Scholar Publishing Group
Machine Learning Theory and Practice
https://doi.org/10.38007/ML.2022.030403
ISSN 2790-0983 Vol. 3, Issue 4: 18-26

Online Education User Behaviour Based on Machine Learning

Yanan Li*
Liaoning Metallurgical Vocational and Technical College, Liaoning, China
nice-you@sohu.com
**corresponding author*

Keywords: Machine Learning, Online Education, User Behaviour, Behaviour Prediction

Abstract: The rapid development of machine learning has made it possible to quickly mine exploitable data from massive amounts of data, and it is important for online education platforms to reasonably track the changing learning behaviour of learners. The aim of this paper is to study the prediction of online education user behaviour based on machine learning. A deep clustering algorithm is proposed based on demographic information about learning behaviour and clickstream data recorded in a university virtual learning environment. Applying the deep clustering model proposed in this paper to the analysis of the distribution of learning activities and the correlation between learning behaviour and education level, it is demonstrated that the deep clustering model proposed in this paper's can analyse and predict the learning behaviour of different student groups. It helps us to further study and analyse students' learning behaviour in depth and comprehensively in order to provide real-time quality feedback and promote the development of online education.

1. Introduction

Education informatization has entered a new stage of development and is shifting from digital education to intelligent education supported by modern information technology such as big data analysis and artificial intelligence [1-2]. At present, large-scale online education has accumulated a huge amount of teaching behaviour data and knowledge resources [3]. Despite the complexity and diversity of user behaviour in the virtual space, certain regularities in online user behaviour can be found by analysing online behaviour such as user web access and web page clicks. At the same time, these patterns have obvious human characteristics [4-5].

Current research on user behaviour has involved research areas from a variety of disciplines such as computer science, sociology and management science, and the involvement of multidisciplinary perspectives has led to a clearer characterisation of Internet user behaviour [6].Saeid SadighZadeh

developed a generic neural representation learning framework for modelling user behaviour, aiming to address the problem of sparsity between different applications. Their problem setting spans both conductive and inductive learning scenarios, where conductive learning models entities seen during training and inductive learning targets entities observed only during inference. They exploit different aspects of information that reflect user behaviour (e.g. interconnectedness, time and attribute interaction information in social networks) to enable large-scale personalised inference. Their proposed model complements simultaneous advances in neural architecture selection and is able to accommodate the rapid addition of new applications in online platforms [7].Malvika Singh proposes a multi-level user behaviour visualisation framework that provides efficient visualisation of user behaviour data collected from production vehicles via telematics. Their approach visualises user behaviour data at three different levels: a Task Level View aggregates event sequence data generated through touchscreen interactions to visualise user flows; a Flow Level View allows comparison of individual flows based on selected metrics; and a Sequence Level View provides detailed insights into touch interactions, line of sight and driving behaviour. Our case study demonstrates that UX experts find their approach a useful addition to their design process [8].Sylvia Chan-Olmsted proposes two approaches to travel user behaviour analysis: a raw sequence constructed from categorical data describing the user's activity over time, and a time sequence, which is an enhanced version of the first approach and includes a non representation of activity time frames. The analysis of user travel behaviour can be used for adaptive gamification strategies. The method was evaluated on a behavioural atomic dataset based on one year of Foursquare check-ins. The results show that the two methods reflect different aspects of travel user behaviour and that both can be used in a complementary way [9]. It is therefore relevant to study machine learning-based prediction of online education user behaviour.

In this paper, the main focus is on online education user behaviour prediction as a research context for the study of user behaviour. Firstly, recorded data is used to characterise basic user behaviour data, such as users' click records. Secondly, clustering methods are used to extract user contextual information, such as common places, to characterise information about features associated with user behaviour. Finally, information about the user's past is extracted and analysed to predict the user's future activities and related information, such as the user's learning goals.

2. A Study of Online Education User Behaviour Prediction Based on Machine Learning

2.1. User Behaviour

User behaviour refers to the behaviour of Internet users who use the Internet to engage in business, life, learning and entertainment activities, including both the textual information they fill in during registration and the operation of the mouse, keyboard and other actions. User behaviour is a broad concept, the essence of which lies in its ability to show some characteristics of the user's own attributes and activities [10-11]. User behaviour tends to be blind and usually random over a certain period of time, but in the long run it tends to have a certain pattern. In addition, the behaviour of individual users is often simple and extremely irregular, but a comprehensive study of the behaviour of a large number of users can reveal regularity [12-13].

