A computational investigation of learning behaviors in MOOCs

Sheng-Hua Zhong1 | Yanhong Li1 | Yan Liu2 | Zhiqiang Wang1

1 College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
2 Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China

Correspondence
Zhiqiang Wang, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China.
Email: wangzq@szu.edu.cn

Funding information
National Natural Science Foundation of China, Grant number: 61502311; Ministry of Education - Microsoft Cooperation in Production and Education Project, Grant number: 201601008033; Research and Practice Project of Computer Foundation Courses in Guangdong Province

Abstract
Massive open online courses (MOOCs) are the latest e-learning initiative to attain widespread popularity in the world. Thus, it is highly required to have a throughout analysis of learning in MOOCs, from theoretical to practical. Our primary goal is to take a detailed and comprehensive investigation into the learning behaviors in MOOCs, as well as to identify issues that have not yet to be adequately resolved. We employed commonly used educational data mining methodologies to analyze and interpret the behaviors in a computer science course based on the questionnaire survey data and daily activity data. We find most of the students could be divided into several groups that are coincident with their learning styles. Moreover, we can easily predict students’ learning styles based on their learning behaviors. This finding means the learning style could be a factor to indicate students’ learning behaviors, or even measure whether a student is appropriate to learn via MOOCs.

Key words
behavior study, computer science education, learning style, massive open online courses

1 INTRODUCTION

Throughout history, educators have always been intrigued by the potential of technology to help transform education and improve student learning [23]. As a concept, educational technology concerns an array of tools, such as media, machines, and networking hardware, as well as considering underlying theoretical perspectives for their active applications [43]. Educational technology can occur in or out of the classroom. It can be self-paced, asynchronous learning, or may be instructor-led synchronous learning. If the development of novel educational methods can provide the tools to help teachers to evaluate, analyze, and understand the educational procedure and results, and even the requirement of students, we could expect the educational process can be optimized [19]. Especially, based on the new teaching techniques, it is possible that we could collect and observe the students’ behaviors and have an understanding of their learning procedure. In recent years, many educators believe it is now a “changing education paradigm” [44]. In their discussions, this paradigm could be called as “Massive Open Online Courses (MOOCs),” has opened up a new revolution in the current era.

MOOCs offer researchers abundant fine-grained data collected on learners’ participation and mutual interactions that have never been available at such scales before [18]. The kinds of data available in learning environments have fuelled a proliferation of interdisciplinary research, bringing together computational and social scientists to collectively ask and answer questions about how learning happens in massive-scale online courses [18]. Unfortunately, handling large amounts of data manually is prohibitive. The information overload in massive-scale online courses requires the introduction and integration of new processing approaches into everyday objects and activities (“ubiquitous and pervasive computing”) [40]. The technique required for
analyzing educational big data must be able to support deeper analytics such as statistical analysis and data mining, scale to extreme data volumes; deliver faster response times driven by changes in behavior; and automate decisions based on analytical models [47]. Recently, educational data mining and analytics are able to discover useful knowledge or interesting patterns from the data coming from MOOC platform [6]. In existing work, detection, identification, and modeling of students’ learning behavior are primary research objectives. More specifically, these authors seek to identify learning strategies and when they occur, and model affective and metacognitive states [1,40]. Another orientation is the discovery and modeling of the respective behaviors within MOOCs, such as identifying meaningful patterns of participation, engagement, and disengagement in learning activities [29]. Some other existing works try to explore, identify, and evaluate various factors as indicators of performance for prediction purposes [40]. Several studies published in this area show that it is indeed possible to predict drop-off or performance in a degree or in a course with a reasonable accuracy [36]. For example, using logistic regression as a classifier, Jiang et al. [26] are able to predict the probability of students earning certificates for completion of the MOOC based on students’ first week assignment performance and social interaction within the MOOC.

Although there are studies indicating that learning styles play significant roles in influencing students’ learning behaviors and academic achievement [49], unfortunately, not much existing works try to explore online learning activities between students’ with different learning styles based on computational methods, in particular for the context of Chinese society. Our work seeks to investigate and understand learning processes and behaviors of students in MOOCs. In this paper, we analyze a MOOC on computer science course using several data mining techniques and provide some indications in terms of useful insights and guidance that could improve the education experience for both teachers and learners.

The remainder of this paper is organized as follows. In the next section, we will provide the literature review. The section following thereafter describes our methodology in this paper. After it, we provide the experimental results obtained by educational data mining and analytics. Finally, we will conclude this paper and outline the future work.

