Vol.2014-CLE-13 No.4
2014/5/16

LMS 蓄積データを用いた学習特徴の抽出における変数の粒度の検討

小川賀代[†1]　　ピトヨ　ハルトノ[†2]

近年，LMS の普及に伴い，アクセスログ，成績，アンケートなどの学習に関する様々なデータの取得が可能になり，膨大なデータとなりつつある．これらの履歴データをデータマイニングすることで学生の理解度や学習パターンの把握ができ，個人に適した学習の提供が期待されている．本報告では，定量的な複数の指標を用いてクラスター数を決定し，各クラスターの学習特徴を解析，可視化する方法を述べる．また，解析に使用する変数の粒度についても検討を行い，粒度の違いにより，異なる解析結果が得られることが確認でき，解析者の見たい情報に応じて粒度を変化させる必要性があることがわかった．

Evaluation of the granularity of variables for extracting learning characteristics using data accumulated by LMS

KAYO OGAWA[†1]　PITOYO HARTONO[†2]

The popularization of Learning Management Systems recently made it possible to obtain educational data such as students' access logs and grades. Now, a vast amount of data is becoming available to researchers. Data-mining of these historical data is expected to contribute to the understanding of the level of comprehension and learning characteristics of students, enabling the provision of education suitable to individual students. This paper reports a method for determining the number of clusters using multiple quantitative indices, and analyzing and visualizing the learning characteristics of each cluster. The paper also examines the granularity of variables used for the analysis. We confirmed that the analyses yields different results depending on the granularity, and found that it is necessary to change granularities according to the information an analyst is interested to obtain.

1. はじめに

近年，学習を裏付け深めるための情報として，学習管理システム(LMS:Learning Management System)から取得できる学習履歴データを活用することが期待されている．特に膨大なデータの活用のためには，機械的に学習者を類似した学習特徴を持つグループに分類し，意味付けを行うことが必要となる．これを可能とする解析手法としてクラスター分析がある．しかし，分類数（クラスター数）は経験的に決定されているため，経験者の判断に左右される．そのため，データの自然な分布を必ずしも反映しているとはいえない．また，ビッグデータの取り扱いにおいては，できるだけ機械的に行える方が望ましい．

これまでの研究において，評価観点の異なる 3 つの指標を用いてクラスター数を決定し，学習者の特徴を抽出する解析手法を提案してきた．[1] また，分類結果の直感的な理解に向け，自己組織化マップ(SOM:Self-organizing map)による可視化の検討を行った．ただし，解析を行うにあたってデータ変数の粒度の影響やどの程度の粒度で用いるべきであるかについての検討はなされていない．そこで本研究では，本研究では学習履歴データの変数の粒度を変化させたときの学習特徴の抽出を行い，粒度の違いによる結果の比較検討を行った．

