Chen, W. et al. (Eds.) (2016). Proceedings of the 24th International Conference on Computers in Education. India: Asia-Pacific Society for Computers in Education

Behaviors and Features Selection of Online Learning Data

Bo JIANG*, Ming GAO
Center for Educational Big Data, College of Education Science and Technology, Zhejiang University of Technology, China
*bjiang@zjut.edu.cn

Abstract: To identify any patterns in learning behaviors, learners' learning behavior data are captured and stored in many online learning platforms. It is crucial to determine which behaviors and which features of a behavior are most related to the specific analytics task. To do this, feature selection method has often been applied to determine a global reduced feature space. However, little attention has been paid to select the behaviors and features within behavior simultaneously. In this work, we propose a two-level feature selection method which can determine the importance of behaviors and features simultaneously. The proposed method is embedded into the classical *k*-means to cluster a famous e-learning dataset. Our experimental results show that the proposed method is an effective way to improve the clustering performance significantly.

Keywords: learning analytics; feature selection; behavior selection; multi-view clustering

1. Introduction

In the past decade, more and more students' learning behavior data were generated and collected from several kinds of tutoring systems and online learning courses(Baker, 2014; Romero & Ventura, 2013). With these learning behavior data, we may find how students learn, what kind of learning materials are more attractive, and even, why some students achieve high academic performance but others fail. The amount of behavior feature captured by log systems, such as keyboard and mouse actions within a behavior, behavior sequence, behavior duration and behavior frequency, make the learning behavior data highly dimensional, which bring challenge for pattern discovery(Jong, Y., & Wu, 2007) (Perera, Kay, Koprinska, Yacef, & Zaiane, 2009).

Most learning behavior data have several behaviors and each of them contains a number of features, which means there are multiple feature groups in the feature space. Figure 1 shows an example of two-level feature group structure of learning behavior data. On one hand, we often want to know which learning behavior(s) dedicate more than others to learning outputs(Ramesh, Goldwasser, Huang, Daumé III, & Getoor, 2013, December). On the other hand, within each learning behavior, it is also crucial to recognize which features are more important than others for a specific analytics task, which provides valuable feedback for us to optimize the learning system(Riley, Miller, Soh, Samal, & Nugent, 2009,July). To the best of our knowledge, there existed several works dedicated to solve the two issues independently, but few works were involved in how to solve them simultaneously. For instance, (Minaei-Bidgoli & Punch, 2003, July) and (Romero, Espejo, Zafra, Romero, & Ventura, 2013) used the behavior's total number to predict student performance; (Perera et al., 2009) investigated students' three behaviors in an online learning environment and used them to group students.

In this work, a simple yet efficient feature selection method was developed to select feature within each learning behavior. We conducted our feature selection method in students clustering task, which cluster students using their learning behavior data recorded by system. In the proposed method, each behavior and each feature belonging to it are assigned a positive weight to express their importance to the clustering result. A weight adaptation strategy was developed to update the behavior and feature weights automatically. The proposed method was tested on a famous learning behavior dataset Digital Electronics Education and Design Suite (Deeds), which was developed by University of Genoa(Vahdat, Oneto, Anguita, Funk, & Rauterberg, 2015). "Deeds" is a learning environment for digital electronics courses, which records several kinds of behaviors and features of each behavior.

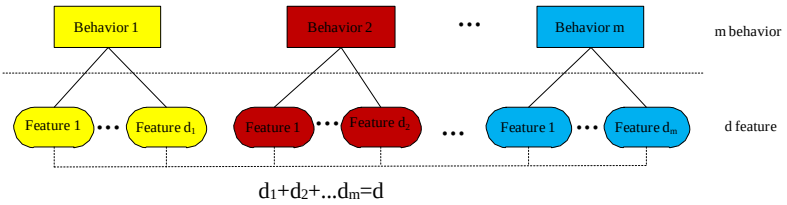

Figure 1. Two-level feature group structure of learning behavior data

The rest of the paper is organized as follows. In Section 2, the proposed behavior and feature selection method is given. The description and pre-processing of Deeds dataset are described in Section 3. The experimental results are reported in Section 4. Finally, the conclusions of this study are given in Section 5.

