1F1-OS-26a-3

時系列クラスタリングを利用した未就学児の学習データ分析

Learning Analytics for Preschool Children Using Time Series Clustering

内藤純平 *1 Junpei Naito　馬場雪乃 *1 Yukino Baba　鹿島久嗣 *1 Hisashi kashima　高木丈智 *2 Takenori Takaki　布野卓也 *2 Takuya Funo

*1京都大学大学院情報学研究科知能情報専攻
Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University

*2しまねソフト研究開発センター
Shimane IT Open-Innovation Center

Learning analytics attempts to utilize data analysis technologies to support students' learning process and improve the quality of education. In spite of the increasing attention to learning analytics in higher education, it has not been fully considered in primary and preschool education. In this research, we apply learning analytics to preschool education, and try to predict the continuation of learning by preschool children. Based on our intuition that time-change patterns of the assessment scores are effective features, we use time-series clustering to extract such patterns. The experiments using the real preschool education dataset show that the use of time-change patterns improves prediction of future continuations of study.

1. はじめに

学習データ分析 (Learning Analytics) とは、教育においてデータ解析の技術を適用し、学習者の学習サポートや、より良い教育を行うための知見を得ようとする分野である。学習時間や、受けている講義の内容、使っている教材などのデータを解析することによって、今後の学習の傾向について予測する、学習に関して問題を抱えている学習者を特定し、対処を行う、学習者に対して適切な教材を推薦すること等を目的として行われる。近年では、MOOC (Massive Open Online Course) の導入等、インターネットを介して教育を行うようになったことで、学習時間や、試験の評点、課題の提出状況等のデータが蓄積されるようになってきたことを受け、学習データ分析は、高等教育を中心に、脱落しそうな学習者を発見すること等を目的として、盛んに行われるようになってきている [Siemens 11]。

一方で、幼児教育、初等教育を対象とした学習データ分析はあまり行われていない。その原因としては、学習時間、試験の点数と言った評点の意味合いが高等教育のそれとは必ずしも同じではなく、全く同じ方法でそのまま評価することができないということが挙げられる。また、インターネットやコンピュータを利用した講義などが行われることもなく、単純に得られるデータの種類、量が少ないことも原因としてあるだろう。また、未就学児の学習の評価に関しては、子供の能力を計測するのに試験を行う必要があるが、ペーパーテスト等を使って客観的な指標を提供することが難しい。試験を行うのに保護者や教師の補助が必要になることが考えられ、さらに、試験結果を出す際にも主観的な判断が入る可能性がある。さらに、高等教育について大きな興味となっているドロップアウト予測に関しても、未就学児の教育に関しては違う解釈をする必要がある。子供が学習を継続するかどうかの判断は、必ずしも本人が行うものではない。本人が学習に対して意欲を示しているかどうかは影響してくるだろうが、子供が興味を示しているか判断することに加え、学習が価値のあるものかどうかを判断する保護者の影響が大きく現れるだろう。

総合的に見て、高等教育と、幼児教育や初等教育とでは、考慮すべき要素が違ってくることに加え、統計処理を目的としてデータを蓄積することも一般的ではなく、実施された回数が少ないこともあり、幼児教育や初等教育における学習分析の有効性はまだ十分に示されたとは言えない。

本研究では、未就学児の学習データの分析を行い、学習の継続を予測することを検討する。これは、高等教育におけるドロップアウト予測に相当する問題である。学習を分析するにあたり、評価基準が客観的でなく、保護者に依る主観を含んだ評価データに対してロジスティック回帰を用いて分析を行い、それに基づいて将来その子供が学習を継続して行うかを予測する。特に、評価データの中でも、評価点数の時間変化が有用な情報であると考え、これを予測のための特徴量として利用する。そのために、時間変化のデータについて時系列クラスタリングを行い、分類されたクラスタを特徴量として予測に加える。時系列クラスタリングには、階層的クラスタリングを用いる。

しちだ・教育研究所から提供されたデータを利用して実験を行う。子供の年齢が 0 歳から 2 歳までのデータを使い、2 歳以降に学習を継続するかどうかをロジスティック回帰で予測し、予測の精度をみたところ、時系列クラスタリングを使わない場合の AUC 値が 0.7549 であったのに対し、時系列クラスタリングを使った場合、最大で AUC 値が 0.8001 と向上した。また、子供の年齢が 2 歳から 4 歳までのデータを使い、4 歳以降に学習を継続するかどうかをロジスティック回帰で予測した結果は、時系列クラスタリングを使わなかった場合の AUC 値が 0.8014 に対し、時系列データを使った場合の AUC 値は最大で 0.8327 と向上した。これにより、未就学児の学習分析には、時系列データのクラスタリング結果が学習の継続を予測するのに有用であることが実験的に示されたと言える。

