 CMC, vol.59, no.2, pp.625-633, 2019

A Multi-Feature Weighting Based K-Means Algorithm for MOOC Learner Classification

Yuqing Yang[1, 2], Dequn Zhou[1, *] and Xiaojiang Yang[1, 3, 4]

Abstract: Massive open online courses (MOOC) have recently gained worldwide attention in the field of education. The manner of MOOC provides a new option for learning various kinds of knowledge. A mass of data miming algorithms have been proposed to analyze the learner's characteristics and classify the learners into different groups. However, most current algorithms mainly focus on the final grade of the learners, which may result in an improper classification. To overcome the shortages of the existing algorithms, a novel multi-feature weighting based K-means (MFWK-means) algorithm is proposed in this paper. Correlations between the widely used feature grade and other features are first investigated, and then the learners are classified based on their grades and weighted features with the proposed MFWK-means algorithm. Experimental results with the Canvas Network Person-Course (CNPC) dataset demonstrate the effectiveness of our method. Moreover, a comparison between the new MFWK-means and the traditional K-means clustering algorithm is implemented to show the superiority of the proposed method.

Keywords: Multi-feature weighting, learner classification, MOOC, clustering.

1 Introduction

The development of massive open online courses (MOOC) has been recognized as one of the most significant innovations in the field of education [Jacoby (2014)]. It provides new courses at an unprecedented scale, both in terms of learner numbers and in terms of global reach [Pursel, Zhang, Jablokow et al. (2016)]. Many data mining techniques have been proposed to group learners based on their learning style, approach, profile, prior knowledge, and so on [Shahir and Husain (2015); Wang, Yang, Wen et al. (2015); Papamitsiou and Economides (2014); Romero and Ventura (2017)].

[1] College of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing, 211106, China.

[2] Office of International Cooperation and Exchanges, Nanjing University of Finance & Economics, Nanjing, 210046, China.

[3] Jiangsu Guidgine Educational Evaluation Inc., Nanjing, 210046, China.

[4] International Education Office of Centennial College, Toronto, P.O. Box 631, Canada.

[*] Corresponding Author: Dequn Zhou. Email: dqzhou@nuaa.edu.cn.

Clustering techniques are the most popular techniques to group learners with similar categories allowing formulation of appropriate learning strategies for each group of learners [Dutt, Ismail and Herawan (2017); Cabedo, Tovar and Castro (2016); He, Ouyang, Wang et al. (2018); Zhang, Zheng and Xia (2018)]. Using cluster analysis as a technical means can effectively identify and characterize the underlying features of MOOC learners [Cabedo, Tovar and Castro (2016)]. Wang et al. [Wang and Fu (2018)] exploited the data mining tools to analyze learners' behavior characteristics and then classify the learners into different groups. In [Gallén et al. [Gallén and Caro (2017)], a set of 26 questions was designed to investigate the learners' motivation to study with MOOC, where the answer options of the questions were treated as cluster characteristic indexes. Yousef et al. [Yousef, Chatti, Wosnitza, et al. (2015)] adopted cluster analysis to analyze the different goals of users and establish a deeper understanding of their behavior. Gadhavi et al. [Gadhavi and Patel (2017)] utilized the data mining technology to group the MOOC learners and predict their final grade. Prabhakar et al. [Prabhakar and Zaiane (2017)] utilized a modified Particle Swarm Optimization technique to group the MOOC learners based on their grades and personal information, where the intra-group heterogeneity and inter-group homogeneity are both included to enhance the classification results. Harwati et al. [Harwati, Alfiani and Wulandari (2015)] exploited the k-means clustering algorithm to reveal the hidden pattern and classify students mainly based on their grade.

As analyzed above, most current methods exploit the learner's final grade to judge and classify them. Note that many factors will influence the learner's final grade in practice, thus it is difficult to obtain a comprehensive view of the state of the learner's performance and simultaneously classify them into proper groups with the single feature. To address this challenge, we design a novel multi-feature weighting based K-means (MFWK-means) algorithm. Correlations between the grade and other features are first investigated, and then the learners are classified based on their grades and weighted features with the proposed MFWK-means. Experimental results with the Canvas Network Person-Course (CNPC) dataset demonstrate the effectiveness and superiority of our method.