2.2. User Behaviour Data Acquisition

User behaviour data can be divided into two parts: explicit behaviour and implicit behaviour, each with different acquisition methods and approaches, and the two acquisition methods will be

introduced separately. Obtaining explicit user behaviour data:Explicit behaviour data is the basic information data of the user, which generally includes the user's age, gender, time of use of a system, place of origin, occupation, education level, contact information and other relevant data indicating user information. The explicit knowledge provided by the user during registration is the basic information reflecting the user's traits [14-15].

Implicit behavioural data of users:Implicit behavioural data of users is mainly the recorded data of users' behaviour on the information platform. This includes when they logged in and browsed what kind of content, and what kind of operations they performed on the page, usually sharing, printing, saving, etc [16]. If a peak study period is defined for a certain exam date and a period of 40 days prior to it, then by analysing the date of a single login by a user it is possible to determine if the user is studying during a peak period. By analysing a certain time period, the total number of peak study sessions and the total length of study sessions can be obtained for that user.

2.3. Machine Learning

Machine learning is an implementation of artificial intelligence that trains a model based on sample data and uses the model to make predictions and decisions about the data. Machine learning gives computers the ability to learn in a human-like manner, allowing machines to learn from examples and knowledge like humans, thus having the ability to anticipate and predict [17-18].

Machine learning distinguishes between supervised and unsupervised learning based on the presence or absence of "signals" or "feedback" in the learning system, as shown in Figure 1.

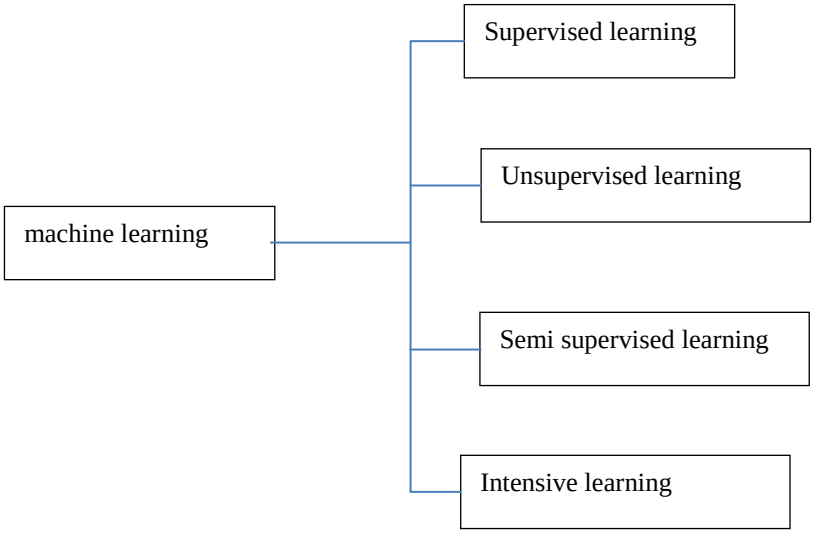

Figure 1. Classification of machine learning

For supervised learning, it can be divided into two categories, classification and regression, according to whether the prediction target is discrete or continuous data. For unsupervised learning, on the other hand, it can be classified into clustering or dimensionality reduction in terms of solving the task.

(1) Classification

Classification, based on some given samples of known categories (i.e. labelled data), enables the computer to classify samples of unknown categories. Classification requires prior explicit

knowledge of the categories and is a supervised learning algorithm for modelling or predicting discrete random variables.

(2) Regression

Regression, a supervised learning algorithm for predicting and modelling numerical random variables, where the predicted outcome is a set of continuous values. Common regression algorithms include random forests and gradient boosting trees.

(3) Clustering

Clustering, as the name implies, is the grouping of similar samples together. Data objects are usually grouped together based on some relationship found in the data, where objects within the same group are similar to each other, while objects in different groups are different. The greater the similarity within groups and the greater the gap between groups, the more effective the clustering is. Common clustering algorithms include divisional clustering, hierarchical clustering, density clustering, grid clustering and hybrid clustering.