†1 日本女子大学
Japan Women's University
†2 中京大学
Chukyo University

2. 学習履歴データの解析手法

本研究で提案する解析手法の概要を図 1 に示し，以下に説明する．

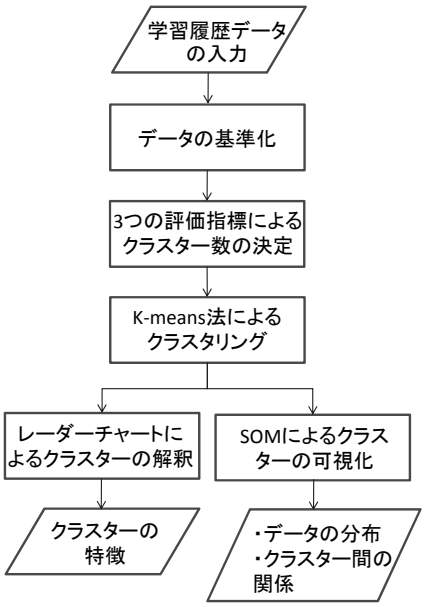


図 1　解析手法の概要

2.1 3 つの指標を用いたクラスター数の決定

学習履歴データを定量的に分類するために，クラスター分析を用いる．クラスター分析は，サンプル間の距離に基

づき類似したサンプルを 1 つのクラスターに分類する手法である．

本論文では，クラスター分析で最も用いられる手法である K-means 法で行う．K-means 法は，予め想定したクラスター数を用いて，式(1)に示す評価関数 を最小化することでデータを分類する．式(1)では，$Z_j \in \{Z_{1j}, Z_{2j}, \cdots, Z_{nj}\}$ を学習者 j の基準化した学習サンプルとし，C_k はクラスター k の重心であり，式（2）で計算する．

$$E = \sum_{k=1}^{K} \sum_{j=1}^{m} \alpha_{kj} \left\| C_k - Z_j \right\|^2 \qquad (1)$$

$$C_k = \frac{1}{m_k} \sum_{z_j \in C_k} z_j \qquad (2)$$

ただし，n：履歴項目数，m：学習者数，K：クラスター数，$\alpha_{kj} = \begin{cases} 1 & Z_j \in C_k \\ 0 & Z_j \notin C_k \end{cases}$，$m_k$：クラスター k に属する学習者数を示す．

従来，クラスター数は解析者の経験と主観に大きく依存して決定することが多いが，そのクラスター数がデータの自然な分布を反映しているとは限らない．また，想定したクラスター数を用いてクラスタリングの実行後に，その妥当性（Cluster Validity）を特定の評価指標を用いて評価するが，データには様々な分布特徴が存在するため，1 つの評価指標のみを適用すると，異なる特徴を持つデータに対し，クラスター評価の妥当性を保証することができない．そこで本解析手法では，評価観点の異なる 3 つの評価指標を用意し，データに潜む自然な分布を反映したクラスター数を決定する．ここでは，先ず，予め定めたクラスター数の範囲で K-means 法を実施し，3 つの評価指標を適応する．そして，最も顕著な結果を示す評価指標を選び，その評価指標の中で顕著に結果を示すクラスター数を決定する．このように，クラスター数の決定は，グラフ化された定量的なデータに基づいて，総合的に判断を行う．本解析手法で用いたクラスター数に対する評価指標を以下に示す．

(a) Dunn index[2]

クラスターの最大直径 D とクラスター重心値間の最小距離 M の比

$$DI = M / D \qquad (3)$$

が最大となるクラスター数を選択する．

(b) DB index[3]

各クラスターにおいて，クラスター内分散 S_i とクラスター重心値間の距離 M_{ij} の比

$$R_{ij} = \left(S_i + S_j\right) / M_{ij} \qquad (4)$$

が最大になる値を R_i として，式(3)に示す類似度

$$\overline{R} = 1/K \sum_{i=1}^{K} R_i \qquad (5)$$

を計算し，この値が最小となるクラスター数 K を選択する．

(c) ベイズ情報量基準 (BIC:Bayesian information criterion)[4]

統計学における情報量基準の 1 つで，最大対数尤度 $\hat{l}_M(D)$ とモデルのパラメータ数 $P_M / 2 \cdot \log R$ の差を計算し，この値が最大となるクラスター数を選択する．

2.2 可視化によるクラスターの分析

2．1 で得られた自然なクラスター数に従い，データを K-means 法によりクラスタリングした後，形成されたクラスターを可視化することで学習者の特性を抽出する．まず，各クラスターの重心を項目毎に基準化し，クラスター別にレーダーチャートを描いて，クラスターの特徴を把握する．レーダーチャートによる視覚的な把握を行うことで，学習者の特性を容易に解釈することができる．これと同時に，Self-Organizing Map（SOM）を用いたクラスターの可視化を行う．1 個人が有する履歴項目は LMS に残された学習ログや成績など多次元に亘るため，個々人を比較することは難しい．また，多次元データのクラスタリングを行っているため，クラスター間の関係を理解するのは困難である．そこで本解析手法では，高次元データの位相関係を保ちながら，低次元空間にそれらのデータを配置できる SOM をデータの可視化手法として用いることにした．SOM は非線形関係を捉えることができ，新しいデータをマップ上に配置することができる．また，SOM の結果の上に，クラスターごとに異なる色で結果を表示することで，クラスター間の関係を把握することができる．