2 The proposed MFWK-means clustering algorithm

2.1 Correlation analysis between the grade and other features

In this paper, we classify the MOOC learners into different categories based on their final grades and other features, such as learning hours, interactions with the course, and so on. In practice, these features are not independent and they may influence the final grade of a MOOC learner. In this part, we first analyze the correlations between them. For a more clear explanation, a widely used MOOC dataset-CNPC dataset is adopted here and in the subsequent experimental parts. This dataset is collected from the Canvas Network open courses (running January 2014-September 2015). These data include over 325000 aggregate records, and each record represents one learner's activity with 26 different features, including course ID, discipline, user ID, and so on. Among these features, some are related to the course information, and others are related to the learners study information. In this paper, we focus on the relationship between the final grade and the

features corresponding to the learners. Thus, four features are selected for the analysis, including "completed", "nevents", "ndays" and "nforum", where the meanings of these features are described in Tab. 1. Note that in the original CNPC dataset, some records of the features are missing. After removing the invalid records, the total number of the records in our experiment is 5280. First, we analyze the correlations between the final grade and the four features by drawing a scatter plot of the feature values versus the grades, where the results are shown in Fig. 1.

Table 1: Feature attributes of the CNPC dataset

Features	Descriptions
completed	percentage of the completion for the homework with the course
nevents	count of distinct interactions with the course
ndays	count of distinct days with one or more events
nforum	number of posts total in discussion forums throughout the course

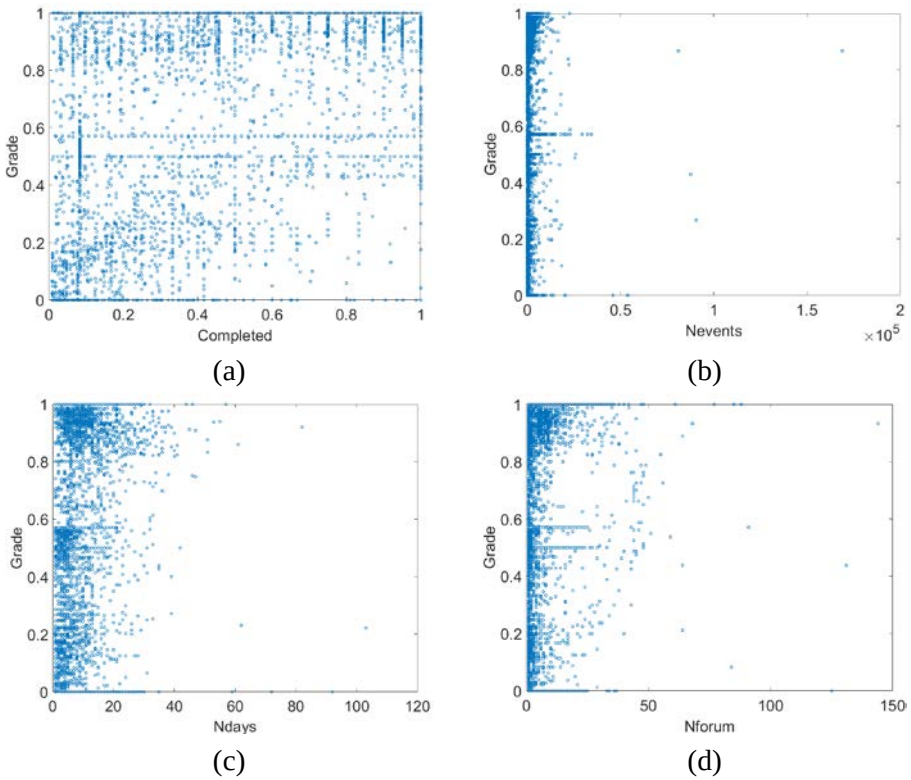

Figure 1: Scatter plots of the feature values versus the grades. (a) Completed versus grade, (b) nevents versus grade, (c) ndays versus grade, and (d) nforum versus grade

From the scatter plots in Fig. 1, we can summarize the following observations: (1) the feature "completed" greatly influence the learner's final grade. In spite of some learners can obtain higher grade with doing less homework, but as a general trend the grade increases with a larger "complete" value. (2) The feature "nevents" slightly influence the learner's final grade. From Fig. 1(b) we can see that with the increase of the "nevents" value, the learner's grade increase slightly. (3) The features "ndays" and "nforum" have the similar degrees of impact to the final grade. As shown in Figs. 1(c) and 1(d), when the values of "ndays" and "nforum" increase, the final grade increases analogously.