2. The proposed feature selection method

In the proposed method, each behavior and feature is assigned a positive weight to express its importance. Different with principal component analytics that is independent with the data mining task, the proposed method is closely related to the specific analytics task because the iterative outputs of task are used to adjust the weight during the feature selection process. In this work, we aim to select the relevant behavior and features in the clustering task. The classical *k*-means algorithm is used to cluster the students based on their learning behavior on Deeds.

Consider the learning behavior dataset $\mathbf{X}$ with N students, H behaviors and D features: $\mathbf{X} = \{X_i\}_{i=1}^{N}$ and $X_i = \{x_i^{(h)}\}_{h=1}^{H}$, where $x_i^{(h)} \in R^{d^{(h)}}$ and $\sum_{h=1} d^{(h)} = D$. The D features are divided into H views $\{T_h\}^H$. The *k*-means algorithm optimized the following Euclidean-based loss function:

$$J(Z,U) = \sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{D} u_{ki}^{kj}(z_{kj} - x_{ij})^2 \quad (1)$$

where $U = [u_{ki}]_{K\times N}$ is the hard partition matrix, $Z = [z_{kj}]_{K\times D}$ is the cluster centroids. To discriminate feature and behavior, a feature weight w and behavior weight v are assigned to each feature and behavior of the entire objects, respectively. The optimization model of this work is defined as follow.

$$\min_{\alpha,\beta} J(Z,W,V,U) = \sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{h=1}^{H}\sum_{j\in T_h} u_{ki}^{kj} w_{kj}^{\alpha} v_h^{\beta}(z_{kj} - x_{ij})^2 \quad (2)$$

$$s.t.\begin{cases} \sum_{k=1}^{K} u_{ki} = 1 & u_{ki} \in \{0,1\}, i = N_{1,K} \\ \sum_{j\in T_h} w_j = 1 & w_j \in [0,1], j = N_{1,D}, h = N_{1,H} \\ \sum_{h=1}^{H} v_h = 1 & v_h \in [0,1] \end{cases} \quad (3)$$

where $W = [w_j]_{1\times D}$ is the feature weight and $V = [v_h]_{1\times H}$ is the behavior weight. α and β, used to control the weight distributions, are two fuzzy exponents that need to be provided by user. To solve this

optimization problem, an iterative algorithm that alternates between updating the clusters and computing the two weight vectors is given. The major steps of the proposed algorithm and the more details about how to determine parameter α and β, you can find them in (Jiang, Qiu, & Wang, 2016).

3. Deeds dataset

Here we used the Digital Electronics Education and Design Suite (Deeds) dataset to evaluate the proposed method. “Deeds” is a learning environment for digital electronics course, which records several kinds of behaviors and features of each behavior. The “Deeds” system generates a high volume of learning behavior data, which include students' time series of behavior during six sessions of laboratory sessions of the course of digital electronics(Vahdat et al., 2015). In this work, only the behavior data in Session 1 is used, which contains students’ behaviors data and the exam scores. Because some students attended the class but did not attend the session test, and some students have grades, however they have no record of this session. Therefore, we only contain the 64 records of students who have both the activity’s features and grades. According to the passing grade 60, the students are divided into two groups: successful group and failed group. The aim of our experiments is to determine which behaviors and features play important role in student grouping.