2. 関連研究

学習分析における主要な目的の一つは、成績不振者や中退者を事前に予測し、学習者の支援に役立てることである。そのため、注意が必要な生徒を予測するための機械学習の利用が検討されている。

連絡先: 内藤純平, naito.junpei.45m@ml.ist.i.kyoto-u.ac.jp

Tamhane らは、Grade 8（日本の中学 2 年生）時点での成績不振者の予測を行った。ロジスティック回帰を使い、Grade 1 ～ 7 の長期間の成績情報が、予測に有効であることを示した [Tamhane 14]。Aguiar らは、Grade 6 ～ 12 を対象に、中退者とそのタイミングを予測した。その際に、「Grade 6 終了時点で中退するか」、「Grade 7 終了時点で中退するか」等の予測問題にすることで、中退者と中退のタイミングを予測できるようにした [Aguiar 15]。Lakkaraju らは、高校を中退する生徒を予測するための枠組みを提案した。この枠組みを使うと、ランダムフォレストや、ロジスティック回帰などの様々な機械学習法を比較評価することができる。また、重要な特徴を可視化することも可能である。生徒のデータとして、GPA の他に、授業の欠席・遅刻率経済状況などを利用した [Lakkaraju 15]。Vihavainen らは、大学のプログラミング講義の成績不振者を、プログラミング時の行動情報から予測した。コードの修正履歴から、プログラミングにかけた時間、修正前後のコードの編集距離、修正の時間間隔などを取得し、行動情報として利用した [Vihavainen 13]。

これらの研究は、全て中等教育、高等教育を対象にしている。一方で、幼児教育の成績不振者・非継続者予測における機械学習法の有効性は、十分に示されていない。

3. 問題設定

3.1 未就学児の学習継続予測問題

本研究では、未就学の子供の学習継続の問題を、ある時期までのデータを利用して、その時期以降に学習を継続するか否かを予測する 2 値分類問題として捉える。2 歳と 4 歳を区切りとして、0 歳から 2 歳のデータを利用して 2 歳以降の学習継続を予測する問題と、2 歳から 4 歳のデータを利用して 4 歳以降の学習継続を予測する問題を扱う。予測に利用するデータは、保護者に依る主観の入った子供の評価データと、子供と保護者に関する情報が含まれるデータである。

3.2 分析対象のデータ

株式会社しちだ · 教育研究所より提供されたデータを用いる。このデータは子供を一意に特定する情報である家庭ごとの会員 ID や、その子供がその家庭の何番目の子供かを表す子供 No、子供の生年月日等に加え、しちだ · 教育研究所が提供している発達検査の結果の時系列データを含んでいる。この発達検査は、「身体的発達」、「知覚的発達」、「言語的発達」、「社会性の発達」の 4 つの科目があり、それぞれの科目に対し、いくつかのチェック項目が用意されている。例えば、身体的発達のチェック項目には、「棒のぼりの棒に 5 秒くらいつかまっている」「野球のボールを 2〜4 メートル投げる」などがある（図 1）。それらのチェック項目を、子供の保護者が記入し、その回答に基づいて、科目の点数が決定する方式である。客観的な指標があるチェック項目もあるが、評価者が子供の保護者であるため、その主観が入ってくる可能性がある。

ある時期以降の学習を継続したかは、ある時期以降にデータが存在するかで定義する。2 歳以降の学習継続を予測する問題では、2 歳以降にデータを 1 つ以上持っている子供を学習を継続した正例であるとし、データを 1 つも持っていない子供は学習を継続しなかった負例であるとする。4 歳以降の学習継続を予測する問題でも同様に、4 歳以降にデータを 1 つ以上持っている子供を学習を継続した正例であるとし、データを 1 つも持っていない子供は学習を継続しなかった負例であるとする。0 歳から 2 歳までには正例 803 件、負例 737 件の、計 1,540 件の子供のデータ、2 歳から 4 歳までには正例 529 件、負例 903 件の計 1432 件のデータがある。これらのデータから特徴を抽出して、特徴ベクトルとした。使った特徴は、子供 No、初回検査日齢、検査時の日齢、検査回数、検査間隔、検査の 4 科目（身体的発達、知覚的発達、言語的発達、社会性の発達）の得点と、4 科目の得点の平均である。子供 No は、家庭において何番目の子供であるかを表す値である。検査日の日齢、検査の 4 科目の得点とその平均に関しては、特徴として取り出すときは、その子供が受けた全ての発達検査から得られる値の平均値とする。