2 | THE LITERATURE REVIEW

MOOCs are seen as a major part of a larger innovation in higher education. Many universities scrambled to join in this new wave, and many MOOC providers were launched. Indeed, most of research works to date are based on the qualitative analysis of data on the leading MOOC platforms or providers [17,24,31,52]. Recently, there has been a growing body of research try to study students’ learning behaviors on MOOCs. Seaton et al. [45] collected valuable data on students’ learning behavior. They examined how the certificate earners allocated their time among the various course components and what fraction of each they accessed. Sharma et al. [46] presented the results of an eye-tracking study on a MOOC lecture showing the relation between gaze variables and students’ performance and learning strategy. Alario-Hoyos et al. [2] provided novel results regarding participants’ profiles and use of built-in and external social tools. Brinton et al. [10] investigated factors that are associated with the decline of activity on MOOC forums, and they used statistical analysis and general model to analyze the forum activity data. Dazo et al. [15] aimed to use grain student log data to understand student video viewing behaviors in computer science education.

There is very little existing work try to investigate the relationship between individual learning styles and learning behaviors on MOOCs. Sinha et al. [48] sought to do the video-watching clickstream analysis to study the watching styles. They applied a cognitive video watching model to explain the dynamic process of cognition involved in MOOC video clickstream interaction. Their model can be effectively used as an operationalization for making predictions regarding critical learner behavior.

Although the relationship between individual learning styles and learning behaviors on MOOCs has received very little attention in the research literature, in traditional online education, some studies have investigated students’ learning styles and there corresponding learning behaviors. Recent studies regarding learning styles in traditional online courses have consistently found that a preference for an active learning style has a positive effect on student’s performance in online learning [34]. Bozionelos [7] found that online students preferring the active learning style performed better on learning tasks, and felt more comfortable than their counterparts with other learning styles. Buerk et al. [11] found that online students tended to have the Convergent learning style. The study of Halsne and Gatta [21] tried to compare the learning styles of students who enrolled in an online course with the learning styles of comparable students who are taking the same course on-campus. They found that online students were predominantly visual learners while the traditional learners were primarily auditory or kinesthetic learners. Downing and Chim [16] found that students with the Reflector learning style tended to be Extraverts in the online learning. Jordanov [28] found that active learning styles had a positive impact on students’ attitudes toward online learning and could further improve the performance on computer-based learning tasks.
As described before, most of these existing studies have indicated that the learning style is a very important factor in learning, especially in online learning. Some researchers have consistently advocated that it will be helpful to design and adapt instruction to meet students’ learning styles in both course design and delivery. However, a majority of those studies are descriptive survey studies [34]. In order to find the direct and precise relationship between learning behaviors and learner characteristics, experimental studies are certainly necessary. Moreover, due to the large and complex behavior data on MOOCs that need to be analyzed, it is better to investigate learning behaviors using educational data mining methods.

The present study was designed to study the experiences of students with MOOCs’ learning based on educational data mining methods. The purpose of our research is not trying to measure the teaching performance in this individual course, or evaluate the learning achievement of each student, but rather the target is on the practical application of MOOCs’ education in Chinese society. Specifically, in this paper, we try to study the related behaviors of students on MOOCs, including registered students’ motivations, students’ daily activities, students’ specific learning behaviors, as well as their attitudes toward this learning new paradigm before and after their participation. Moreover, this paper seeks to investigate the relationship between students’ learning styles and learning behaviors in MOOCs via computational methods.

3 | METHODOLOGY

Our work seeks to investigate and understand learning processes and behaviors of students in MOOCs. Figure 1 shows the overview of the proposed framework. In this section, we will introduce each part of our methodology.

3.1 | Research questions

We ask the following research questions in this work:

Research question 1: MOOC education attracts many advocates and skeptics. Based on its natural advantage in being opened to any person with an Internet connection, proponents of MOOC argue that it can increase the supply of education services and improve the prestige of a higher education institution. Criticism of MOOC focuses on its low completion rate and loose structure. What are the advantages and challenges of learning in a MOOC on computer science course for the context of Chinese society?

Research question 2: Do the learning styles have a direct influence on students’ learning behaviors?

Research question 3: Whether the students’ learning behaviors could be automatically divided into several groups that are coincident with the learning styles?

Research question 4: Is it possible to predict students’ learning styles based on their own behaviors with a reasonable accuracy?