The original dataset contains 15 behaviors, but it regards the same activity on different objects as different behaviors. For example, exercise study and related material study are two independent behaviors. The same behaviors on different objects are merged in our experiment. Nine behaviors are rested as follows: (1)*Aulaweb:* Students are using “Aulaweb” as a learning management system (based on Moodle) which is used for the course. In “Aulaweb”, the students might access the exercises, download them, upload their work, check the forum news, etc. (2)*Blank:* When the title of a visited page is not recorded. (3)*Deeds:* All activities students executed on Deeds simulator. (4)*Diagram: U*sing “Simulation Timing Diagram” of “Deeds” simulator to test the timing simulation of the logic networks. It also contains these components: "Input Test Sequence" and "Timing Diagram View Manager ToolBar". (5)*FSM:* When the student is working on a specific exercise on 'Finite State Machine Simulator'. (6)*Other:* This includes, for majority of the cases, the student’s irrelevant activities to the course. (7)*Properties:* “Deeds” simulator, Simulation Timing diagram, and FSM contain the properties window, which allows to set all the required parameters of the component under construction. For instance, the Properties can contain: "Switch Input", "Push-Button", "Clock properties", "Output properties", "textbox properties". We label all as 'Properties'. (8)*Study:* It indicates that a student is studying / viewing the content of a specific exercise or material. (9)*TextEditor:* It shows that the student is using the text editor. At the same time, each behavior has six features: (1)*MW:* It shows the amount of mouse wheel during an activity. (2)*MWC:* It shows the number of mouse wheel clicks during an activity. (3)*MCL:* It shows the number of mouse left clicks during an activity. (4)*MCR:* It shows the number of mouse right clicks during an activity. (5)*MM:* It shows the distance covered by the mouse movements during an activity. (6) *KS:*It shows the number of keystrokes during an activity.

In addition, the normalization was also conducted on the dataset because we found that the feature values varied significantly in different features and even in the same feature in different behaviors. The huge difference on the feature values brought challenges for our feature selection tasks. Figure 2 shows the feature map before and after normalization. Obviously, most data mining methods favor the normalized feature space.

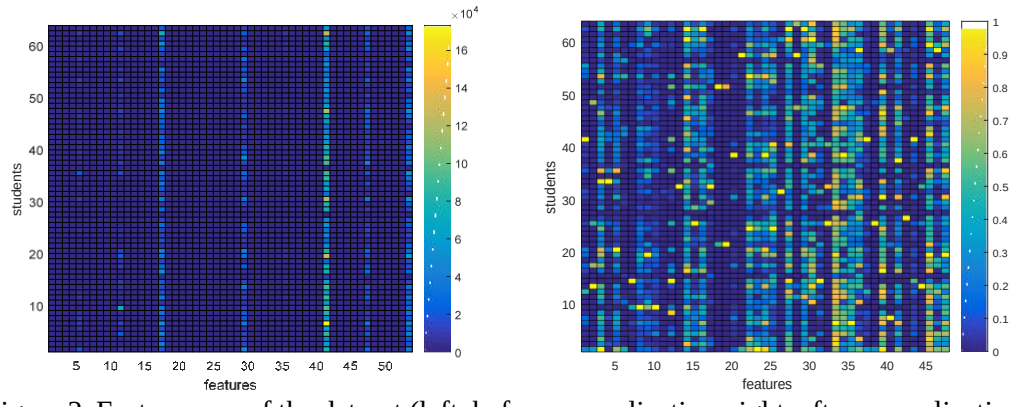

Figure 2. Feature map of the dataset.(left: before normalization, right: after normalization)

4. Experimental analytics

4.1 Behavior selection

The first experiment is to select the most important behavior(s) for our clustering task. In order to select behaviors, we assumed that in each behavior all features are equally important, so β is given a small value and α is assigned an enough large value. We set α and β in {1.25, 1.5, 2, 2.5, 5, 10, 20, 40, 80, 120} to find the best parameter combinations. After running the proposed algorithm 100 times on each different parameter combinations independently, we found the best combination is α=80 and β=2. The average clustering accuracy is only 52.52 and standard deviation is 2.03. The weight of each behavior generated by the proposed algorithm is given in Figure 3 (a). It is obviously that all the weights for behavior *Blank, FSM, Properties,* are almost close to zero in the 100 runs, which indicate that these three behaviors have almost no dedication to the clustering performance. On the contrary, the behavior *Aulaweb, Deeds, Diagram, Other, Study,* and *TextEditor* are the six important behaviors to our clustering task.

