Study on behaviour anomaly detection method of English online learning based on feature extraction

Feng Wei

Yongcheng Vocational College,
Yongcheng 476600, Henan Province, China
Email: fengwei@mls.sinanet.com

Abstract: There are many problems in abnormal detection of online English learning behaviour, such as large error and high detection time. Therefore, a detection method based on feature extraction is proposed. Firstly, frequent pattern mining method is used to collect learners' behaviour data, and the data is collected and preprocessed. Then, the classification constraints are set by support vector machine to complete the data classification. Finally, the sequence minimum eigenvalue method is used to train the abnormal data, extract the high frequency features of the abnormal data, establish the anomaly detection model, and realise the anomaly detection. Experimental results show that the highest detection error of this method is 1.2%, and the highest time cost is 1.8 s. Therefore, this method can effectively reduce the detection error and time cost, and is feasible.

Keywords: feature extraction; English online learning behaviour; anomaly detection; threshold; K-means clustering; Lagrange function.

Reference to this paper should be made as follows: Wei, F. (2023) 'Study on behaviour anomaly detection method of English online learning based on feature extraction', *Int. J. Reasoning-based Intelligent Systems*, Vol. 15, No. 1, pp.41–47.

Biographical notes: Feng Wei is currently an Associate Professor and a Dean of the Department of Art and Culture in Yongcheng Vocational College. He is also a chief editor of five textbooks of high education as well as an author of more than ten papers. His research interests include English education and normal education.

1 Introduction

Internet technology has been widely used in many fields. Among them, it has also been applied in the field of education (Bi et al., 2018). Through the application of Internet technology, the teaching of various subjects in the school has made great progress. Among the existing teaching modes, online teaching mode has been better applied. Students and users can improve their academic performance through online video classes and video resources (Du and Chen, 2018). Among them, English online learning is the subject with the most applications and the largest resource scale among all subjects. Students learn English words, English composition writing and oral pronunciation through online English learning video, which effectively improves the quality of students' classroom learning (Chen et al., 2018). However, with the increasing number of English learning data in the network, there are more fraud, intrusion data and dangerous data. The emergence of these data is easy to lead to abnormalities in students' online learning of English and affect the learning effect (Guo et al., 2018). Therefore, through the detection and observation of students' abnormal English online learning behaviour, we can improve the safety of users' learning and ensure the quality of learning. So, researchers have made a lot of exploration on this issue

Wang and Gao (2020) proposed an abnormal behaviour detection method though spatiotemporal fusion convolution method. In this method, vggnet16 is used to design the dual flow model, and the continuous optical flow frames are used as the basic data of the pinch model. The input data are trained for the first time, and then input into the abnormal behaviour recognition again to realise the abnormal recognition of learning behaviour. The recognition model constructed by this method has good recognition effect, but the normal data in the recognition can not be distinguished, resulting in some errors in the output results, which needs further improvement. Ye et al. (2020) proposed a semantic enhanced online learning behaviour prediction method. Firstly, this method analyses the learning state of learners, and designs an abnormal behaviour prediction method for this state. This method determines the vector of learners' short text with the help of memory network algorithm, then constructs an abnormal learning state recognition model according to the relevant characteristic data of learners' learning statistics, and inputs the above collected states into it to complete the recognition of online learning behaviour abnormalities. The recognition speed of this method is fast, but there are large recognition errors and some shortcomings. Nie et al. (2020) proposed a video abnormal behaviour recognition with deep learning. That extracts video images of learners and users, analyses the video data to learn the spatial dimension information, constructs the

Copyright © 2023 Inderscience Enterprises Ltd.

relationship model between video images by using deep learning algorithm, and completes the recognition method of users' abnormal learning behaviour. This method can improve the performance of abnormal behaviour detection, but the recognition time is long and has some shortcomings.

Based on the existing research methods, a new detection method is designed in this paper. The specific detection process designed by this method is: firstly, the behaviour data is represented in the frequent pattern item set, and the learner behaviour data is regarded as the target data in the region, which is divided into multiple non-overlapping space-time blocks to get the behaviour data preprocessing;

Secondly, set constraints for classification, and Lagrange function is introduced to solve the dual form of data.