3.2 | Participants and course

This study is based on a MOOC course, Foundation of Computer Science (FOC), offered by Shenzhen University (SZU) on online learning platform in 2014. This online course is structured similarly to a traditional course that all contents are predefined by course instructors [39,53]. Shenzhen University was founded as a public university in 1983 and has undergone rapid growth and expansion in past 3 decades just like the city of Shenzhen, China’s most successful Special Economic Zone. As one of the fastest-growing local
Our investigation is based on two different kinds of data: the quizzes and the final online examination. The second is the activities in the discussion forum, including assignments, including exercises within the videos, course, we used two types of evaluation metrics. The first is the lecturers, other students, or teaching assistants. In this course, students can ask and answer questions proposed by the corresponding chapters. In the discussion forum of this course, we used two types of evaluation metrics. The first is the assignments, including exercises within the videos, quizzes in each chapter, and the final online examination. The second is the activities in the discussion forum, including the number of posts, votes, and views. The final grade gives priority to the quizzes and the final online examination.

3.3 Instruments

3.3.1 Data from MOOCs

Our investigation is based on two different kinds of data:

(1) Questionnaire survey data. A questionnaire survey was conducted that studied attitudes and feeling of the students toward learning via MOOCs. We asked students to supply details about their learning process, such as their educational backgrounds, learning objectives, attitudes, schedule, and styles. They were asked to evaluate strong and weak points of MOOCs, and gave their preferred ways of constructing knowledge. The questionnaire survey data contained 183 undergraduate students from Shenzhen University, and the average age of them was 19.8 years old. Seventy-six percent of students reported their gender were male. The major of these students includes architecture, civil engineering, computer science, optoelectronics engineering, and so on.

(2) Activity data. Our online learning provider recorded the daily behavior data of all learners signed up for this course. For each student, the number of visits, the number of posts, the number of votes, the number of reviews for each video, the grades for each assignment, quiz, and the grade for the final online exam. In our paper, we recorded, collected and statistically analyzed the daily activities from all learners in this course. Totally, 1,783 participants were registered for this course. Thus, our daily activity data included all the activities on our online learning platform.

3.3.2 Educational data mining model

As we know, the development of computational methods that can automate processes to analyze the activities of learners and to guide them toward a more effective learning experience [38]. In this study, we try to investigate the human behaviors on MOOCs based on three commonly used educational data mining models:

(1) Clustering model: k-means clustering. The goal of clustering is to naturally group data points together and split the whole data set into a set of clusters. k-means clustering algorithm [35] is one of the simplest unsupervised learning methods for clustering problem. The procedure follows a simple way to cluster a given data set through a center number of groups (assume k clusters) fixed apriori.

(2) Classification model: Support vector machine. With the increasing number of enrollments, the demand to identify, analyze, and classify the activities in MOOCs has arisen and become a critical issue [3]. Some classical classification models in educational domains include support vector machine, neural networks, decision trees, logistic regression, and random forest. Support vector machine (SVM) is a classical supervised learning algorithm that analyzes data and recognizes patterns, used for classification and regression analysis [14].

(3) Classification model: Artificial neural networks. Artificial neural networks (ANNs) are a family of statistical learning models inspired by biological neural networks. ANNs are presented as systems of interconnected “neurons” which send messages to each other. The greatest advantage of ANNs is their ability to be used as an arbitrary function approximation mechanism that “learns” from observed data [37].

3.3.3 Learning style model

Learning styles research has many implications and is a central element of modern teacher training courses [42]. In general, a learning style model is used to classify students according to where they fit on a number of scales pertaining to the ways they receive and process information [20]. There are
over 70 learning styles theories and models in the literature, such as Felder–Silverman learning style model [20], Kolb’s model [30], Honey and Mumford’s model [22,25], and Visual, auditory, and kinesthetic (VAK) model [33]. The Felder–Silverman’s model denotes four areas of personality that contribute to learning [20]. They are active or reflective, sensing or intuitive, visual or verbal, inductive or deductive, and sequential or global. Kolb’s model is based on his experiential learning model [30]. Kolb’s model gave rise to the Learning style Inventory, an assessment method used to determine an individual’s learning style. Honey and Mumford’s model, Learning Styles Questionnaire (LSQ) is directly derived from Kolb’s theory with only a couple of differences. They substitute the terms “reflector” for divergers, “theorist” for assimilators, “pragmatist” for convergers, and “activist” for accommodators [22,25]. Visual, auditory, and kinesthetic (VAK) learning style uses the three main sensory receivers: Visual, auditory, and kinesthetic to determine the dominant learning style. It is sometimes known as VAKT (Visual, Auditory, Kinesthetic, & Tactile). Due to its simplicity, VAK seems to be about the most popular model. But the research has so far been unable to prove the using one’s learning style from VAK model provides the best means for learning a task or subject [13].