A natural problem arose is whether the three unimportant behaviors can be reduced from the dataset? To answer this question, we deleted all features in the three behaviors from the initial dataset and then run the proposed algorithm on it 100 times again. The third column in Table shows the clustering accuracy. Table 3 also shows the results of *t*-test at the confidence level of 5%. From Table 1, the *t*-test result (t=1.05<1.984) shows that there is no significant difference on clustering performance if we reduce the three behaviors from the initial dataset. Figure 3(b) shows the weights of the six retained behaviors. Although the weight of each behavior change a lot during the 100 runs, all behaviors' weights are between 0.1 and 0.2, which indicates that they are equally importance to the clustering task.

4.2 Feature selection

In the second experiment, we aim to select the most relevant features in each retained behaviors. To do this, parameter α was given a small value and β was assigned an enough large value. The best parameter combination α=1.25 and β=80 was found using the same parameter selection method used in behavior selection experiment. Figure 4 shows the feature weight distribution within each of the six behaviors. We can see that the feature weight distributions in each behavior have significant difference. For example, the *MCR(Mouse_click_right)* feature is very important in *Aulaweb* behavior, but that's not the case in other behaviors. In *Aulaweb* behavior, MW and MCR are more important than others. In *Deeds* behavior, MW, MWC and MM are crucial. In *Diagram* behavior, MWC is the most important feature. In *Other* behavior, MW and MWC play important roles. In *Study* and *TextEditor* behavior, MW, MWC and KS are more crucial than others.

Table 2 provides the average value and standard deviation of the feature weight in each behavior during the 100 runs. If a feature's average weight is lower than 0.01, we call this feature a *dispensable* feature, otherwise, it's *important*. All values in bold in Table 2 are dispensable. From Table 4, *MCL* (*Mouse_click_left)* is dispensable for all behaviors, which means that there is no necessity to record the mouse left click action in the platform. On the contrary, the *MWC (Mouse_wheel_click)* feature is important for all kinds of behaviors. It is also found that *MCR* (*Mouse_click_right)* is dispensable for behaviors except Aluweb, *MM (Mouse_movement)* is importance for *Aluweb* and *Deeds* behavior, and *MW (Mouse_wheel)* is important for all behaviors except *Diagram*.

For the feature selection problem, we also need to answer the question: are these dispensable features not important that whether they can be deleted from the dataset? To answer this question, we discarded all the 18 dispensable features from the dataset and ran the proposed algorithms 100 times again. The last two column of Table 3 provides the clustering accuracy and *t*-test result before and after the feature reduction. Surprisingly, the average clustering accuracy was enhanced to 61.38% and the maximal accuracy exceeded 65%. The t-test result (t=9.59) also showed that reducing these dispensable features can improve the clustering performance significantly.

The experimental results show that the proposed method can select the behaviors and features successfully. After behaviors and feature selection, the dimension of the initial dataset decreased from 54 to only 18, which not only makes the data analytics more easily but also improve the clustering performance significantly. The results indicate that the proposed method is effective.

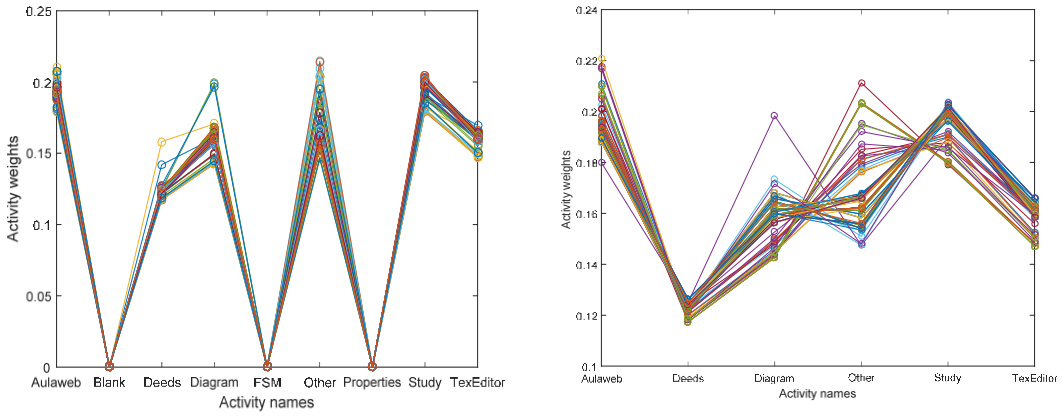

Figure 3. The behavior (activity) weight provided by the proposed algorithm. (a) weights distribution on the initial dataset with 9 behaviors; (b) weights distribution on the reduced dataset with 6 behaviors