3. データ解析における変数の粒度の検討

3.1 使用する学習履歴データの概要

企業が提供している e ラーニングコンテンツの学習履歴データを用いて評価実験を行った．受講者数が 238 名，講座は 10 章からなっており，3〜4 章ごとに全 3 つの章末テストと全 4 つの学習後アンケート，また最後に修了試験，模擬試験が組まれている．講義，テスト及びアンケートの受講に制限はない．変数の詳細を以下に示す．括弧内の数値は，変数の数を示している．

- 講座の受講回数(10)
- アンケートの回答回数(4)
- 各章末テストの初回点数(3)
- 各章末テストの最高点数(3)
- 各章末テストの受験回数(3)
- 修了テストの初回点数(1)
- 修了テストの最高点数(1)
- 修了テストの受験回数(1)
- 模擬試験の受験回数(1)

今回は講座の受講回数とアンケートの回答回数について，合計値を用いた粒度の粗い場合と，講座ごと，アンケートごとに分けた値を用いた粒度の細かい場合の 2 通りについて解析を行った．よって，粒度の粗い場合は 15 変数、粒度の細かい場合は 27 変数で解析を行った．また，これら全てのデータに対し基準化を行い解析を行った．

3.2 評価実験の結果

3.2.1 クラスター数の決定

データのクラスター分析を行うために，まずクラスター数を決定する．今回のデータについては解析条件をクラスター数範囲 2〜10，繰り返し回数 100 回として指標の計算を行った．その結果を図 2，3 に示す．

粒度の粗い場合は，指標値結果から極大・極小値は見当たらないが，指標値の変化量を見ると全ての指標でクラスター数 4 以降で値が小さくなっている．よって，クラスター数を 4 と決定した．一方，粒度の細かい場合は，Dunn Index と BIC においてクラスター数 8 で極大値を示している．よって，クラスター数を 8 と決定した．

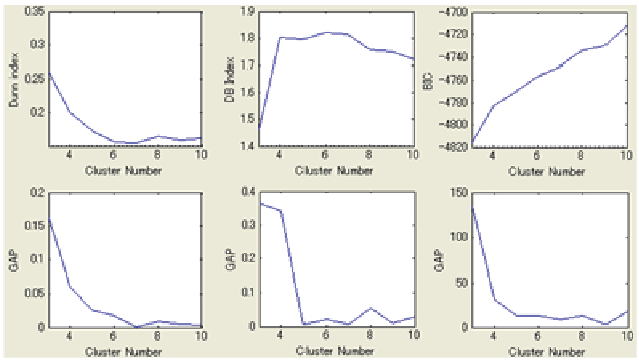


図 2　指標計算結果(上)と指標値変化量(下)
(粒度の粗い場合)