The version of the learning style assessment used on FOC was Honey and Mumford’s model [25]. A total of 183 undergraduate students from Shenzhen University took part in assessing their own learning styles by agreeing or disagreeing with given statements relevant to their learning habits. Following the experimental setting in Ref. [42], we make the distinction between those students whose learning styles are characterized as predominantly active, predominantly passive, and both active and passive. To those students whose learning style is both active and passive, most of them claim that they will become active learners in a more competitive environment.

4 | RESULTS AND ANALYSIS

4.1 | Results for research question 1

Totally, 1,783 participants are registered for this course. A total of 1,190 of them are from Shenzhen University, 554 learners are from other universities or institutes, and another 39 students are online learning users on UOOC. Although the number of participants is less than other MOOC courses, it is still much bigger than the number of students in traditional classes. We first show the exploded doughnut chart of the registrants in FOC (Figure 2). The arc length of each slice represents the number of participants. Shenzhen University is not very famous, and all of the video lectures are in Chinese. Even though these facts are real, about 600 persons from outside Shenzhen University still are attracted to join in this course. In existing work, MOOCs are a useful method to alleviate insufficient computer science (CS) education in some universities or institutes [32]. This phenomenon also exists in our case. Most of these students are from the universities that do not supply CS-related courses, such as Southern Medical University. One of the students is from Changchun, which is a city in the northernmost part of China. The distance of Shenzhen and Changchun is about 1,900 miles. This distance is further than Paris to Moscow. From this observation, we could say that MOOCs could supply the convenience of learning at their own time and location.

From our questionnaire survey data, about 85% participants in our course are familiar with Internet, but only 6% participants are familiar with programming or some other computer science-related skills. Compared with the popularity of MOOCs in some western countries, most of participants in China did not know MOOCs before. In our survey data, only 4% participants had experience on international platforms of MOOCs, and another 4% used some Chinese MOOCs platforms. Seventy-three percent participants heard the term of “MOOC” before, while 27% participants did not know anything at all. There exists no difference between majors or genders. In our survey data, 55% of the participants select MOOCs because they think MOOC is interesting and is a good way to acquire knowledge. Twenty-seven percent participants select MOOCs because they believe MOOC allows them to acquire novel skills, which will be useful to job hunting. To the advantages of MOOCs, 85% participants agree that it provides the convenience of learning at their own pace, time, and location. Sixty-one percent participants indicate the repeatability is helpful to understand the content.

Although learning with MOOCs has many conveniences, in many courses of MOOCs, up to 90% students drop out. The number of participants in different periods of FOC is provided (Figure 3). In this figure, A1 to A8 represent the assignments of FOC from Chapters 1 to 8. As we have known, in most of MOOCs, such as the course Bioelectricity, Fall 2012 at Duke University [9], there was a steep drop of participation in the first several weeks. In FOC, this phenomenon is also existing, an obvious jump happens between the registration and the first assignment. And after the first assignment, the number of participants becomes stable.

4.2 | Results for research question 2

As we know, the learning style of participants can have direct and substantial effects on learning behavior, especially on online learning. In this section, we seek to investigate the influences on learning behaviors in MOOCs of different learning styles.

We list some question samples from the questionnaire in Table 1. Q stands for Question. In these questions, Q.12 is directly related to learning style. All of the choices for this
question are listed in the second row of Table 1. We first analyze the correlations between the interactive activity and the learning style. Q.11 is about the interactive activity on the discussion forum of FOC. All of the choices of this question are listed in the first row of Table 1. The matching matrix with the correlations of the responses between Q.11 and Q.12 are provided (Figure 4a). The red number in the corresponding choice shows the number of students. Only less than one-third of participants are active learners. More than half of them have participated in the discussions. Another 25% participants are passive learners. None of these passive learners has asked or answered questions on the forum. And 76% of these students have not joined in the forum. Nearly half of students’ learning style is both active and passive. About 65% of these students have not taken part in the discussion forum.

Although 65% is smaller than that of passive learners, this ratio is still high enough.

We want to investigate the correlations between the adverse arguments and the learning types. Thus, the matching matrix with the correlations of the responses between Q.12 and Q.16 are provided (Figure 4b). Q.16 is about the criticisms of MOOCs, all of whose choices are listed in Table 1. From Figure 4b, we find about 60% active learners argue the online practices or other homework are not enough. Less of active learners think the MOOCs are lack of monitoring and management, or lack of enthusiasm and ambition, or difficult to focus attention. It means active participants already have the ability to learn by themselves. They do not require the pressure and supervision from outside. In this case, they are eager to different challenges, maybe more practices or other types of homework is a good way to assist them to learn the content effectively. This phenomenon did not happen to other two types of learners. To the potential active learners and passive learners, more complaints aim to hard to focus attention on the content, and the lack of teachers’ monitoring and management.