身体的発達

チェック欄	チェック項目
	85. 棒のぼりの棒に5秒くらいつかまっている
	86. 野球のボールを2〜4メートル投げる
	87. とんで前後左右に移動できる
	88. 20センチの高さのゴムひもをとび越える
	89. ブランコを立ってこぐ
	90. 立ち幅とびで70〜80センチとぶ

図 1: チェック項目例（しちだ · 教育研究所 発達検査ブック 3 より引用）

4. モデリング手法

子供が学習を継続するかどうかという問題を、ロジスティック回帰を使ってモデル化する。その中で、時系列クラスタリングを用いて、予測に利用する。

一人の子供に対して継続的に発達検査を行うため、データの時間変化を観察することができる。この時間変化が似ている子供は、学習を継続するか継続しないかに関して、似たような傾向があるだろうという発想に基づき、データの時間変化を予測に利用する。発達検査の 4 科目の得点について、時間ごとの変化を時系列データとみなし、4 科目のそれぞれについて、時系列クラスタリングを行う。科目ごとに子供をクラスタに分類し、子供が属しているクラスタを特徴として予測に利用する。仮に科目ごとに 10 個のクラスタに分類する場合、属しているクラスタに対応する要素が 1、残りの要素が 0 となるような 10 次元のベクトルを 4 つ、特徴として予測に利用することになる。

クラスタリングには、階層的クラスタリングを用いる。階層的クラスタリングは、全てのデータ、およびクラスタ間の距離（類似度）を計算し、その中で最も距離が小さい（類似度が大きい）２つのデータまたはクラスタを統合し、新たなクラスタとすることを繰り返す方法である。最終的には全てのデータを１つのクラスタに統合する。仮に 10 個のクラスタに分ける場合、統合されていないデータ、およびクラスタの数が 10 個になった時点で終了する。階層的クラスタリングをする際には、クラスタ間の距離、あるいは、データとクラスタ間の距離を計算する必要がある。これには、単連結法（最近隣距離法）、完全連結法（最遠隣距離法）、群平均法の 3 種類の方法を使う。

時系列データ間の距離には、まず、時系列データを単に固定長のベクトルと見て、それらの差のノルムをとるユークリッド距離がある。ユークリッド距離の計算は簡単であるが、計算したい 2 つの時系列データについて、データ数が同じでないと、ユークリッド距離を定義することができない。また、データが

記録された時間が違うと、計算することができない。そこで、データ数が違ったり、記録された時間が違ったりしても計算が可能な DTW 距離も利用する。また、DTW 距離であれば、形が似ていても時間がずれている、というデータ間の距離を計算するとき、ユークリッド距離と比較して距離が小さくなる。

データ点の数が違ったり、データが記録された時間が違ったりしても距離を計算する別の方法として、欠損しているデータの補間がある。2 つのデータ点の間にある値を利用したいときに、その近辺のデータ点を使って値を導出する。本研究では、線形補間を用いた。線形補間は、ある時間に取られた 2 つのデータ点の間で、値が線形に変化しているものと仮定してデータを補間する方法である。具体的に式で表すと、2 つのデータ点 $(a_1, b_1), (a_2, b_2)$ $(a_1 \leq a_2)$ があり、$a_1 \leq a \leq a_2$ となるようなデータ点 (a, b) は、次のようになる。

$$b = \frac{(a - a_0)b_1 + (a_1 - a)b_0}{a_1 - a_0}$$

本研究では、子供の学習データ間の距離を計算するのに、線形補間の使い方を 2 通り試した。1 つ目は、あるデータ点以前に、データが存在しない場合、時間 0 におけるデータの値を 0 と仮定して線形補間を行い、あるデータ点以降にデータが存在しない場合、最後のデータ点から値が変化しないものと仮定して補間を行う方法である。2 つ目は、あるデータ点以前にデータが存在しない場合は、それ以前の範囲については距離計算に利用せず、更にあるデータ点以降にデータが存在しない場合は、それ以降の範囲については距離計算に利用せず、データが存在する範囲内のみを使って距離計算を行う方法である。1 つ目を線形補間 A、2 つ目を線形補間 B と呼ぶことにする。

ユークリッド距離と DTW 距離のどちらも、時系列データのデータ点の数が多くなると距離が大きくなってしまうため、データ点の数で正規化を行うことで、データ点の数が多い形が似ている時系列データ間の距離が小さくなるようにする。実験では、正規化した距離と正規化をしない距離の両方について検討する。