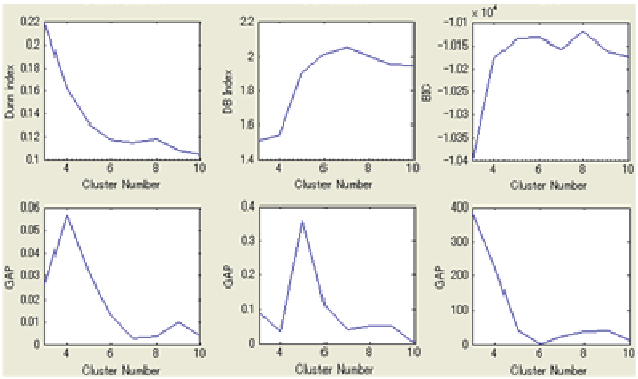
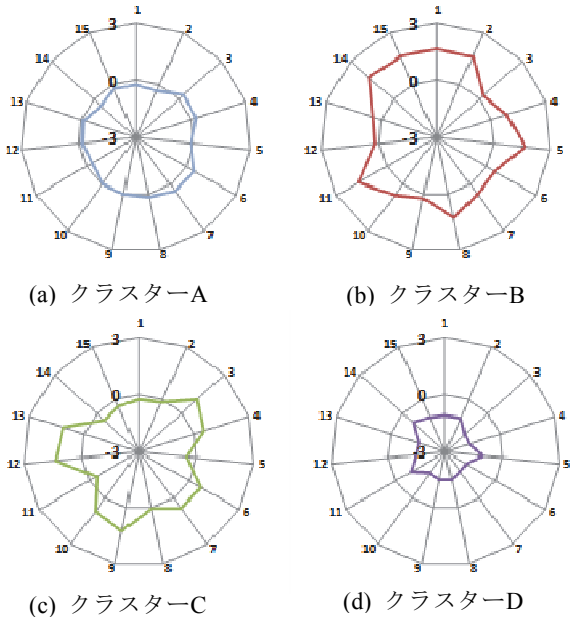


図 3　指標計算結果(上)と指標値変化量(下)
(粒度の細かい場合)

3.2.2 クラスターの特徴の把握

クラスターの学習特徴を把握するため，クラスターごとにレーダーチャートを描き，学習特徴の解釈を行った．レーダーチャートを図 4，5 に，レーダーチャートから読み取れる学習特徴と所属人数を表 1，2 に示す．なお，クラスター名は A-H とした．また，図 4，5 の軸は解析に用いた変数を表しており，軸の値の 0 はデータの平均値を示している．



(a) クラスターA　(b) クラスターB

(c) クラスターC　(d) クラスターD

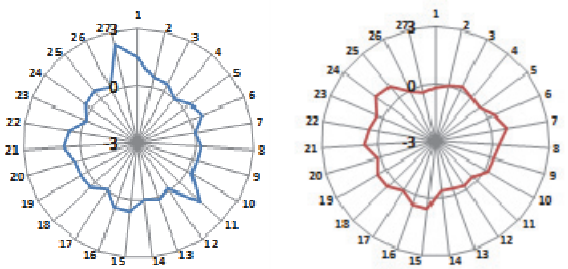
1：　10 章からなる講座の合計受講回数
2：　4 つのアンケートの合計回答回数
3-5：　章末テスト 1 の初回点数，最高点数，受験回数
6-8：　章末テスト 2 の初回点数，最高点数，受験回数
9-11：　章末テスト 3 の初回点数，最高点数，受験回数
12-14：　修了テストの初回点数、最高点数，受験回数
15：　模擬試験受験回数

図 4　各クラスターのレーダーチャート
(粒度の粗い場合)

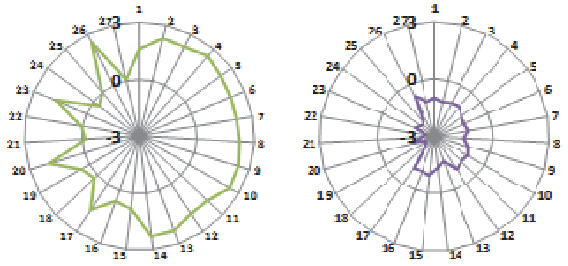
表 1　各クラスターの学習特徴と人数（粒度の粗い場合）

Cluster	学習特徴	人数(人)	割合(%)
A	平均的なクラスター	61	25.6
B	講義・テスト共に繰り返し取り組む	19	8.0
C	講義・テスト共に繰り返し取り組まないが，テスト点数は高い	145	60.9
D	全体的にコンテンツにしっかり取り組まない	13	5.5