By using the scatter plots, we analyze the relationship between the grade and the other features roughly. In order to quantitatively evaluate the correlations between these features, we adopt the Pearson Correlation Coefficient (PCC) measure [Zou, Zeng, Cao et al. (2016)], which can be written as:

$$PCC_{X,Y} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - \left(\sum X\right)^2}\sqrt{n\sum Y^2 - \left(\sum Y\right)^2}} \tag{1}$$

where X and Y represent two vectors and n denotes the number of the variables in each vector. We calculate the PCCs between the grade and the other four features by using Eq. (1), where the results are recorded in Tab. 2. From the experimental results in Tab. 2 we can find that the feature "completed" is more relevant to the final grade, while the feature "nevents" is less relevant to the grade, and the features "ndays" and "nforum" obtain similar PCC values. The conclusion is consistent with the scatter plot analysis.

Table 2: Pearson Correlation Coefficient between the grade and other four features

PCC	completed	nevents	ndays	nforum
grade	0.514	0.114	0.234	0.269

2.2 The multi-feature weighting based K-means algorithm

K-means is a widely used clustering algorithm, which partitions a data set into K clusters by minimizing the sum of squared distance in each cluster. In traditional methods, the MOOC learners are usually classified based on their final grade. The using of a single feature in clustering algorithm may limit the objectivity and comprehensiveness of the classification process. To cover the shortage of the traditional clustering manner, we propose a novel multi-feature weighting based K-means algorithm in this paper. Based on the correlation analysis between the grade and other features, the proposed MFWK-means clustering algorithm can be implemented with the following four steps:

MFWK-means clustering algorithm

Step 1: initialization

Randomly select K points as initial cluster centers.

Step 2: assignment

Calculate the multi-feature weighting distance between each data point and each cluster center based on Eqs. (2) and (3), and then assign each point to the closest cluster center.

Step 3: update

Calculate the mean value of the data points for each cluster and update the cluster center, and then repeat Step 2 and Step 3.
Step 4: convergence
Stop when there is no change of the cluster centers or reaching a predefined number of iterations
In the proposed MFWK-means clustering algorithm, the multi-feature weighting distance can be formulated as:

$$D = \sum_{k=1}^{K} \sum_{i=1}^{n} \left\| g_i(k) - c(k) \right\|^2 \tag{2}$$

where $c(k)$ represents the cluster center k ; $g_i(k)$ represents the i th multi-feature weighting data point in the cluster k, which is composed of the learner's final grade and other related features. In the proposed method, the multi-feature weighting data vector G can be defined as:

$$G = \frac{1}{T+1}\left(\sum_{t=1}^{T} w_t F_t + F_0 \right) \tag{3}$$

in which F_0 represents the value of the grade, and F_t denotes the utilized related features, and T is the number of the related features. In Eq. (3), the weight w_t is defined by measuring the correlation between the selected feature F_t and the final grade F_0. In our method, we use the PCC defined in Eq. (1) to calculate the weights.

3 Experimental results

Equations and mathematical expressions must be inserted into the main text. Two different types of styles can be used for equations and mathematical expressions. They are: in-line style, and display style. In order to verify the effectiveness of the proposed MFWK-means clustering algorithm, the widely used CNPC dataset is utilized in our experiment. First, the MFWK-means clustering algorithm is adopted to classify the MOOC learners in to different groups. Besides the feature "grade", another four features "completed", "nevents", "ndays", and "nforum" are also used in the proposed algorithm. The weight of each feature is calculated according to Eq. (1), and the MFWK-means algorithm is implemented based on the steps described in Section 2.2. Note that due to the various scales of the utilized features, we normalize each feature to the range $[0,1]$ based on the Eq. (4):

$$\hat{F} = \frac{\left(F - \min(F)\right)}{\left(\max(F) - \min(F)\right)} \tag{4}$$

where $\hat{F}$ is the normalized feature. $\min(F)$ and $\max(F)$ represent the minimum and maximum values of the feature F, respectively. To demonstrate the superiority of the proposed MFWK-means algorithm, the traditional K-means clustering algorithm is also applied for comparison. Classification results of K-means and our algorithm are shown in Figs. 2(a) and 2(b), respectively, where the group number is set as $K = 3$ for the two compared algorithms.