The future attitude toward learning in MOOCs is one of the important factors to evaluate the quality of MOOCs. In our survey, Q.18 is about it. We also list all the choices of Q.18 in Table 1. To demonstrate the correlations between the future attitude and the learning styles, the matching matrix with the correlations of the responses between Q.18 and Q.12 is provided (Figure 4c). From this figure, we can find, after a semester, most of the participants have accepted MOOCs as an effective learning paradigm. Only 9% of posts agree that classroom teaching is better than MOOCs.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Q.6</td>
<td>(Multiple answers) What do you learn from this MOOC course?</td>
<td>Have some interests in computers, familiar with the use of some popular software</td>
<td>Have the elementary knowledge, construct &quot;computational thinking&quot;</td>
<td>Acquire novel skills, and it will be useful to job hunting</td>
<td>Generate great interests in computer science, want to do some related work in future</td>
</tr>
<tr>
<td>Q.7</td>
<td>(Single answer) How long does it take to learn FOC in each week?</td>
<td>Less than one hour</td>
<td>One hour to two hours</td>
<td>Two hours to three hours</td>
<td>More than three hours</td>
</tr>
<tr>
<td>Q.8</td>
<td>(Multiple answers) Do you know anything about MOOCs before?</td>
<td>Had experiences on international platforms</td>
<td>Had experiences on Chinese platforms</td>
<td>Heard the term or the concept of &quot;MOOC&quot; before</td>
<td>Not knowing anything at all</td>
</tr>
<tr>
<td>Q.9</td>
<td>(Single answer) How is your study progress?</td>
<td>Follow the requirement of the course</td>
<td>Try to follow the schedules, fall behind one or two weeks</td>
<td>Cannot catch up with the requirements of this course</td>
<td>Start studying before the final exam</td>
</tr>
<tr>
<td>Q.10</td>
<td>(Single answer) How do you solve the problem?</td>
<td>Discuss with other students</td>
<td>Ask questions on the discussion forum</td>
<td>Send an email to the lecturer</td>
<td>Find answers online</td>
</tr>
<tr>
<td>Q.11</td>
<td>(Single answer) What is your interactive activity on the discussion forum of FOC?</td>
<td>Ask and answer questions on the discussion forum</td>
<td>Post questions on the discussion forum</td>
<td>Answer questions on the discussion forum</td>
<td>Do not participate in the discussions</td>
</tr>
<tr>
<td>Q.12</td>
<td>(Single answer) What is your learning style?</td>
<td>Active learning</td>
<td>Passive learning</td>
<td>Both active and passive learning</td>
<td>–</td>
</tr>
<tr>
<td>Q.13</td>
<td>(Multiple answers) What are the advantages of learning in MOOCs?</td>
<td>Using a variety of teaching materials and methods</td>
<td>The convenience of learning at own pace, time and location</td>
<td>The repeatability is helpful to learn the content broadly and deeply</td>
<td>The discussions on the forum are helpful</td>
</tr>
<tr>
<td>Q.14</td>
<td>(Multiple answers) Which parts do you think are better to learn in MOOCs?</td>
<td>The basic knowledge of computer technology</td>
<td>Algorithm</td>
<td>Introduction of software and hardware</td>
<td>New development in computer technology</td>
</tr>
<tr>
<td>Q.15</td>
<td>(Multiple answers) Which parts are better to learn by traditional classroom teaching?</td>
<td>The basic knowledge of computer technology</td>
<td>Algorithm</td>
<td>Introduction of software and hardware</td>
<td>New development in computer technology</td>
</tr>
<tr>
<td>Q.16</td>
<td>(Multiple answers) What are the disadvantages of learning in MOOCs?</td>
<td>Monitoring and management are not enough</td>
<td>Learn by myself makes me lack of enthusiasm and ambition</td>
<td>Difficult to pay attention to the online contents, it makes me hard to learn</td>
<td>Online practices or other homework are not enough</td>
</tr>
<tr>
<td>Q.17</td>
<td>(Multiple answers) What can be improved in FOC?</td>
<td>Provide more offline courses as a supplement</td>
<td>Increase network bandwidth to solve network congestion problem</td>
<td>Increase the diversity of exercises and quizzes to consolidate the knowledge</td>
<td>Provide more tools to help students learn on the platform</td>
</tr>
<tr>
<td>Q.18</td>
<td>(Multiple answers) What is your future attitude toward learning in MOOCs?</td>
<td>Understand what MOOCs are, I wouldn’t exclude the possibility of learning in MOOCs</td>
<td>Be familiar with it, prefer MOOCs to traditional classroom teaching</td>
<td>Select MOOCs or traditional classroom teaching based on the nature of the course</td>
<td>Traditional classroom teaching is better than MOOCs</td>
</tr>
</tbody>
</table>
About 24% participants in active learners are familiar with MOOC, and they will prefer MOOCs to classroom teaching. To the potential active learners and passive learners, this number decreases to 20% and 16%, respectively. This difference in the number demonstrates that the active learners are more open to exploring the new education methods and technologies. We could expect most of the students have an understanding of MOOCs. In the next time, they can choose MOOCs or traditional classroom education based on the nature of the course. Although we can also find the number of recommendation for MOOCs is slightly lower than the reported number from one existing paper [24], considering the fact that many Chinese students get used to traditional classroom learning paradigm for a long time, we can say that they had a good experience in learning MOOCs.