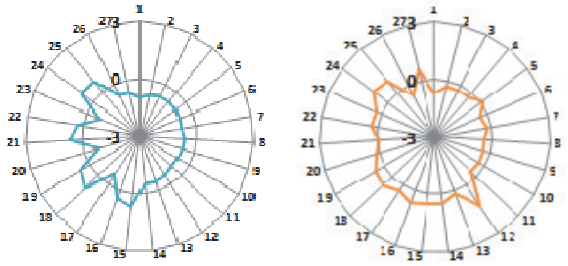
変数の粒度を細かくすることにより，クラスター数が増え，さらに細かい学習特徴を読み取ることができた．例えば，粒度の細かい場合のクラスターG と H からは，それぞれ最初のコンテンツのみよく取り組む，後半のコンテンツのみよく取り組む，といった粒度が粗かった時には見えなかった結果が得られた．これにより，学習者のコンテンツに取り組む姿勢の変移を把握することができた．



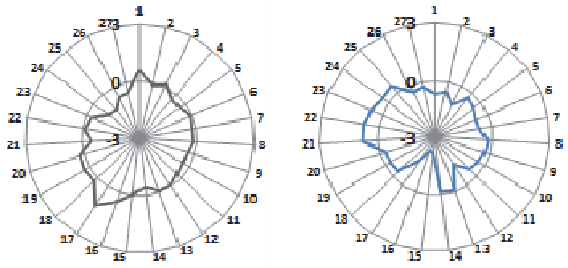
(a) クラスターA　(b) クラスターB



(c) クラスターC　(d) クラスターD



(e) クラスターE　(f) クラスターF



(g) クラスターG　(h) クラスターH

1-10： 10 章からなる講座の各受講回数
11-14： 4 つのアンケートの各回答回数
15-17： 章末テスト 1 の初回点数，最高点数，受験回数
18-20： 章末テスト 2 の初回点数，最高点数，受験回数
21-23： 章末テスト 3 の初回点数，最高点数，受験回数
24-26： 終了テストの初回点数、最高点数，受験回数
27： 模擬試験受験回数

図 5　各クラスターのレーダーチャート
(粒度の細かい場合)

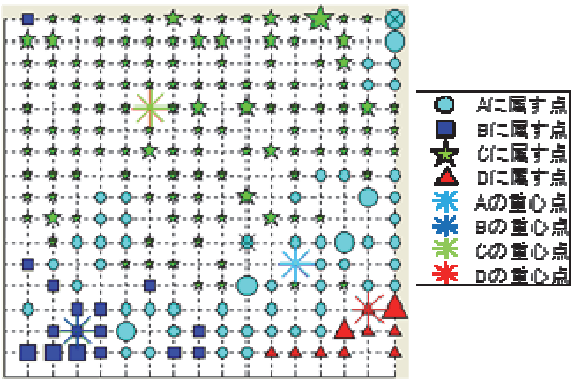
また，特に章末テストによく取り組む，講義よりテストに力を入れて取り組む，といった特徴は粒度の粗い場合にも得られてもよい特徴であるが，粒度の粗い解析結果からは，このような特徴は見られなかった．よって，変数の粒度を上げたことでより明確に特徴抽出を行うことができるといえる．

表 2　各クラスターの学習特徴と人数 (粒度の細かい場合)