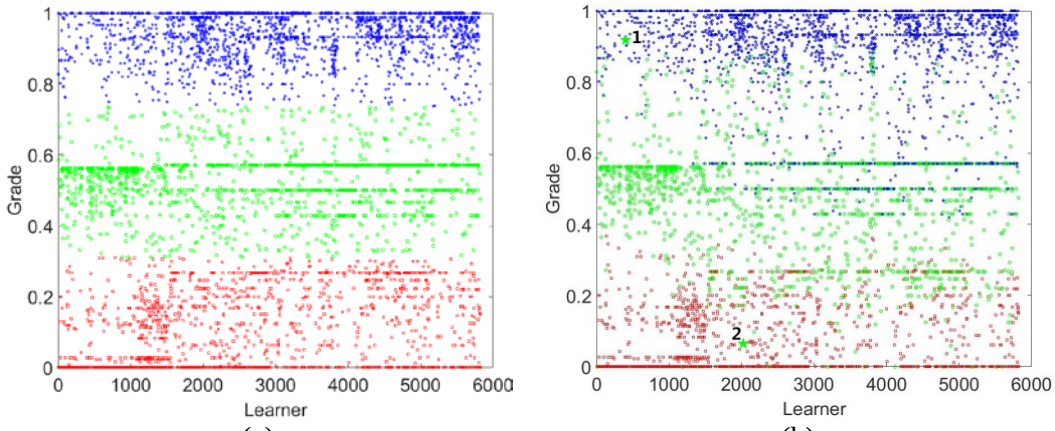

Figure 2: MOOC learner classification results with different clustering algorithms. (a) K-means, (b) MFWK-means

In Figs. 2(a) and 2(b), the blue points represent the "Good Learner" group, the green points corresponding to the "Ordinary Learner" group, and the red points denote the "Poor Learner" group. As shown in Fig. 2(a), the traditional K-means algorithm classifies learners strictly according to their final grade. From Fig. 2(b) we can see that the classification results of our algorithm is similar to Fig. 2(a) in general, but some learners are classified into different groups with the following 4 cases: (1) a learner is classified into a "Ordinary Learner" group with high grade; (2) a learner is classified into a "Good Learner" group with medium grade; (3) a learner is classified into a "Poor Learner" group with medium grade; (4) a learner is classified into a "Ordinary Learner" group with low grade. This is mainly because in the proposed MFWK-means clustering algorithm, we utilize various features besides the learner's grade, and meanwhile, each feature is assigned with a weight factor based on the correlation between the feature and the grade. To better analyze the classification results with the proposed MFWK-means algorithm, we choose two typical points in Fig. 2(b) for a detailed analysis. As shown in Fig. 2(b), the points 1 and 2 are denoted with green pentagon, in which point 1 is corresponding to the case (1), and point 2 is corresponding to the case (4). For a better explanation, we plot the scatter plots to show the various aspects of learner's conditions, which are shown in Figs. 3 and 4. From Fig. 3 we can observe that although the learner got a high grade (0.918), the values of the related features are extremely low compared to other learners. Taking into consideration of various aspects of the learner's study process, it is more proper to classify the learner into "Ordinary Learner" class. For the point 2 in Fig. 4, the opposite is happened. In spite of the learner obtained a low grade (0.042), the other aspects of this learner are excellent, thus the learner is also classified into "Ordinary Learner" class. As analyzed above, the proposed MFWK-means clustering algorithm can obtain a more comprehensive view of the state of the MOOC learners, and further result in a more correct classification.

