4ZC-03

プログラミング初学者教育における要支援者予測のための ログデータクラスタリング解析

橋本 玄基 †　　大枝 真一 ‡
木更津工業高等専門学校 制御・情報システム工学専攻 †　　木更津工業高等専門学校 情報工学科 ‡

1. はじめに

日本の教育システムは一対多のものが多く，定期的に行われる試験での評価が主流である．そのため，学生の状態把握が難しい．しかしプログラミング教育においては UNIX コマンド履歴や，ソースコードの編集履歴などのログデータという形で容易かつ自動的に学生の行動の一部を確認することができる．ログデータの活用法を考える研究はいくつか存在する [1, 2, 3]．しかし，成績評価関数の検証が不十分であったり，評価機が集団内の分類を行っているものであり，集団が変わった際に適用可能であるかが考慮されていなかったり，また，スキル評価を行うことは困難としているものもある．

そこで本研究では成績評価は従来どおり試験評価の役割とし，要支援者の予測に焦点を絞り解析を行った．教師なし学習であるクラスタリングを用いた外れ値検出を利用して要支援者の予測方法を検討した．

2. 提案手法

学生の授業中のコマンド履歴から累積入力数を 5 分毎に集計した時系列ベクトルを特徴ベクトルとする．次に特徴ベクトル群に対して距離計算を Dynamic Time Warping(DTW) で行い，初期値決定に k-means++[4] を用いた k-medoids 法 [5] を適用する．DTW による M, N 次元時系列ベクトル $\boldsymbol{x}_M, \boldsymbol{x}_N$ の距離計算方法は以下のとおりである．

1. $M \times N$ 行列 $\boldsymbol{D}$ を，$\boldsymbol{D}_{1,1} = 0$，それ以外の要素を ∞ と定義する．
2. 全ての $i(= 2, \cdots, N)$，$j(= 2, \cdots, M)$ について以下の式で $\boldsymbol{D}_{i,j}$ を求める．

$$\boldsymbol{D}_{i,j} = \sqrt{(x_{Ni}, x_{Mj})^2} + \min(\boldsymbol{D}_{i-1,j}, \boldsymbol{D}_{i,j-1}, \boldsymbol{D}_{i-1,j-1}) \quad (1)$$

3. $\boldsymbol{D}_{N,M}$ を距離とする．

クラスタ数が k，それぞれのクラスタに属するデータ集合を $C_1, C_2, \cdots, C_k$，クラスタの medoid(クラスタの中心に最も近いデータを指し，クラスタ 1 の medoid は $\boldsymbol{x}_{med_1}$ となる) を $med_1, med_2, \cdots, med_k$ とする．このとき観測データ集合 $X = \{\boldsymbol{x}_i \mid i = 1, 2, \cdots, n\}$ であるとすると提案手法全体のアルゴリズムは以下のとおりである．

1. $d(\boldsymbol{x}_i, \boldsymbol{x}_j)(i, j = 1, 2, \cdots, n)$ を求める．$d(\boldsymbol{x}, \boldsymbol{y})$ は $\boldsymbol{x}$ と $\boldsymbol{y}$ の距離を表す．
2. クラスタの medoid のインデックス med_1 を $\{1, 2, \cdots, n\}$からランダムに 1 つ選ぶ．
3. $D(\boldsymbol{x})(\boldsymbol{x} \in X)$ を求める．$D(\boldsymbol{x})$ は決定されたクラスタの medoid と $\boldsymbol{x}$ 間の最短距離を表す．
4. 下式の確率分布に従って選ばれたデータのインデックス j を次のクラスタの medoid のインデックス med_i とする．これを medoid が k 個できるまで繰り返す．

$$P(\boldsymbol{x}) = \frac{D(\boldsymbol{x})^2}{\sum_{\boldsymbol{x} \in X} D(\boldsymbol{x})^2} \quad (2)$$

5. 全データ $\boldsymbol{x}_i(i = 1, 2, \cdots, n)$ を最も近い mediod $\boldsymbol{x}_{med_l}$ に対応するクラスタ集合 C_l に所属させる．
6. 各クラスタの mediod のインデックスを式 (3) の通りに更新する．

$$med_l = \underset{\substack{i \\ \boldsymbol{x}_i \in C_l}}{\operatorname{argmin}} \sum_{\boldsymbol{y} \in C_l} d(\boldsymbol{x}_i, \boldsymbol{y}) \quad (3)$$

"Logdata Clustering Analysis for Dropout Prediction in Beginner Programming Class"