From these results, we can find that the learning styles have a direct influence on some learning habits, such as interactive experience on MOOCs, disadvantages of MOOCs, and future attitudes toward MOOCs.

### 4.3 Results for research question 3

In the previous section, we found that the learning styles have a direct influence on some learning behaviors. But we do not know whether it only occurs in a small, specific set of behaviors or most of the learning behaviors. If it is the latter, subjects could be automatically divided into several groups that are coincident with the learning styles.

To answer these questions, we use k-means clustering model to group the subjects’ responses. Let \( X \) be responses from all subjects, and \( x_i \in X \) is used to denote the vectorized response datum for subject \( i \). For example, to question \( l \), which includes four choices, if subject’s answer is “A,” the vectorized sample for this question is \([1,0,0,0]\); if the answer is “AD,” the vectorized sample is \([1,0,0,1]\). In our clustering, we only consider the responses from Questions 6 to 18 for all 183 subjects. Thus, the dimension of \( x_i \) is \( 1 \times 53 \), and the dimension of \( X \) is \( 183 \times 53 \). Then, we use \( k \)-means clustering model to group the subjects’ responses \( X \). We set the number of cluster centers \( C \) to be 3, which is equal to the number of learning styles. To each group, we calculate the number of subjects that are consistent with the corresponding learning style. Then, we obtain the best corresponding group between the learning styles and the clusters. \( N_s \), which is the maximum value of this consistent number, as Eq. (1):

\[
N_s = \max \left( \sum_{j=1}^{C} \sum_{i=1}^{C} I(y_i = k_j) \right)
\]

where \( I \) is the sign function, \( y_i \) is the learning style corresponding to \( x_i \). \( C_j \) is the number of \( x_i \) in the \( j \)th cluster, and \( C \) is the number of clusters. \( k_j \) is defined as Eq. (2):

\[
k_j = [k_j^1, k_j^2, ..., k_j^c, ..., k_j^C], k_j^i \neq k_j^t, k_j^c
\]

\[
k_j = \begin{cases} 
  1 & c = t \\
  0 & c \neq t 
\end{cases}, \quad t \in \{1, 2, ..., C\}
\]

All the statistical experiments are repeated for ten times, and the average results are reported. The average accuracy (\( N_s \), 183) across ten trials is 56.52%. The average accuracy of the clustering is significantly better than the random guessing baseline (33.3%). From this result, we can say, if two subjects share the same learning style, it is highly possible that they demonstrate similar learning behaviors.
4.4 Results for research question 4

In the previous section, we found that the learning styles have a close relationship with the learning behaviors. Students’ learning behaviors can be clustered into several groups, and the division is the learning style. In this part, we want to move one step further to predict students’ learning styles based on their learning behaviors.

Let \( X \) be a set of responses from all subjects, and \( x_i \in X \) is used to denote the vectorized response datum for subject \( i \). In this part, \( X \) does not include the responses for learning styles. Thus, the dimension of \( x_i \) is \( 1 \times 50 \), and the dimension of \( X \) is \( 183 \times 50 \). Let \( Y \) be a set of labels corresponding to \( X \), and \( y_i \in Y \) is used to denote the label vector of \( x_i \).

\[
y'_i = \begin{cases} 
1 & \text{if } x_i \in c_j \text{ class} \\
0 & \text{if } x_i \not\in c_j \text{ class} 
\end{cases}
\]  

(3)

Here, different class labels are corresponding to different learning styles. Thus, the number of classes \( N_C \) is equal to 3.

We divide the data into two groups, including 90 training data and 93 test data. Based on the given training set, the aim is to learn a mapping function from the set \( X \) to the label set \( Y \), and then classify the new coming data points according to the learned mapping function. All the statistical experiments are repeated for ten times with randomly selected training sets, and the average results are reported. In our experiments, we choose two classical supervised learning methods to predict students’ learning styles based on their learning behaviors, including support vector machine (SVM) and artificial neural networks (ANN). To SVM, we set the penalty factor as 1,000. For parameters such as the learning rate and the momentum in the ANN, we follow the general setting of previous work. Learning rate is set to be 2, and momentum is set to be 0.5. And we simply use two-layer ANN. The number of neurons in the input layer is set as the number of responses; the number of neurons in the output layer is set as the number of classes (the number of learning styles).