Finally, the abnormal data are trained with the method of sequence minimum eigenvalue to determine the high-frequency characteristics of the abnormal data; and the distance between data is calculated to complete the high-frequency feature extraction of abnormal data; According to the extracted features, the high-frequency features are reconstructed to reduce the difference between the data. On this basis, the English online learning behaviour anomaly detection model is constructed. The high-frequency features of the abnormal data are input into the model, and the threshold of the abnormal feature data is calculated to complete the anomaly detection.

2 Data collection and classification of English online learning behaviour

2.1 Data collection and preprocessing of English online learning behaviour

This paper first collects online behaviour data. That is mainly includes behaviour data such as learning log, browsing track, study result, English video browsing volume and browsing duration (Nie et al., 2020). These behaviour data can reflect a learner's learning habits and behaviour state. For the above learning behaviour research, to collects learners' behaviour data with that help of frequent pattern mining method. The probability of frequent occurrence of this mining method in a large relational dataset can be a frequent behaviour dataset, which includes the relevant data of the research object. That was the certain connection between behaviour data, and the collection of data WAS completed according to the connection (Tao et al., 2019).

Set the expression of online behaviour data with frequent mode item set as follows:

$$A = \{a_1, a_2, \ldots, a_m\} \tag{1}$$

Among these, a_m represents the number of M terms of the behavioural data. m indicates the number of behaviour data.

According to the determined pattern set, set the transaction database of its behaviour data as:

$$D_i = \{d_1, d_2, \ldots, d_n\} \tag{2}$$

where D_i represents a set of learned behavioural feature items in i in the set of things items, and this value is unique in the set of behavioural feature frequent items. n represents the number of learned behavioural feature items.

When the set of behavioural feature items meets $B \neq \varphi$, and the number length in $B \in A$ is L. The length of L is 12.

At this time, the support of the total number of behaviour feature data in the English online learning behaviour pattern item set affects the final collection of data (Qiu and Li, 2018). Therefore, calculate the support of the total number of feature data in the frequent item set of behaviour features in the total number of things in the database as follows:

$$support(B) = \frac{\left|\{d \in D_i | B \in d\}\right|}{\left|\{d \in D_i\}\right|} \tag{3}$$

According to the data support of the online English learners behavioural data in the frequent term set determined above, to better collect the data from the online English learners, the learner behavioural data is regarded as the target data (Yu et al., 2019) in a region, divided into non-overlapping $N \times Q$ regions, and the temporal block $N \times Q \times n$ is determined. At this point, the learner behaviour data in different space-time blocks are aggregated together to obtain the data target set as:

$$K_{xyz} = \{K_{x_i y_i z}\} \tag{4}$$

where $(x_i y_i z)$ represents the central point of the behavioural data in different space-time blocks, and K_{xyz} represents the frame moment of the state behaviour at the current moment.

Considering that the learning behaviour characteristics of English online learners are in a changeable state, in order to eliminate its impact on data collection, for to complete this preprocessing with the help of threshold setting. The results are as follows:

$$K_p(d_2) = u \sum K_{xyz} \sum_{i=1}^{n} N \times Q \times n \tag{5}$$

Among them, u represents the behaviour data, and $K_p(d_2)$ is a pre-preprocessed data, Q represents the threshold.

In the behaviour data collection and preprocessing, the data was characterised in the frequent pattern item set, the support of the total number of characteristic data in the set is determined, and the learner behaviour data is regarded as the target data in a region, which is divided into multiple non-overlapping space-time blocks to complete the English online behaviour data collection, with the help of threshold setting, preprocess English online learning behaviour data to complete data collection and preprocessing.

2.2 Classification of English online learning behaviour data

On the basis of the above collected English online learning behaviour data, for get abnormal detection, it is necessary to classify the above English online behaviour data. Classify the normal data, and get the abnormal data, remove the

normal data and reduce the complexity of the detection work. The classification of general online behaviour data is mainly to classify all the data collected above into general data and suspected abnormal data. Therefore, to support vector machine method (Li et al., 2018b) to set the classification constraints to classify behaviour data.