†Genki HASHIMOTO・Advanced course of Control and Information Engineering, National Institute of Technology, Kisarazu College
‡Shinichi OEDA・Department of Information and Computer Engineering, National Institute of Technology, Kisarazu College

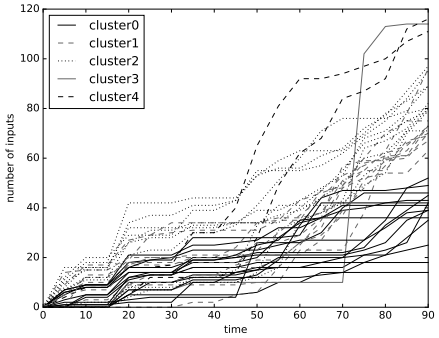

図 1　時系列クラスタリング結果.

7. クラスタ集合に変化がなくなるまで手順 (5), (6) を繰り返す.

3. 計算機実験

3.1 データセット

プログラミング初学者 39 名を対象に行われた 90 分の授業，試験中のコマンド入力履歴を使用した.

3.2 結果と考察

代表として，ある授業のクラスタリング結果を図 1 に示す．クラスタ番号はクラスタの medoid の授業終了時の入力数が小さいものから 0-4 としている.

クラスタの傾向としては基本的に入力量によって分類されていることがわかる．例外として，後半に急激に入力数が増えたユーザは単独でクラスタを形成している．データ数が全体の 10% 未満のクラスタを外れクラスタとしたとき，外れクラスタの傾向には (1) 他のクラスタよりにコマンド入力数が少ない (2) 他のクラスタより入力数が多い (図 1 の cluster4) (3) 短時間で急激に入力が増えている (図 1 の cluster3) の 3 通りが存在した.

クラスタリング結果と，授業担当教員の主観による学生の 4 段階評価 (A〜D) および関心意欲などの特徴を用いて，外れクラスタに所属する学生の特徴を確認する.

特徴 (1) には教員評価が D の学生が所属している．このことから明らかに授業内容が分からない学生，課題に取り組むことを放棄している重度の要支援者といえる.

特徴 (2) には A〜C 評価の学生，特に A と C の学生が多く所属している．考えられる可能性としてはプログラムがうまく動かずに，デバック作業で多くコマンドを入力している，授業課題をすぐに終え，個人的な学習を進めている等があげられる.

特徴 (3) は 1 回の授業と試験で各 1 名ずつ外れクラスタに所属した．評価はバラけたものの関心意欲が低い学生である点が共通している．試験においては筆記とプログラム作成の 2 種の問題があったため，筆記に時間を多く割いたことが考えられる．授業中においては途中まで寝ていて起きてから課題を進めたことなどが予想できる.

ここから外れクラスタに所属した学生はその授業において全体よりも理解が遅れているものか，非常に速い学生が所属していると予測される.

4. まとめ

本研究の目的はプログラミング初学者のログから要支援者検知を行うことである．ログデータは授業内容に大きく依存し，評価機を作成することが難しいため，クラスタリングにより，集団内の外れを検出することを考えた．クラスタリングによって，外れクラスタには 3 つの傾向があることがわかったそれぞれの原因が予想できた．よって外れクラスタに所属した学生についてのみ詳細にログデータ等から状態の把握を行い，必要な学生に支援を行うことでより高い授業効果が期待される.

今後の課題として，学生の自己評価アンケートを用いた外れクラスタの意味付けの検証，最適なクラスタ数の検討，他手法での解析があげられる.

謝辞　本研究は JSPS 科研費 16K01095 の助成を受けたものです.

参考文献

[1] 朽木 拓, 山田 敬三, 佐々木 敦 “プログラミングスキルレベル評価手法の研究”, 情報処理学会全国大会講演論文集, Vol.72, pp.521-522, 2010.
[2] 伊藤 暁人, “ニューラルネットワークによる学生の成績予測とその学習指導への適用可能性の検討”, 平成 20 年度名古屋工業大学卒業研究論文, 2010.
[3] 橋本 玄基, 清野 真理子, 大枝 真一. “プログラマのスキル評価のためのログデータ解析”, 情報処理学会全国大会講演論文集, Vol.78, pp.899-900, 2016.
[4] David Arthur,Sergei Vassilvitskii, “k-means++: The Advantages of Careful Seeding”, SODA ’07 Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, pp.1027-1035, 2007.
[5] Christopher Michael Bishop, “Pattern Recognition and Machine Learning”, Springer, pp.424-428, 2006.