We report the average classification accuracy over ten trials on test dataset (Table 2). From the results, we can find both of methods achieve excellent classification accuracy. It means it is easy to predict students’ learning styles based on students’ learning behaviors. On the other hand, it also means that the learning style is a factor to indicate students’ learning behaviors, or even measure whether a student is appropriate to learn via MOOCs.

<table>
<thead>
<tr>
<th>Classification model</th>
<th>Average accuracy</th>
<th>STD</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>99.79</td>
<td>0.4347</td>
</tr>
<tr>
<td>ANN</td>
<td>100</td>
<td>0</td>
</tr>
</tbody>
</table>

5 DISCUSSION

Compared with other work, we make three main contributions in this paper:

1. Develop data mining method to explore the relationship between students’ learning styles and learning behaviors with a reasonable accuracy.

   Based on the unsupervised clustering method, we find most of the students could be automatically divided into several groups that are coincident with their learning styles. By using two classical supervised learning methods, we can easily predict students’ learning styles based on their own behaviors. Both of methods achieve excellent classification accuracies. Similar with the performance of our model, a number of studies in education field have shown that data mining and analytics are indeed possible to explore students’ behaviors and learner’s engagement patterns with a reasonable accuracy. All these results mean that data mining has potential as a valuable tool for discovering how people learn, predicting learning, and understanding real learning behavior [5]. By achieving these goals, educational data mining should be used to design better and smarter learning technology and to improve both the quality and delivery of MOOCs. In next decade, the key goals for the education community include bringing better practices to the whole field and more deployment into intelligent education systems.

2. Enhance the understanding of behavior differences between students belonging to different learning styles.

   New insight should not just come from analyzing new problem or task, but they should also from analyzing it within the context of the old to provide new perspectives on old problems. Learning style is not a new concept or theory. It is an important and necessary element of modern education and it has many pedagogical implications. Unfortunately, there is very little existing work try to explore the learning activities between students’ with different learning styles. MOOCs provide researchers abundant data from diverse population of MOOC participants that have never been available at such scales before. It offers good opportunity for researchers from cross-disciplinary fields to investigate this important educational topic and its roles in influencing students’ learning behaviors and academic achievement. Our work seeks to explore the relationship between students’ learning styles and learning behaviors in MOOCs via computational methods. Based on data mining methods, we find students belonging to different learning styles behave differently in MOOCs. As we have known, how to provide the instructions to guide the interactions in MOOCs is highly important but tricky [41]. Based on our work, the learning style can be used as an effective factor to measure and predict learners’ behaviors in MOOCs. Therefore, we suggest the teachers in MOOCs could
talk to students about learning styles. Explaining to the learners with different learning styles how they learn most efficiently could be an important step in helping them improve their learning experiences [20].

3) Provide empirical support for learning in MOOCs with a Chinese education context.

Although a number of studies have explored students’ behaviors and learner’s engagement patterns in MOOCs to understand issues of persistence, most of them refer to MOOCs that have been taught in western countries, but we consider the particularities of Chinese society in this new way of e-learning. As we have known, the low completion rate calls into questions about the quality of MOOCs [12]. We find that there exists a difference of the completion rates between the students from Shenzhen University and other universities. In the students from Shenzhen University, about 74% of them complete the course and obtain the certificates of FOC. About 30% of students from other universities complete the course and obtain the certificates. This difference is resulted from two factors. First, video lectures serve as the primary content in MOOC, thus the lack of engagement is a known problem. In Shenzhen University, the instructors, mentors, and lecturers of FOC hold some serials of off-line lectures on campus. These lectures support an engaging and interactive mode of content delivery. From the questionnaire survey data and the feedbacks from students, these off-line lectures are especially helpful for passive learners. In all chapters of FOC, students responded that these off-line supplementary lectures are useful to learn the Algorithm part. CS activities for young students are widely used, particularly with visual programming environments [4]. Existing work has evidenced that students prefer to watch hand-writing codes on a visual programming environment than stare at a slide of video with codes already on it [50]. Therefore, we expect the off-line lectures of Algorithm part could be able to convey the sense of immediacy. Second, it is believed that Chinese students are used to ask questions after they have learned rather than during the process of learning as western students do [27]. To the students in the same campus of Shenzhen University, they are convenient to share and discuss the content of the course than other students. Therefore, in the next year of our MOOC course, we have provided more online and offline learning resources to students. To other MOOC providers or university partners, we suggest them to understand the differences of learners and develop more customized services for them.