5. 実験

未就学の子供の学習について分析を行い、時系列クラスタリングを行うことで、2 値分類を使った予測の精度が向上するかを見る。0 歳から 2 歳までのデータを利用した、2 歳以降の学習継続の予測と、2 歳から 4 歳までのデータを利用した、4 歳以降の学習継続の予測を行い、特徴量やクラスタリング方法の違いにより予測精度がどのように変化するか比較する実験を行う。

5.1 比較手法

時系列クラスタリングを使う予測と、使わない予測の両方を行い、時系列クラスタリングによって予測精度が向上するかを見る。時系列データ間の距離を計算するのに、DTW 距離とユークリッド距離を使い、それぞれについて正規化を行った場合と、行わなかった場合の両方を実験する。線形補間は、4 章に示した A、B を使い、また、線形補間を行わない場合についても実験を行う。線形補間を行わない場合は、ユークリッド距離を計算できないデータが多いので、ユークリッド距離の計算はせず、DTW 距離のみ計算する。階層的クラスタリングの方法に単連結法、完全連結法、群平均法を使う。クラスタの数は 10、20、50、100、200 とした。合計で、時系列クラスタリングを 150 通り行う。比較対象である時系列クラスタリングを使わないものを加え、151 の手法について AUC の値を出し、比較する。

表 1: クラスタリング無しの AUC と特徴の重み

AUC 値		0.7549	
特徴名	重み	特徴名	重み
子供 No.	-0.0440	初回検査日齢	-1.0622
検査時の日齢	2.3269	検査間隔	0.2380
検査回数	0.3985	得点平均	0.1282
身体的発達の得点	0.0437	知覚的発達の得点	-0.1585
言語的発達の得点	0.3448	社会性の発達の得点	0.1993

表 2: 0 歳から 2 歳のデータを使った予測の精度

	AUC 値
クラスタリングなし	0.7549
クラスタリング最大値	0.8001
クラスタリング平均値	0.7516

5.2 実験手順

時系列データ間の距離を計算し、距離行列を作成する。DTW 距離とユークリッド距離のそれぞれについて、正規化した距離と、正規化していない距離を計算する。距離計算の際に、線形補間を行う場合は適用する。作成された距離行列を元に、単連結法、完全連結法、群平均法のいずれかで階層的クラスタリングを行う。その結果を元に、データを 10、20、50、100、または 200 のクラスタに分割し、分類されたクラスタを表すベクトルを子供のデータに追加する。これは、属しているクラスタに対応する要素のみが 1、その他の要素は全て 0 となる 10、20、50、100、または 200 次元のベクトルである。知覚的発達の得点、言語的発達の得点、社会性の発達の得点に関しても同様にする。

子供のデータを、分類器を作成するための学習データと、作成した分類器の精度を検証するための検証データに分ける。学習データと検証データの割合は 80 : 20 とする。学習データに対してロジスティック回帰を行い、それによってできた分類器を使って、検証データに関して予測を行う。予測の精度を、AUC を使って比較する。

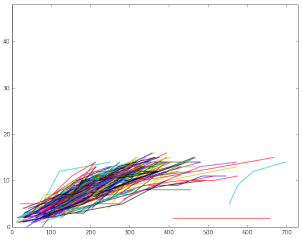
5.3 結果

まず、時系列クラスタリングを利用せずに、0 歳から 2 歳までのデータを利用した、2 歳以降での学習継続を予測した結果を示す。AUC 値と、各特徴の重みを表 1 に示す。これを見ると、重みの絶対値が大きい、すなわち、予測に大きい影響を与えているのは、検査時の日齢である。正の重みであるため、日齢が高い時点で多くの検査を受けているほど、学習を継続しやすいという傾向が分かる。また、次に重みの絶対値が大きいのは初回検査日齢だった。負の重みであるため、初めて検査を受けた時の日齢が低いほど学習を継続しやすいという傾向が分かる。

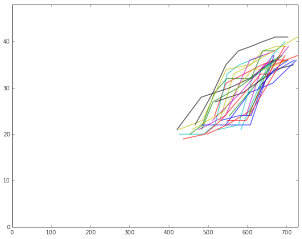
続いて、クラスタリングを行った結果について見ていく。0 歳から 2 歳のデータを利用した、2 歳以降での学習継続の予測について、クラスタリングを行った場合と、行わなかった場合での予測精度の違いを表 2 に示す。表 2 から、クラスタリング手法を適切に選ぶと、予測精度が向上していることが分かる。

