018).
[15] Park, J., Denaro, K., Rodriguez, F., Smyth, P. and Warschauer, M.: Detecting changes in student behavior from clickstream data, *Proc. 7th International Learning Analytics & Knowledge Conference* (*LAK '17*), pp.21–30, DOI: 10.1145/3027385.3027430 (2017).
[16] Milliron, M.D., Malcolm, L. and Kil, D.: Insight and action analytics: Three case studies to consider, *Research and Practice in Assessment*, Vol.9, pp.70–89 (2014).
[17] 藤本　徹，荒　優，山内祐平：大規模公開オンライン講座（MOOC）におけるラーニング・アナリティクス研究の動向，日本教育工学会論文誌，Vol.41, No.3, pp.305–313 (2017).
[18] Kahan, T., Soffer, T. and Nachmias, R.: Types of Participant Behavior in a Massive Open Online Course, *The International Review of Research in Open and Distributed Learning*, Vol.18, No.6, pp.1–18, DOI: 10.19173/irrodl.v18i6.3087 (2017).
[19] Hanan, K. and Martin, E.: MOOCs Completion Rates and Possible Methods to Improve Retention—A Literature Review, *World Conference on Educational Multimedia, Hypermedia and Telecommunications*, Vol.2014, No.1, pp.1305–1313 (2014).
[20] Cheng, Y. and Gautam, B.: Early Prediction of Student Dropout and Performance in MOOCs using Higher Granularity Temporal Information, *Journal of Learning Analytics*, Vol.1, pp.169–172, DOI: 10.18608/jla.2014.13.14 (2014).
[21] gacco, available from ⟨http://gacco.org⟩ (accessed 2019-07-01).
[22] Open edX, available from ⟨https://open.edx.org/⟩ (accessed 2019-07-01).

古川 雅子　（正会員）

2015 年国立情報学研究所情報社会相関研究系助教．オープンサイエンス基盤研究センター助教．筑波大学大学院図書館情報メディア研究科博士後期課程．専門分野は，ラーニングアナリティクス，MOOC 等 e ラーニング教材開発・評価．

逸村 裕

1980 年慶應義塾大学文学部図書館・情報学科卒業．慶應義塾大学文学研究科図書館・情報学専攻修士課程修了．愛知淑徳大学文学研究科図書館情報学専攻博士後期課程修了．1980 年上智大学図書館員．1991 年愛知淑徳大学文学部助教授．2002 年名古屋大学附属図書館研究開発室助教授．2002 年文部科学省研究振興局学術調査官（併任）（2008 年まで）．2006 年筑波大学大学院図書館情報メディア研究科教授．現在，筑波大学図書館情報メディア系教授．学術情報流通と評価，大学図書館，情報探索行動研究に従事．

山地 一禎　（正会員）

1994 年豊橋技術科学大学工学部情報工学課程卒業．2000 年同大学大学院工学研究科電子・情報工学専攻修了．博士（工学）．理化学研究所脳科学総合研究センター研究員等を経て，現在，情報・システム研究機構国立情報学研究所コンテンツ科学研究系教授．オープンサイエンス基盤研究センターセンター長．高等教育機関における ICT 基盤整備に関する研究開発に従事．文部科学省平成 30 年度文部科学大臣表彰科学技術賞（開発部門）受賞．

情報処理学会論文誌　教育とコンピュータ　Vol.6　No.2　52–60 (June 2020)

正誤表

下記の箇所に誤りがございました．お詫びして訂正いたします．

訂正箇所	誤	正
52 ページ	受付日 2019 年 8 月 7 日	受付日 2019 年 7 月 16 日