Set the data training set of English online learning behaviour data to be classified as:

$$F = \{(a_1, b_1), \ldots, (a_n, b_n)\} \tag{6}$$

Among these, $a_n \in R$, $b_n \in [0, 1]$, represent different components of the data to be categorised.

According to the principle of maximising the classification interval of support vector machine, the classification constraints of different types of data are set as follows:

$$Y = g \min_{a,b} \frac{1}{2}\|p\|^2 + E\sum s_i \tag{7}$$

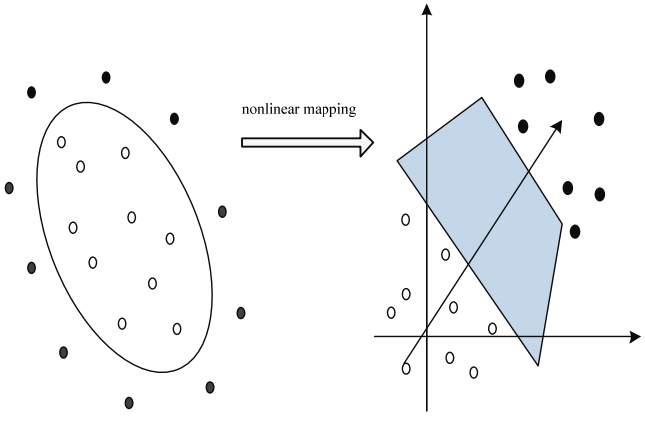
where p is the normal vector representing the optimal classification surface of the SVM, E represents the intercept between English behavioural data, s_i represents the probability that the data to be classified is assigned correctly, and g represents the penalty parameter of the error term.

It is necessary to solve the behaviour data to be classified by introducing Lagrange function (Zhao et al., 2019):

$$H(a, b) = \varepsilon \frac{1}{n}\|p\|^2 + E\sum 1 - s_i - \tau_i \tag{8}$$

where τ_i represents the non-negative vector values of the data to be classified, ε representing the Lagrange function multiplier.

Figure 1 Schematic diagram of behaviour data mapping process to be classified (see online version for colours)



According to the dual problem of the behaviour data to be classified (Wang et al., 2020), classify the corresponding behaviour data to be classified, and map each data to support vector machine. The mapping process is shown in Figure 1.

According to the above different mapped dimension spaces of the behaviour data to be classified, the calculation formula for classifying the English online learning behaviour data into general data is as follows:

$$\gamma = x_i - \sum_{i=1}^{n} v_i \alpha_i (x_i, x_j) \tag{9}$$

Among them, γ represents the general data classification results, and v_i represents the proportion.

The results of abnormal behaviour data classification are:

$$\delta = \sum_{i=1}^{n}\sum_{i=1}^{n} v_i \alpha_i (x_i, x_j)/b \tag{10}$$

Among them, the set of abnormal behaviour data is δ, α_i represents the kernel function, and b represents the classification plane copying decision function.

In the classification, support vector machine was used to set the constraints of classification, and Lagrange function is introduced to solve the duality of the behaviour data to be classified, so as to loss the complexity, complete the classification of English online learning behaviour data, and determine the abnormal data of online behaviour, It lays a foundation for effective detection of subsequent behaviour abnormalities.

3 Feature extraction-based behaviour anomaly detection method for English online learning

After the above classification, the abnormal data is determined to improve the effectiveness of users' English learning behaviour anomaly detection. This paper further uses the method of feature extraction to accurately detect English online learning behaviour. Feature extraction method mainly refers to the research object realised according to the relevant feature extraction algorithm. The algorithm covers a wide range, the research effect of data features is obvious, and its extraction accuracy and work efficiency are also high, so it has certain advantages. Therefore, to improve detection effect of English online learning behaviour abnormalities, this paper uses this method for detection research (Sun and Lu, 2018). First, train English online learning behaviour abnormal data with the help of the sequence minimum feature value method, treat the trained abnormal data results as a dictionary Z of a global analysis, and regard the characteristics of these abnormal data as the target block, set to f_i, at this time, represent the high frequency feature (Li et al., 2018a) of English online learning behaviour abnormal data under the feature extraction algorithm, obtained:

$$Z(x) = f_i \times \vartheta \tag{11}$$

Among them, $Z(x)$ represents the high-frequency features of the extracted anomalous data and ϑ represents the global analysis dictionary.