表 3: 2 歳から 4 歳のデータを使った予測の精度

	AUC 値
クラスタリングなし	0.8104
クラスタリング最大値	0.8327
クラスタリング平均値	0.7597



2-1: 検査時の日齢が小さいクラスタ。学習を継続しない子供が少なかった。



2-2: 検査時の日齢が大きいクラスタ。学習を継続する子供が多かった。

図 2: 検査時の日齢の傾向を示す例

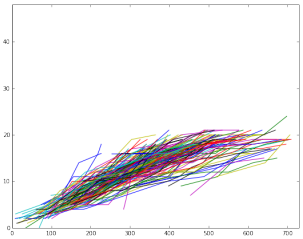
表 4: 継続した子供が多いクラスタと継続した子供が少ないクラスタにおける、知覚的発達の得点の比較

	平均	最大値
継続した子供が多いクラスタ	20.98	26.46
継続した子供が少ないクラスタ	11.51	15.61

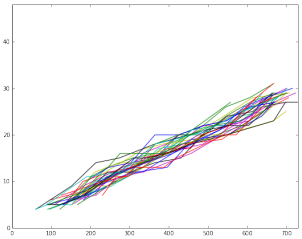
同様に、2 歳から 4 歳までのデータを利用した 4 歳以降での学習継続の予測について、クラスタリングを行った場合と行わなかった場合での予測精度の違いをを表 3 に示す。0 歳から 2 歳までのデータを使った実験よりも、クラスタリングの影響が小さいという様子が見て取れる。

クラスタの様子について詳細に見ていく。図 2 は、線形補間を行わずに DTW 距離を計算し、群平均法で 50 のクラスタに分類するクラスタリングを行った結果できたクラスタのうち、ある 2 つのクラスタである。横軸が検査時の日齢、縦軸が知覚的発達の得点となっている。なお、この時の AUC 値は 0.7701 であった。図 2-1 は正例 15、負例 157 を含む学習を継続した子供が少なかったクラスタであり、図 2-2 は正例 19、負例 5 を含む学習を継続した子供が多かったクラスタである。検査時の日齢が小さいクラスタに学習を継続した子供が少なく、検査時の日齢が大きいクラスタに学習を継続した子供が多くなっていて、日齢が高い時点で検査を多く受けているほど継続しやすいという傾向がクラスタリングの結果にも現れていることが分かる。

また、重みだけでは見えない傾向も、クラスタリングで見ることができた。図 3-1 は正例 34、負例 133 を含む学習を継続した子供が少ないクラスタ、図 3-2 は正例 27、負例 4 を含む学習を継続した子供が多いクラスタである。図 3 に見るように、学習を継続した子供が多いクラスタは得点が高くまで伸びている子供がいることが多く、学習を継続した子供が少ないクラスタは得点が低く止まっている子供がいることが多かった。実際に、学習を継続した子供が多いクラスタと、そうでないクラスタについて、それぞれの子供の平均得点と最大得点を平均したものを比較してみたところ、表 4 のようになった。知



3-1: 得点が小さいクラスタ。学習を継続する子供が少なかった。



3-2: 得点が大きいクラスタ。学習を継続した子供が多かった。

図 3: 平均点数の傾向を示す例

覚的発達の得点を、線形補間を行わない DTW 距離で、群平均法で 50 個のクラスタにクラスタリングしたものについてのみ示したが、他の科目の得点についても同様の傾向が見られ、クラスタリング方法を変えても、群平均法と完全連結法を使っているものに関しては程度の差はあるが同じ傾向が見られた。

6. おわりに

本研究では、未就学の子供に対して学習データ分析を行い、子供の学習継続の予測を行った。学習継続の予測に時系列クラスタリングを利用する場合としない場合で予測精度を比較する実験を行い、クラスタリングによって予測精度が上がっていることを観察した。加えて、クラスタの様子を見ることによって、学習を継続した子供の得点の傾向と、学習を継続しなかった子供の得点の傾向を見ることができた。

参考文献

[Aguiar 15] Aguiar, E., Lakkaraju, H., Bhanpuri, N., Miller, D., Yuhas, B., and Addison, K. L.: Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time, in *LAK 15* (2015)

[Lakkaraju 15] Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., and Addison, K. L.: A machine learning framework to identify students at risk of adverse academic outcomes, in *KDD 15* (2015)

[Siemens 11] Siemens, G. and Long, P.: Penetrating the FOG: Analytics in learning and education, *EDUCAUSE Review* (2011)

[Tamhane 14] Tamhane, A., Ikbal, S., Sengupta, B., Duggirala, M., and Appleton, J.: Predicting student risks through longitudinal analysis, in *KDD 14* (2014)

[Vihavainen 13] Vihavainen, A., Luukkainen, M., and Kurhila, J.: Using students' programming behavior to predict success in an introductory mathematics course, in *EDM 13* (2013)