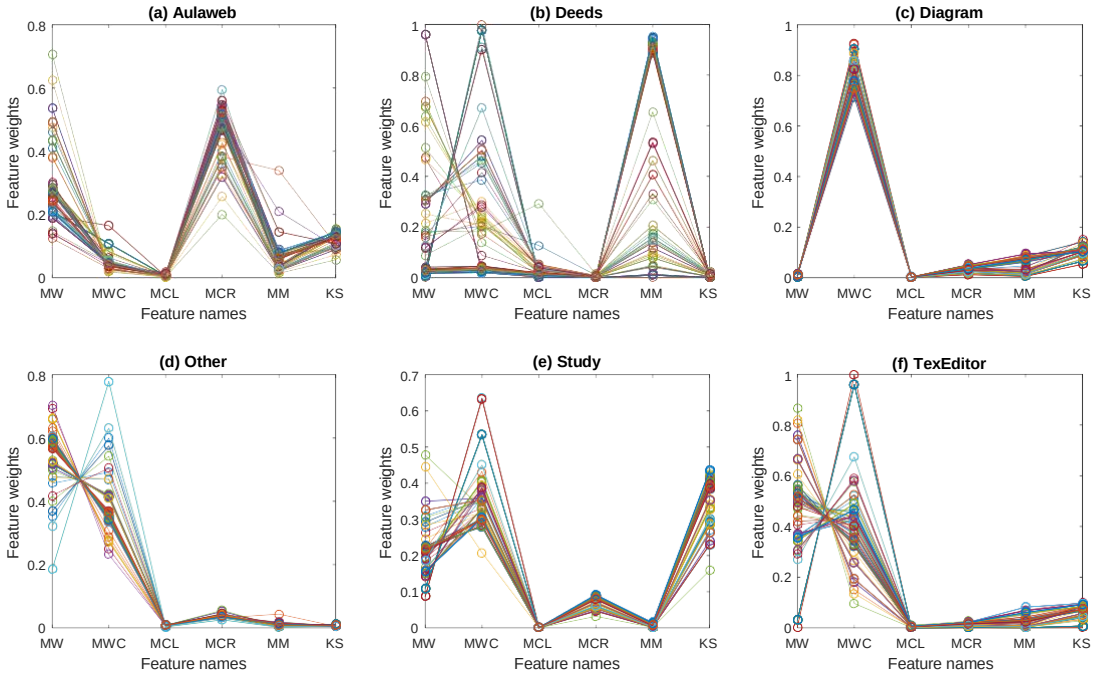

Figure 4. Feature weights distribution in the six selected behaviors

Table 1: Clustering performance of the proposed algorithm on Deeds with different behaviors and features

	Initial behaviors	Reduced behaviors	Initial features	Reduced features
Mean of accuracy	52.52%	52.73%	56.67%	61.38%
Std. of accuracy	2.03%	2.00%	4.91%	5.93%
t test results	t = 1.05		t = 9.59	

Table 2: The average value and standard deviation (in bracket) of feature weight in each behavior.