Therefore, it is necessary to cluster the high-frequency feature sample data of the above extracted abnormal data, and take a key data in the clustered high-frequency feature data as a core data to construct an optimal abnormal data extraction dictionary.

On this basis, to use K-means clustering algorithm, train the high-frequency features of the above extracted abnormal data. It is assumed that there are n target block features in the high-frequency features of the trained abnormal behaviour data, which need to be clustered. It is assumed that there are n target block features in the high-frequency features after clustering, and the feature set is:

$$U = \{u_1, u_2, \ldots, u_k\} \tag{12}$$

Among them, u_k represents the clustering centre of abnormal data and high-frequency characteristic data. k indicates the number of target block features

Then, the feature set obtained by each high-frequency feature data corresponding to an analysis dictionary is expressed as:

$$G = \{g_1, g_2, \ldots, g_k\} \tag{13}$$

The high frequency feature target block feature to be reconstructed is represented as s_i, and the Euclidean distance from all high frequency feature data is h_i calculated as:

$$S = R \arg\min \|s_i - u_k\| O^2 \tag{14}$$

Specifically, R argmin represents the absolute value of the Euclidean distance between the data and the O adaptive value of the high-frequency features.

Obtain the distance between each sample and the nearest cluster centre and select the maximum value based on the distance between each sample as far as possible.

In principle, this sample is selected as the new clustering centre of this cluster, so the clustering centre under each cluster can be calculated as:

$$S_0 = \arg\max \|S\| O^2 \tag{15}$$

According to the distance between the determined high-frequency features and abnormal data, considering the differences between the features of all abnormal data, reconstruct the high-frequency features to improve the accuracy of extraction. The characteristics of the reconstructed English online learning abnormal behaviour data are as follows:

$$\widehat{\aleph}_i = \arg\min \|s_i - u_k\| O^2 + \theta \|s_i - u_k\|^2 \tag{16}$$

Among these, $\widehat{\aleph}_i$ represent the reconstructed anomalous feature data, θ representing the target structural difference factor.

The above process is repeated, and the clustering centre is constantly moved until the clustering error function converges or reaches the maximum number of iterative steps. The squared error SSE function is:

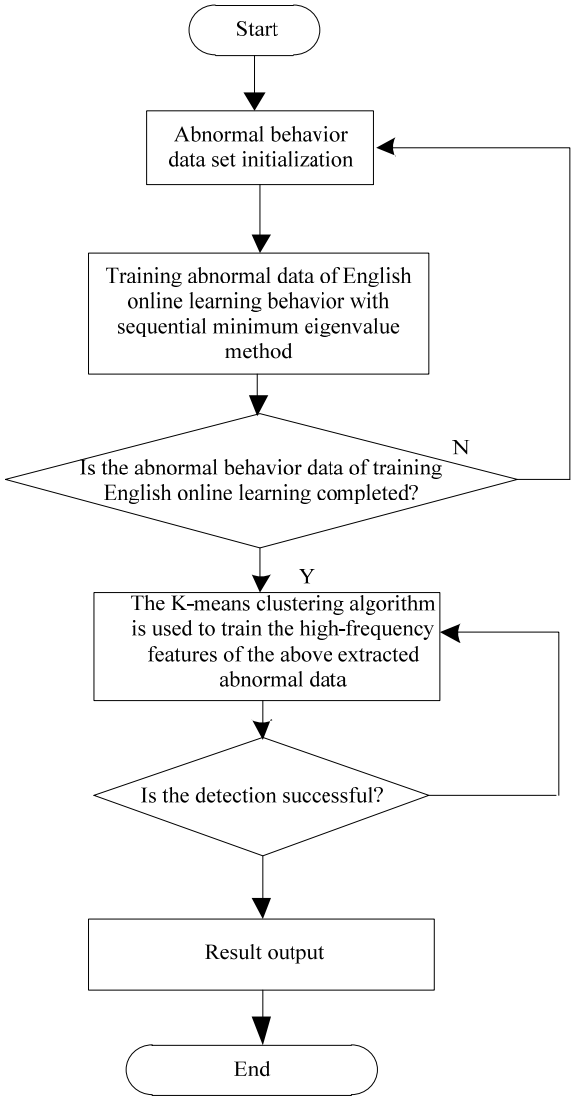
$$SSE = \sum_{i=1}^{n} \left(\widehat{\aleph}_i - u_k \right)^2 \tag{17}$$