A limitation of the current study is the homogeneity of the participants. As we have known, most of the participants in Foundation of Computer Science are undergraduates from Shenzhen University. The questionnaire survey data and daily activity data are almost from them too. Therefore, the population of our study seems homogeneous than other MOOCs. But when we go back to the Questionnaire survey data, we could find, even these students are undergraduates, only 4% of them had study experiences on the international platforms of MOOCs, and another 4% had learning experience on Chinese MOOCs platforms. Unlike the popularity of MOOCs in some western countries, most of the users in China are students from University. Thus, homogeneity of the participants is a property that holds with the respect of the particular case of Chinese students.

In education, big data is already reality. What is not a reality yet is the analysis of educational data to understand learning and teaching better and to improve them [36]. Although numerous data mining and analytics approaches have been successful at modeling a range of phenomena relevant to student learning in online systems, some important tasks have been tackled and interesting findings have been discovered, at least two challenges are on the way. One is security and privacy challenge. Great power of data means great responsibility. A big data initiative should not only focus on the Velocity, Variety, and Volume of the data, but also on the best way to protect it. Security issues challenges are being amplified by Velocity, Variety, and Volume of big data, such as large-scale cloud framework, distinction of data source and pattern, cascading nature of data acquisition [8]. Each of these Vs of big data is growing at an astounding rate and required a shift in how security vendors manage threats [47]. Moreover, users in MOOCs platform have to trust what happens with their daily data that platform store and analyze. A reasonable answer is opt-in: Interactions are stored only when users opt for it [36]. But it will seriously limit the available data, and also has a negative impact on the data mining and analytics approaches. Another important challenge is generalizability: Is a data mining approach valid for real time application? Currently, my answer is probably not. This reality will slow down the adoption of educational data mining and learning analytics in daily life. A real-time big data mining system should go deeper into the length of each learning time pieces [47]. It is highly required to the development of fast and efficient algorithms for real-time big data analytics to deliver accurate predictions of various kinds, such as students’ performance, detector of behaviors.

6 | CONCLUSION AND FUTURE WORK

MOOCs have captured the attention of higher education institutes around the world. Our paper does not aim to supply solutions to the problems in MOOCs but rather offers recommendations for improving the teaching and learning experience. This work presents an analysis of students’ learning behaviors in a MOOC on computer science course for the context of Chinese society. Although there are several existing work already published in the literature on this topic, most of them refer to MOOCs that have been taught in Western countries, but we consider the particularities of Chinese society.
in this new way of e-learning. Both positive and negative arguments of MOOCs could be obtained in this paper. Based on its open access via the web, FOC has attracted a larger number of students than traditional courses. Thus, it can increase the prestige of a higher education institution. On the contrary, we also find most of students could not follow the expected schedule of this course. Although the completion rate of FOC is high, it is still lower than the traditional courses. More importantly, according to the analyses of students’ learning behaviors and learning styles by classical clustering and classification models, we find it exists behavior differences between students belonging to different learning styles. We suggest that the students’ behaviors are indicative of their own learning styles. Investigating these relationships is one way of helping to improve the quality of MOOCs course delivery since it is a precursor to advanced instructor analytics and individualization.

Computational simulations and other interactive support methods can increase students’ understanding of difficult concepts in online engineering education [51]. In future, we try to explore more interactive support methods to help students construct more understanding about their own learning styles and study habits. Another meaningful future work is to investigate whether different cultures have an impact on the results of the study and whether the results of our work could be generalized in other MOOC environments.

REFERENCES


**Y. Li** received her MS in Computer Systems and Structure in Computer College from Chongqing University in 2001. She is an assistant professor in College of Computer Science & Software Engineering at Shenzhen University in Shenzhen. Her research interests include multimedia information processing and MOOC education.

**Y. Liu** received her BEng from Department of Electrical Engineering at Southeast University in 1996 and MSc from School of Business at Nanjing University in 1999. She got her PhD from Department of Computer Science at Columbia University in 2005. Currently, she is an associate professor in Department of Computing at The Hong Kong Polytechnic University. She is the director of cognitive computing lab. Her research interests include multimedia content analysis, machine learning, cognitive science and brain modeling.

**Z. Wang** received his MSc degree in Theoretical Electrical Engineering from Hubei University of Technology, Wuhan, China, in 1991. He is a full professor in College of Computer & Software Engineering at Shenzhen University in Shenzhen. His research interests include multimedia content analysis and computer education.