(4) Dimensionality reduction

Dimensionality reduction, which is essentially learning a mapping function, refers to the use of some mapping method to map data points originally in the high latitude space to the low latitude space. Common dimensionality reduction methods include PCA principal component analysis and LDA discriminant analysis, etc.

3. A Survey and Study on Predicting User Behaviour in Online Education Based on Machine Learning

3.1. Online Learning Dataset

The Online Education University Learning Analytics dataset OULAD collects relevant learning data for the years 2020 and 2021 and records it in tables. Each table represents different information separately, and the data in each table can be related by using identifiers such as primary or foreign keys in the database to identify the relevant columns. The structure of the data contained in the dataset is shown in Figure 2.

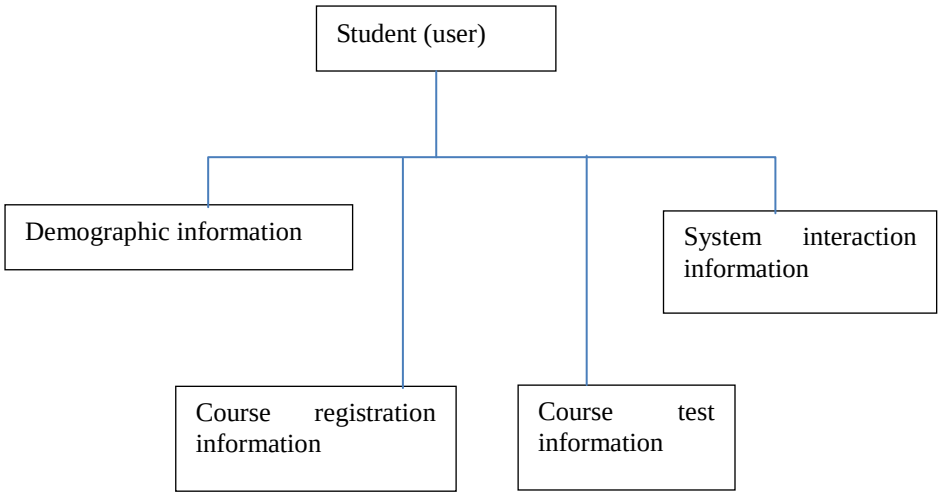

Figure 2. Dataset structure

3.2. Deep Clustering Model

The overall framework of the model consists of a combined GMM clustering layer and a latent feature extraction layer. A stack of four layers of autoencoders are first stacked to form a stack autoencoder for extracting deeper features while maintaining a local data distribution structure. Significant latent features are identified from the input layers of each autoencoder. The depth features extracted by the encoders are then reconstructed by a decoder consisting of a four-layer fully connected layer.

The clustering loss used in this model is the K-L scatter of the decoder's defined distribution and the predefined target distribution, while clustering training of the features is introduced and the clustering loss is calculated. The network parameters are eventually updated by jointly training the reconstruction loss of the stacked autoencoder and the clustering loss of the Gaussian mixture model to optimise the final parameters. The final objective of this deep clustering model can be represented as follows:

$$L = L_r + \gamma L_c \tag{1}$$

As shown in equation (1) above, where Lr and Lc denote reconstruction loss and clustering loss respectively. The formula for reconstruction loss Lr is shown in (2).

$$L = \left\| x_i + \hat{x}_i \right\|^2 \tag{2}$$

The reconstruction loss represents the difference between the input data xi and the decoder's reconstruction data $\hat{x}_i$ of the encoder input data.

4. Analysis and Research on the Prediction of Online Education User Behaviour Based on Machine Learning

4.1. Distribution of Learning Activities

In order to determine the intensity of students' learning behaviors at the early stage of the course, this study analyzed the behavioral data recorded before submitting the first class assignment. The distribution of participants' learning behaviors relative to their final performance is shown in Figure 3. As shown in the figure, there are four types of students who pass, excel, withdraw and fail. Prior to the first class assignment submission, the learning patterns of students in the four final performance categories were similar prior to the first assignment submission. Therefore, it is difficult to distinguish the final performance of the four groups of students based on the average clickstream of each learning resource. At this point, we can easily conclude that the students who failed or dropped out did not work as hard before submitting their first assignment as those who received outstanding grades.