According to the abnormal data extracted from the above features, the abnormal detection model is constructed, and to following results are obtained:

$$\beth(i) = \frac{1}{m} \sum_{i=1}^{n} \frac{\theta \|s_i - u_k\|^2}{\arg\min \|s_i - u_k\| O^2} \ln(T) \tag{18}$$

Among them, the English online learning behaviour anomaly data representing the input model, $\beth(i)$ the English online learning behaviour anomaly detection results, and T represents the threshold of abnormal behaviour detection of abnormal data. If this value is greater than T, it will be regarded as abnormal in the online behaviour. That implementation process to anomaly detection based on feature extraction is shown in Figure 2.

Figure 2 Implementation process of anomaly detection based on feature extraction



In the implementation of English online learning behaviour anomaly detection, the sequential minimum eigenvalue method is used to train behaviour anomaly data, the high-frequency characteristics of abnormal behaviour data, and complete the high-frequency feature extraction of abnormal data; According to the extracted features, the high-frequency features are reconstructed to reduce the difference between the data. On this basis, the English online learning behaviour anomaly detection model is

constructed. The high-frequency features were input to model, and threshold with the abnormal feature is calculated to complete the English online learning behaviour anomaly detection. This method introduced feature extraction algorithm, and successfully combined with other algorithms, has a simple idea, fast convergence speed, good clustering effect, so it can effectively reduce the detection time and detection error of abnormal behaviour data.

4 Experimental analysis

4.1 Scheme design

The experimental scheme is to determine the research object, initialise the abnormal behaviour data of the research object, extract the behaviour characteristics according to the behaviour data, detect the behaviour information of English learners, design performance indicators, and analyse the performance of the detection method through the indicators. In the test, 50 English learning users registered for more than 1 year in an English teaching platform were selected, including 20 males and 30 females. The data types of English learning users include learning performance, learning log, browsing track, English video views, browsing duration and other behavioural data. These behavioural data can reflect the learning habits and behaviour states of learners. According to their user behaviour characteristics, the abnormal behaviour of the subjects was determined

1 The duration of online browsing resources within three months was extracted. The normal browsing duration was 20-30 minutes, including more than 30 minutes for 30 people;

2 Half of the total duration of English online stay learning is 60 minutes, and the behaviour that exceeds the duration and lives shorter than the duration is regarded as abnormal behaviour;

3 Extract the online English test scores of the subjects, and their online test scores are less than 60 points, which is learning abnormal behaviour.

Initialise the abnormal behaviour dataset, as shown in Table 1.

Table 1 Data parameter settings

Serial number	*Parameters of the symbol*	*The parameter name*	*Set the value*
1	m	Number of behaviour data	12
2	n	The number of behavioural feature item sets learned	30
3	k	Target block characteristics	30
4	T	Abnormal behaviour detection threshold	60
5	L	Number and length of feature item sets	10