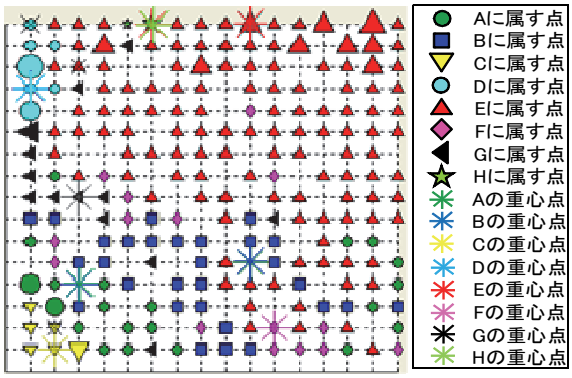
Cluster	学習特徴	人数 (人)	割合 (%)
A	全体的によく取り組むが，特に模擬試験を繰り返し取り組む	24	10.1
B	特に章末テストに取り組む	29	12.2
C	全講義を繰り返し受講し，テストも何回も取り組む	7	2.9
D	全体的にコンテンツにしっかり取り組まない	11	4.6
E	講義・テスト共に繰り返し取り組まないが，テスト点数は高い	134	56.3
F	講義よりテストに力を入れて取り組む	15	6.3
G	講義・テスト共に最初のコンテンツは頑張るが，後半になるにつれて取り組まない	14	5.9
H	全体的にあまり取り組まないが，章末テストに関しては部分的によく取り組む	4	1.7

3.2.3 SOM を用いたクラスターの可視化

続いて，SOM を用いてデータ分布の様子を把握する．マップを生成するにあたり，マップサイズを 16×16，繰り返し回数を 1500 回, 学習率を 0.1 として解析を行った．また，



(a) 粒度の粗い場合



(b) 粒度の細かい場合

図 6　SOM による学習履歴データの可視化

3.2.1 の結果で得られた各学習者の所属クラスターが，どこにマッピングされたかわかるように, SOM の結果の上に，クラスターごとに異なるマークと色で示し，各クラスタの重心を「＊」で示した．更に，同じユニットに同クラスターに属するデータが重なる場合はプロット点を大きくし，異なるクラスターに分類されるにもかかわらず，データが同じユニットに重なる場合は「×」印を付けた．結果を図 6 に示す．図 6(a)を見ると，クラスターA については，性質が最も異なるとされる（2 次元空間において最も距離が離れている）対角線の位置にも分布していることがわかる．一方，図 6(b)では，クラスター数が増えたことも要因だと考えられるが，全体的にデータがクラスターごとにまとまって分布する結果が得られた．このことから，変数の粒度を細かくして解析を行ったことで，類似した特徴を持つデータがよりまとまるようにクラスタリングされたことが，SOM の結果より把握することができる．

4. まとめ

本報告では，企業が提供している e ラーニングコンテンツから取得した LMS に蓄積されている学習履歴データを用いて学習者の特徴を抽出するために，定量的な複数の指標を用いてクラスター数を決定し，各クラスターの学習特徴を抽出，可視化する方法を用いて解析を行った．更に，蓄積学習履歴データの変数の粒度を変化させて解析したときの学習特徴の抽出を行い, 比較を行った. この結果より，変数の粒度の細かい方が，より明確な学習特徴を得られることが確認できた．また，SOM による可視化の結果においても，粒度が細かい方が，よりクラスターごとにまとまって分布する結果が得られた．これらの結果より，変数の粒度は解析結果の精度に影響を及ぼすことがわかった．しかし，学習履歴データを蓄積するサーバーの容量には制限があるため，どの変数をどの粒度で用いればよいかは，目的に応じて粒度を変化させて解析を行う必要がある．

謝辞 本研究で使用した解析データは，株式会社プロシーズからご提供頂いたものである. ここに感謝申し上げる. また，本研究は，科学研究費補助金基盤研究(C)（課題研究番号 25330419）の助成を受けたものである．

参考文献

1) 石川晶子，小川賀代，ピトヨ ハルトノ: 学習履歴データを活用した学習者の特徴抽出方法の検討，教育システム情報学会誌，Vol．31， No．2， pp．185-196 (2014).
2) Dunn, J. C．: A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters，Cybernetics and Systems，Vol．3，No．3，pp．32-57(1973).
3) Davies， D． and Bouldin，D．: A cluster separation measure，IEEE Trans. Pattern Analysis and Machine Intelligence，Vol. PAMI-1，No．2，pp．224-227(1979).
4) Schwarz, G. :Estimating the dimension of a model，Annals of Statistics，Vol．6， No．2，pp．461-464(1978).