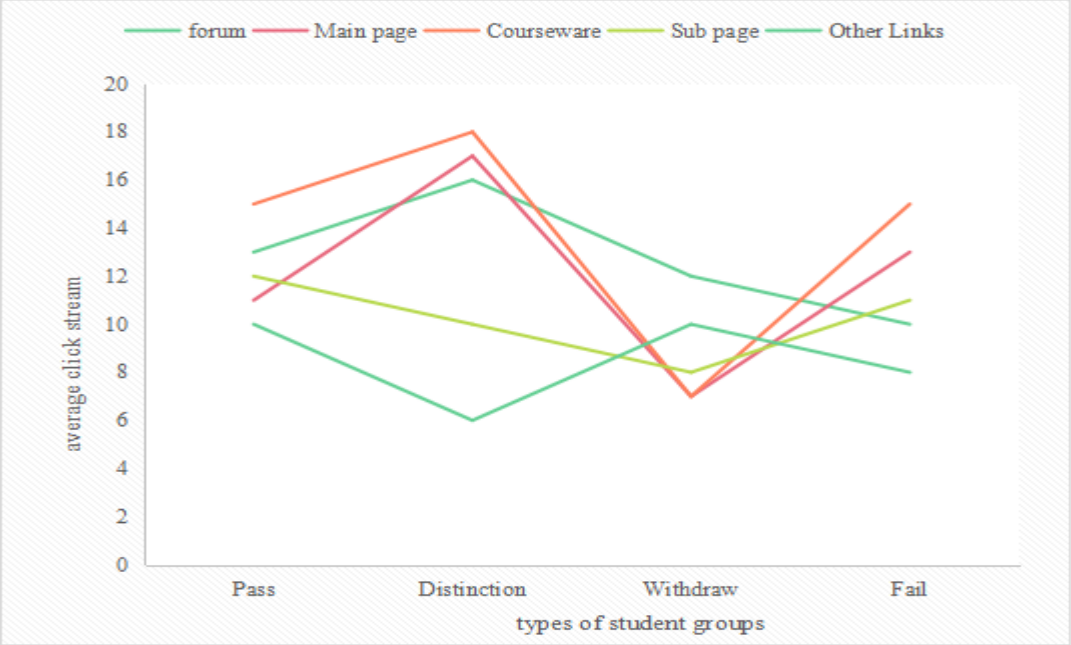

Figure 3. Distribution of learning activities of different final performance student groups before submitting the first assignment

4.2. Correlation Experiment between Learning Behavior and Education Level

This paper also analyzes the relationship between different education levels and learning behavior. In this data set, the educational attainment of the participants was broadly divided into four categories, namely: "A level or equivalent to A level", "higher education qualification", "below A level" and "postgraduate qualification". Table 1 shows the distribution of students with different education levels in different behavior groups. It can be found that the distribution of students of the four educational levels of the online viewing courseware group is relatively average, as shown in Figure 4. In the group of students without learning behavior, "higher education" and "graduate education" account for a small proportion of students, which is also in line with what we know daily that most students with higher education level have strong learning motivation or learning goals.

Table 1. Correlation between learning behavior and education level

Education level	Watch courseware online	Forum discussion	No learning behavior
Postgraduate education	38	45	6
Below Level A qualification	44	15	28
Higher education qualification	55	57	21
Class A or equivalent	36	32	38