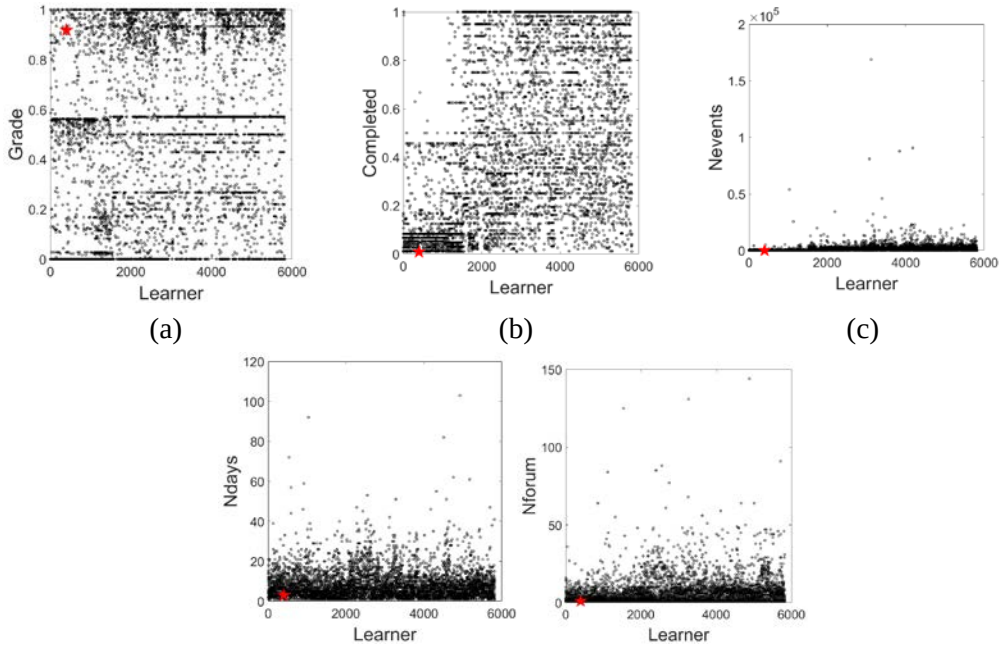

Figure 3: Various aspects of learner's conditions with point 1. (a) Completed, (b) nevents, (c) ndays, and (d) nforum

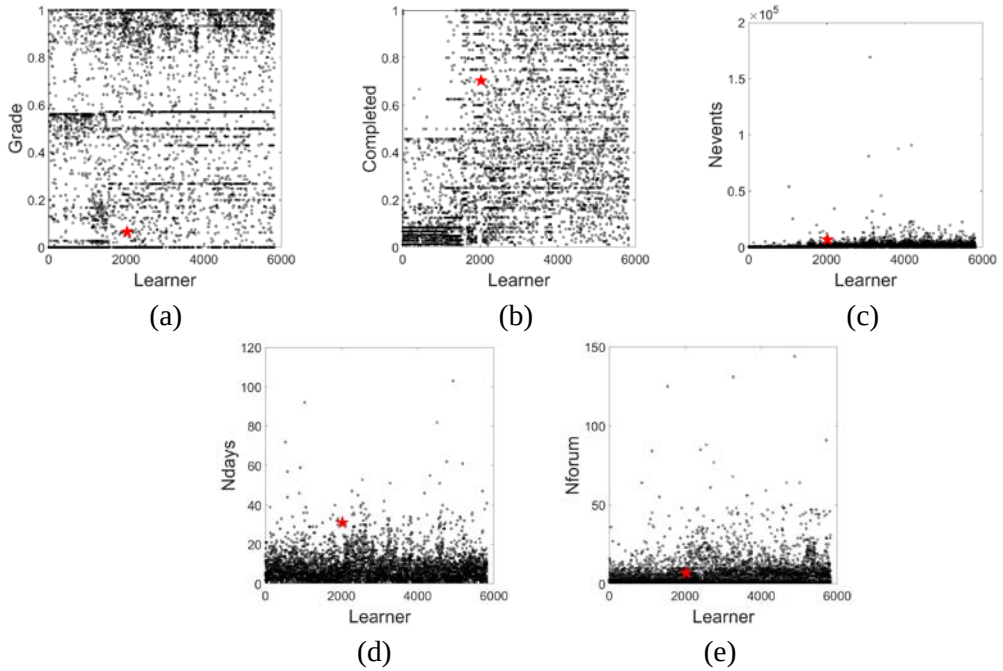

Figure 4: Various aspects of learner' conditions with point 2. (a) Completed, (b) nevents, (c) ndays, and (d) nforum