	MW	MWC	MCL	MCR	MM	KS
Aluweb	0.275(0.106)	0.054(0.027)	**0.007(0.002)**	0.479(0.068)	0.059(0.041)	0.127(0.019)
Deeds	0.185(0.274)	0.271(0.336)	**0.022(0.032)**	**0.003(0.003)**	0.512(0.407)	**0.007(0.007)**
Diagram	**0.009(0.003)**	0.807(0.058)	**0.003(0.001)**	**0.033(0.010)**	**0.045(0.031)**	0.102(0.021)
Other	0.548(0.086)	0.390(0.090)	**0.007(0.001)**	**0.038(0.004)**	**0.008(0.005)**	**0.008(0.001)**
Study	0.197(0.072)	0.366(0.097)	**0.001(0.000)**	**0.074(0.015)**	**0.007(0.004)**	0.357(0.066)
TextEditor	0.438(0.202)	0.457(0.217)	**0.004(0.002)**	**0.015(0.007)**	**0.024(0.024)**	0.062(0.027)

5. Discussion

Despite significant progress of educational data mining achieved in recent years, the topic of learning behavior analytics is still very challenging, due to the complicated structure of learning behavior. One of the challenges for learning behavior analytics is how to select the most relevant behaviors and features for the specific analytics task. This paper proposes two-level feature selection framework to the construction of effective feature space for online learning data. Experimental results show that the proposed method can discriminate the behaviors and features simultaneously. More importantly, the results also show that the reduction on behaviors and features wouldn't degrade performance, instead it improves clustering accuracy significantly.

We note some limitations of the current study that highlight opportunities for future method improvement. Firstly, the feature selection results and clustering results are sensitive to the two parameters α and β which need to run the algorithm several times to determine the best parameter combination. Secondly, the proposed *k*-means based algorithm failed to preform very well on the Deeds dataset. This may be caused by the sparsity in the feature space, because we found that there are large amounts of zero values in the feature space. And there is no evidence supporting that there must have a direct relationship between these learning behaviors and the test performance. Lastly, although we found that some behaviors and features are not important for our analytics, we fail to discover the reason behind it.

In future work we plan to continue to investigate the application of the proposed method to other sessions of the Deeds dataset. We also plan to combine the proposed method with the classification task to predict the learning performance. Further, we aim to develop a sparse feature selection method using regularization terms and evaluate the methodology on other learning datasets.

Acknowledgements

This work is partly supported by the National Natural Science Foundation of China under Grant No. 61503340 and Zhejiang Provincial Natural Science Foundation of under Grant No. LQ16F030008.

References

Baker, R. S. (2014). Educational Data Mining: An Advance for Intelligent Systems in Education. *IEEE Intelligent Systems, 29*(3), 78-82.

Jiang, B., Qiu, F., & Wang, L. (2016). Multi-view clustering via simultaneous weighting on views and features. *Applied Soft Computing, 47*, 304-315.

Jong, B. S., Y., T., & Wu, Y. L. (2007). Learning log explorer in e-learning diagnosis. *IEEE Transactions on education, 50*(3), 216-228.

Minaei-Bidgoli, B., & Punch, W. F. (2003, July). *Using genetic algorithms for data mining optimization in an educational web-based system.* In *Genetic and evolutionary computation conference* (pp. 2252-2263). Springer Berlin Heidelberg.

Perera, D., Kay, J., Koprinska, I., Yacef, K., & Zaiane, O. R. (2009). Clustering and sequential pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering, 21*(6), 759-772.

Ramesh, A., Goldwasser, D., Huang, B., Daumé III, H., & Getoor, L. (2013, December). *Modeling learner engagement in MOOCs using probabilistic soft logic.* In *NIPS Workshop on Data Driven Education* (Vol. 21, p. 62).

Riley, S. A., Miller, L. D., Soh, L. K., Samal, A., & Nugent, G. (2009,July). Intelligent Learning Object Guide (iLOG): A Framework for Automatic Empirically-Based Metadata Generation. *Artificial Intelligence in Education, 200*, 515-522.

Romero, C., Espejo, P. G., Zafra, A., Romero, J. R., & Ventura, S. (2013). Web usage mining for predicting final marks of students that use Moodle courses. *Computer Applications in Engineering Education, 21*(1), 135-146.

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3*(1), 12-27.

Vahdat, M., Oneto, L., Anguita, D., Funk, M., & Rauterberg, M. (2015). *A Learning Analytics Approach to Correlate the Academic Achievements of Students with Interaction Data from an Educational Simulator.* In *Design for Teaching and Learning in a Networked World* (pp. 352-366). Springer International Publishing.