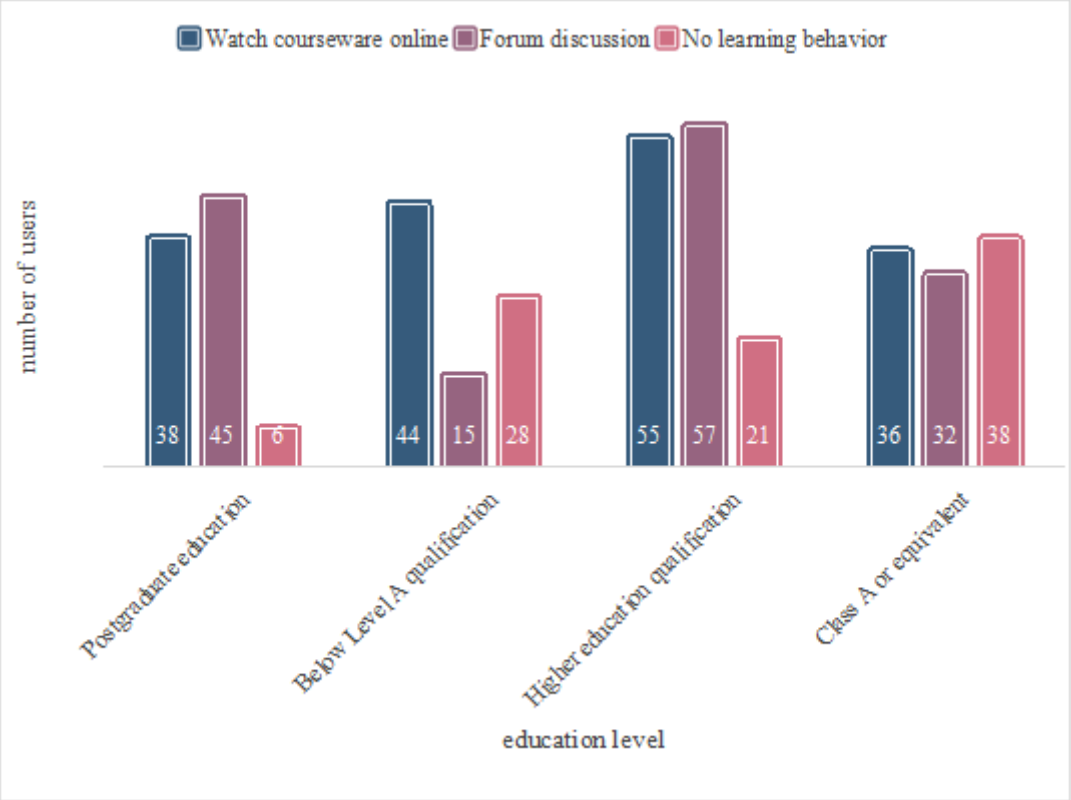

Figure 4. Distribution of education level in different learning behavior groups

5.Conclusion

A lot of information of users in the process of using the Internet, such as visitors' personal information, click process, click start and end time, are often recorded in user behavior logs, which provides the possibility to analyze users' online behaviors with the help of behavior log big data. The research in this paper verified the prediction of online education user behavior based on machine learning, identified and differentiated user groups with high similarity, and finally explored the distribution of education level in different learning behavior groups. Although this paper has made some breakthroughs in the research on the prediction of user behavior and the construction of algorithm model, there are still unresolved gaps in the research work, which need to be further explored. Due to the complexity of human beings, their online behaviors are not only affected by external factors but also internal factors. In the next work, We will continue to expand the amount of experimental data and further select higher dimensional feature vectors to observe the accuracy of the user behavior prediction model.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this

study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] *Vashek Matyas, Kamil Malinka, Lydia Kraus, Lenka Knapova, Agata Kruzikova: Even if users do not read security directives, their behavior is not so catastrophic. Commun. ACM 65(1): 37-40 (2022) https://doi.org/10.1145/3471928*

[2] *R. Geetha, S. Karthika, Ponnurangam Kumaraguru: 'Will I Regret for This Tweet?' - Twitter User's Behavior Analysis System for Private Data Disclosure. Comput. J. 65(2): 275-296 (2022) https://doi.org/10.1093/comjnl/bxaa027*

[3] *Noureddine Amraoui, Belhassen Zouari: Anomalous behavior detection-based approach for authenticating smart home system users. Int. J. Inf. Sec. 21(3): 611-636 (2022) https://doi.org/10.1007/s10207-021-00571-6*

[4] *Husna Sarirah Husin, James A. Thom, Xiuzhen Zhang: Evolution of user navigation behavior for online news. Int. J. Web Inf. Syst. 18(1): 1-22 (2022) https://doi.org/10.1108/IJWIS-06-2021-0064*