4.2 Experimental index design

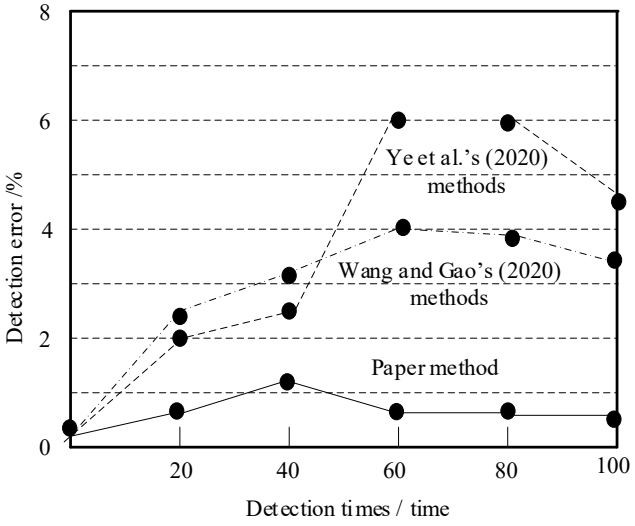
In order to verify the effectiveness of the proposed method, experiments are compared and analysed. The effectiveness of this method is verified by comparing the detection result error and detection time cost of the three methods.

4.3 Results

4.3.1 Error analysis of abnormal behaviour detection in different methods of English online learning

Firstly, the experiment analyses the error of this method, Wang and Gao's (2020) method and Ye et al.'s (2020) method in the detection of abnormal English behaviour with the sample research objects. The result is the accuracy of each experimental iteration. The error of the detection result is shown in Figure 3.

Figure 3 Abnormal detection errors of English online learning behaviour in different methods

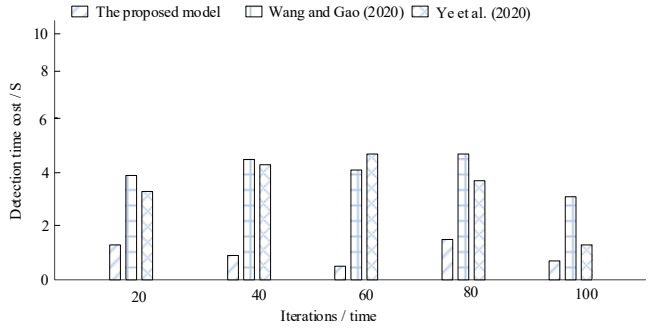


According to the curve above, using the methods of this paper, Wang and Gao (2020) and Ye et al. (2020) get the error of detecting has differences. Among them, detection error of English online learning behaviour abnormalities of the sample subjects using this method is always lower than that of the other two methods, and the minimum detection error is about 1%, while the detection error of the other two traditional methods is always higher than that of this method. In addition, most of the errors of the traditional methods are above 2.0%, among which the highest errors of 2222 and 111 are 6% and 4%, respectively. The errors of the literature methods are relatively large. In contrast, this method has better detection effect. This is because this method uses the feature extraction method to extract the features of abnormal behaviour data, and detects them according to the feature threshold, which improves the detection accuracy and reduces the detection error.

4.3.2 Analysis of time cost of abnormal behaviour detection in different methods of English online learning

The detection time is also an important part of measuring the feasibility of the method. Therefore, based on the above test, the detection time overhead of the three methods compared in this paper is tested, and the following results are obtained:

Figure 4 Comparison of time cost of abnormal behaviour detection of English online learning by different methods



By analysing the experimental results in Figure 4, there is a large gap in the time cost results of English online learning behaviour anomaly detection of the sample subjects by using the methods. Among them, under the background of different test times, the detection time overhead obtained by the three methods is quite different. Among them, under the background of different test times, the detection time overhead obtained by the three methods is quite different. However, from the overall trend of the image, the detection time cost of these three methods has changed to some extent, but the change is small. The method designed in this paper has a shorter time cost, and the detection time of the other two methods is greater than that of this method, the maximum detection time cost of the method in this paper is 1.8 s, and the minimum detection time of the method in the literature is 1.8 s. However, the detection time of the method in this paper is only 1.0 s when the number of tests in the literature is the least, with a difference of 0.8 s. Therefore, the detection speed of this method is faster.

5 Conclusions

An anomaly detection method was designed to solve the problems of large error and high cost of detection time in English online learning behaviour anomaly detection. In this method, the sequence minimum eigenvalue method is introduced to extract abnormal behaviour data. Before the data extraction, frequent pattern mining method and support vector machine are used to collect and analyse the data, and the abnormal detection method of feature extraction is optimised to complete the abnormal behaviour detection of English online learning. The experimental results show that the detection error of this method is reduced by more than 1.0% and the detection time cost is reduced by more than 0.8 s compared with the contrast method. Therefore, the detection time and error are effectively reduced. It provides convenient conditions for online English learning.