4 Conclusion

In this paper, we propose a novel multi-feature weighting based K-means algorithm to classify the MOOC learners into different groups. In order to comprehensively exploit the final grade and other various features of the learners, correlations between the grade and different features are first investigated. Then, the learners are classified based on their grades and weighted features with the proposed MFWK-means algorithm. Experimental results demonstrate the effectiveness and superiority of our method. In future works, more advanced data miming technologies can be investigated to analyze the learner's characteristics, such as deep learning networks, which may further improve the MOOC learner classification accuracy.

Acknowledgement: This work was supported in part by the Teaching Reform Foundation of Nanjing University of Finance & Economics and in part by Postgraduate Education Reform Project of Jiangsu Province.

References

Cabedo, R.; Tovar, E.; Castro, M. (2016): A benchmarking study of clustering techniques applied to a set of characteristics of MOOC participants. *ASEE'123rd Annual Conference*.

Dutt, A.; Ismail, M. A.; Herawan, T. (2017): A systematic review on educational data mining. *IEEE Access*, vol. 5, pp. 15991-16005.

Gadhavi, M.; Patel, C. (2017): Student final grade prediction based on linear regression. *Indian Journal of Computer Science and Engineering*, vol. 8, no. 3, pp. 274-279.

Gallén, R. C.; Caro, E. T. (2017): An exploratory analysis of why a person enrolls in a massive open online course within MOOC knowledge data collection. *Global Engineering Education Conference*, pp. 1600-1605.

Harwati; Alfiani, A. P.; Wulandari, F. A. (2015): Mapping student's performance based on data mining approach (a case study). *Agriculture and Agricultural Science Procedia*, vol. 3, pp. 173-177.

He, L.; Ouyang, D.; Wang, M.; Bai, H.; Yang, Q. et al. (2018): A method of identifying thunderstorm clouds in satellite cloud image based on clustering. *Computers, Materials & Continua*, vol. 57, no. 3, pp. 549-570.

Jacoby, J. (2014): The disruptive potential of the massive open online course: a literature review. *Journal of Open, Flexible, and Distance Learning*, vol. 18, no. 1, pp. 73-85.

Papamitsiou, Z.; Economides, A. A. (2014): Learning analytics and educational data mining in practice: a systematic literature review of empirical evidence. *Journal of Educational Technology & Society*, vol. 17, no. 4, pp. 49-64.

Prabhakar, S.; Zaiane, O. R. (2017): Learning group formation for massive open online courses (MOOCs). *International Conference on Educational Technologies*, pp. 129-136.

Pursel, B. K.; Zhang, L.; Jablokow, K. W.; Choi, G. W.; Velegol, D. (2016): Understanding MOOC students: motivations and behaviours indicative of MOOC completion. *Journal of Computer Assisted Learning*, vol. 32, no. 3, pp. 202-217.

Romero, C.; Ventura, S. (2017): Educational data science in massive open online courses. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 7, no. 1, pp. 1-12.

Shahiri, A. M.; Husain, W. (2015): A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, vol. 72, pp. 414-422.

Wang, G.; Fu, G. (2018): The cluster analysis of online learners' behavior characteristics from the perspective of data mining. *Modern Distance Education Research*, vol. 154, no. 4, pp. 106-112.

Wang, X.; Yang, D.; Wen, M.; Koedinger, K.; Rosé, C. P. (2015): Investigating how student's cognitive behavior in MOOC discussion forums affect learning gains. *Proceedings of the 8th International Conference on Educational Data Mining*, pp. 226-233.

Yousef, A. M. F.; Chatti, M. A.; Wosnitza, M.; Schroeder, U. (2015): A cluster analysis of MOOC stakeholder perspectives. *International Journal of Educational Technology in Higher Education*, vol. 12, no. 1, pp. 74-90.

Zhang, G.; Zheng, Y.; Xia, G. (2018): Domain adaptive collaborative representation based classification. *Multimedia Tools and Applications*, pp. 1-22.

Zou, Q.; Zeng, J.; Cao, L.; Ji, R. (2016): A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing*, vol. 173, pp. 346-354.