[5] *Lori Baker-Eveleth, Robert W. Stone, Daniel M. Eveleth: Understanding social media users' privacy-protection behaviors. Inf. Comput. Secur. 30(3): 324-345 (2022) https://doi.org/10.1108/ICS-07-2021-0099*

[6] *Jongpil Park, Jai-Yeol Son, Kil-Soo Suh: Fear appeal cues to motivate users' security protection behaviors: an empirical test of heuristic cues to enhance risk communication. Internet Res. 32(3): 708-727 (2022) https://doi.org/10.1108/INTR-01-2021-0065*

[7] *Saeid SadighZadeh, Marjan Kaedi: Modeling user preferences in online stores based on user mouse behavior on page elements. J. Syst. Inf. Technol. 24(2): 112-130 (2022) https://doi.org/10.1108/JSIT-12-2019-0264*

[8] *Malvika Singh, Babu M. Mehtre, S. Sangeetha: User behavior based Insider Threat Detection using a Multi Fuzzy Classifier. Multim. Tools Appl. 81(16): 22953-22983 (2022) https://doi.org/10.1007/s11042-022-12173-y*

[9] *Sylvia Chan-Olmsted, Rang Wang: Understanding podcast users: Consumption motives and behaviors. New Media Soc. 24(3): 684-704 (2022) https://doi.org/10.1177/1461444820963776*

[10] *E. Karthik, T. Sethukarasi: Sarcastic user behavior classification and prediction from social media data using firebug swarm optimization-based long short-term memory. J. Supercomput. 78(4): 5333-5357 (2022) https://doi.org/10.1007/s11227-021-04028-4*

[11] *Mahyar Kamali Saraji, Abbas Mardani, Mario Köppen, Arunodaya Raj Mishra, Pratibha Rani: An extended hesitant fuzzy set using SWARA-MULTIMOORA approach to adapt online education for the control of the pandemic spread of COVID-19 in higher education institutions. Artif. Intell. Rev. 55(1): 181-206 (2022) https://doi.org/10.1007/s10462-021-10029-9*

[12] *Jaroslav Majern ́k, Andrea Kacmarikova, Martin Komenda, Andrzej A. Kononowicz, Anna Kocurek, Agata Stalmach-Przygoda, Lukasz Balcerzak, Inga Hege, Ioan-Adrian Ciureanu: Development and implementation of an online platform for curriculum mapping in medical education. Bio Algorithms Med Syst. 18(1): 1-11 (2022) https://doi.org/10.1515/bams-2021-0143*

[13] Steven J. Greenland, Catherine Moore: Large qualitative sample and thematic analysis to redefine student dropout and retention strategy in open online education. Br. J. Educ. Technol. 53(3): 647-667 (2022) https://doi.org/10.1111/bjet.13173

[14] Kyungmee Lee, Mik Fanguy: Online exam proctoring technologies: Educational innovation or deterioration? Br. J. Educ. Technol. 53(3): 475-490 (2022) https://doi.org/10.1111/bjet.13182

[15] Caleb Or, Elaine Chapman: Development and validation of an instrument to measure online assessment acceptance in higher education. Br. J. Educ. Technol. 53(4): 977-997 (2022) https://doi.org/10.1111/bjet.13180

[16] Henriikka Vartiainen, Hanna Vuojärvi, Kaija Saramäki, Miikka Eriksson, Ilkka Ratinen, Piritta Torssonen, Petteri Vanninen, Sinikka Pöllänen: Cross-boundary collaboration and knowledge creation in an online higher education course. Br. J. Educ. Technol. 53(5): 1304-1320 (2022) https://doi.org/10.1111/bjet.13186

[17] Subramani Jegadeesan, Mohammad S. Obaidat, Pandi Vijayakumar, Maria Azees, Marimuthu Karuppiah: Efficient privacy-preserving anonymous authentication scheme for human predictive online education system. Clust. Comput. 25(4): 2557-2571 (2022) https://doi.org/10.1007/s10586-021-03390-5

[18] Anuj Garg, Sharmila A, Pramod Kumar, Mani Madhukar, Octavio Loyola-González, Manoj Kumar: Blockchain-based online education content ranking. Educ. Inf. Technol. 27(4): 4793-4815 (2022) https://doi.org/10.1007/s10639-021-10797-5