Acknowledgements

The research topic of this paper is achievements of ‘Research on the Integration of Higher Vocational Normal Education and Basic Education’ (No. 2018-JSJYZD-055) which belongs to the Fund Project – Key Research Project on Teacher Education Curriculum Reform in Henan Province: chaired by Feng Wei.

References

Bi, M., Wang, A-D., Xu, J. and Zhou, F-C. (2018) ‘Anomaly behavior detection of database user based on discrete-time Markov chain’, *Journal of Shenyang University of Technology*, Vol. 40, No. 1, pp.70–76.

Chen, H., Wang, G. and Song, J. (2018) ‘Research on anomaly behavior classification algorithm of internal network user based on cloud computing intrusion detection data set’, *Netinfo Security*, Vol. 15, No. 3, pp.1–7.

Du, G.Y. and Chen, M.J. (2018) ‘Research on anomaly detection algorithm of moving objects based on intelligent video analysis’, *Video Engineering*, Vol. 42, No. 12, pp.23–26.

Guo, Z., Peng, H., Niu, S., Shao, K., Lyu, Z. and Wang, W. (2018) ‘Analyzing user and network behaviors for host-based anomaly detection’, *Journal of Beijing Jiaotong University*, Vol. 42, No. 5, pp.40–46.

Li, H-B., Li, Q., Tang, R-M., Wu, J., Lv, Z-Y., Pei, D., Shi, J-J., Dong, X., Fang, S-D., Yang, Y-F. and WU, Y. (2018a) ‘User behavior anomaly detection for database based on unsupervised learning’, *Journal of Chinese Computer Systems*, Vol. 39, No. 11, pp.464–2472.

Li, S., Li, R-Q. and Yu, C. (2018b) ‘Evaluation model of distance student engagement： based on LMS data’, *Open Education Research*, Vol. 24, No. 1, pp.91–102.

Nie, H., Xiong, X., Guo, Y., Chen, X. and Zhang, S. (2020) ‘Video abnormal behavior identifying algorithm based on deep learning’, *Modern Electronics Technique*, Vol. 43, No. 24, pp.110–112, 116.

Qiu, H-W. and Li, X-Y. (2018) ‘Simulation of accurate detection of abnormal data in interactive network’, *Computer Simulation*, Vol. 35, No. 5, pp.375–378.

Sun, L-H. and Lu, Y. (2018) ‘Multivariate autoregression based algorithm for anomaly detection in rating data’, *Computer Engineering and Design*, Vol. 39, No. 6, pp.1629–1632, 1652.

Tao, T., Zhou, X., Ma, B. and Zhao, F. (2019) ‘Abnormal time series data detection of gas station by Seq2Seq model based on bidirectional long short-term memory’, *Journal of Computer Applications*, Vol. 39, No. 3, pp.924–929.

Wang, X., Feng, A., He, F., Ma, H. and Yang, J. (2020) ‘Research of database user behavior anomaly detection based on K-means and naive Bayes’, *Application Research of Computers*, Vol. 37, No. 4, pp.1128–1131.

Wang, Z-W. and Gao, B-P. (2020) 'Spatio-temporal fusion convolutional neural network for abnormal behavior recognition', *Computer Engineering and Design*, Vol. 41, No. 7, pp.2052–2056.

Ye, J-M., Luo, D-X., Chen, S. and Liao, Z-X. (2020) 'Semantic enhanced behavior prediction method for online learners', *Journal of Chinese Computer Systems*, Vol. 41, No. 1, pp.51–55.

Yu, L-H., Zhang, K., Cai, Y. and Jing, H-F. (2019) 'Abnormal behavior detection algorithm of moving target', *Computer Engineering and Design*, Vol. 40, No. 12, pp.3443–3450.

Zhao, H., Liu, Y., Li, S., Xu, P. and Zheng, Q. (2019) 'Online learning behavior based personality recognition', *Education Research*, Vol. 25, No. 5, pp.110–120.