Learners' Autonomous Learning Behavior in Distance Reading Based on Big Data

https://doi.org/10.3991/ijet.v17i09.31373

Shuliang Huan[1,2](✉), Cuibi Yang[3]
[1] College of Arts, Chongqing Three Gorges University, Wanzhou, China
[2] School of Humanities, Southwest Jiaotong University, Chengdu, China
[3] Institute of Public Administration, Chongqing Three Gorges University, Wanzhou, China
20020011@sanxiau.edu.cn

Abstract—Some learners couldn't get ideal learning results from distance education, to find out the root of this problem, their learning behavior and initiative should be analyzed in real time. However, in existing research results, the collected distance learning behavior data of learners are not pertinent enough, the extraction methods of big data are not proper enough, and the analysis models of learning behavior are not scientific enough. For these reasons, this paper analyzed the autonomous learning behavior of learners during distance reading based on big data, and gave the analysis contents and methods. In the text, a sequence diagram transformation method that is self-adaptive to long and short sequences had been introduced to process the learning behavior data of learners, who were then classified according to the features of different autonomous learning behavior sequences. Then, an attribute reduction algorithm based on improved Bayesian fuzzy rough set was adopted for attribute reduction, and the behavior indexes that are closely related to the autonomous learning effect of learners were selected for correlation analysis. After that, this paper proposed a method for detecting non-autonomous learning behavior based on multiple time scales, combining short-term autonomous learning pattern with long-term autonomous learning behavior and resource access behavior, this paper also analyzed the typical autonomous learning pattern of learners through the learning of hidden layer features. At last, experimental results proved the effectiveness of the proposed analysis method.

Keywords—big data, distance reading instruction, autonomous learning, behavior analysis, rough set

1 Introduction

As the learning requirements of learners are growing constantly, the traditional education mode is being updated accordingly, the emerging distance education is a convenient method but it has a few shortcomings in terms of interaction and supervision, so it has certain limitations in real applications of intensive training and practice, and the teaching of reading is a typical example [1-7]. As the processing technologies of big data are advancing in recent years, new modes of distance education are being developed further [8-12]. The newly developed distance education courses can provide multifarious learning resources and teaching assistant services, which can well adapt to the requirements of different learning scenarios [13-18]. However, some learners couldn't get ideal learning results from distance education [19-23]. During the course teaching process, some learners choose to give up halfway, and some can only keep a low learning efficiency. To solve these problems with distance education, it's necessary to analyze their learning behavior and initiative in real time.

Tjhin [24] studied the behavior pattern of online learning users in Indonesia, especially that of student users, to find the most suitable and invigorative learning resources for students while improving the difficulty of the learning materials at the same time. Their research data were collected from a private university in Indonesia, 730 student users were selected from 3870 online course users and taken as samples. At present, distance education and e-learning are popular technologies, and the use of the monitoring system allows teachers to evaluate students' learning status; Dow et al. [25] carried out experiment to demonstrate the feasibility and effectiveness of the monitoring system, and the results showed that with the help of this system, students' academic performance had been improved, abnormal learning phenomena became less, and the monitoring system had provided teachers with a simple interface to instruct their students and improve their academic performance. In recent years, the Internet has become an important learning environment for modern students, and it has overcome the constraints of time and space, Chang [26] constructed an EL-RFM model to measure learners' learning motivation, based on the results of the model, teachers can formulate better learning strategies to attract students' interest in learning. Based on the learning data of cadre platform courses of Chongqing Electric Power College, Chen [27] studied the relationship between the characteristics of students and their online learning behavior, the differences in the learning behavior of cadres were analyzed from several aspects including age, gender, title, unit, number of logins, number of learning, learning completion situation, correlation between gender, title, and elective courses, and correlation between age, title, and elective courses. Umbleja and Ichino [28] constructed a model for predicting students' final exam performance and completion time based on students' behavior in the past; the modeling process applied the data mining technology, regression analysis, principal component analysis, and the hierarchical clustering of the data of the symbol histogram.

After reviewing relevant literatures, it's found that the research on the learning behavior analysis of learners in the big data environment is still in an initial stage. the collected distance learning behavior data of learners are not pertinent enough, the extraction methods of big data are not proper enough, and the analysis models of learning behavior are not scientific enough. For these reasons, this paper analyzed the autonomous learning behavior of learners during distance reading based on big data. The second chapter gave the analysis contents and methods. The third chapter introduced a sequence diagram transformation method that is self-adaptive to long and short sequences to process the learning behavior data of learners, who were then classified according to the features of different autonomous learning behavior sequences. The fourth chapter adopted an attribute reduction algorithm based on improved Bayesian fuzzy rough set for attribute reduction, and selected behavior indexes that are closely related to the autonomous learning effect of learners for correlation analysis. The fifth chapter proposed a method for detecting non-autonomous learning behavior based on multiple time scales, combined the short-term autonomous learning pattern with long-term autonomous learning behavior and resource access behavior, and analyzed the typical autonomous learning mode of learners through the learning of hidden layer features. At last, experimental results proved the effectiveness of the proposed analysis method.

2 Contents and methods of behavior analysis of learners' autonomous learning in distance reading

Learners' learning behavior on distance reading instruction platforms includes various activities such as: select reading content, carry out teaching links, review and take exams, and reflect and summarize, specifically, the activities include: read texts and browse catalogue, teach reading content and students' learning experience, interpret reading texts according to learning situations, understand the teaching process, select course, withdraw course, arouse and supplement students' reading ability, instruct students to master new reading skills, organize students to communicate and share reading and learning experiences, take class notes, take online notes, post on forum, finish homework, review the teaching content of reading appreciation, understand reading rules and strategies, master the reading method of combining reading with writing, take exams and evaluate, understand "language tastes" and "reading with emotions", as shown in Figure 1.

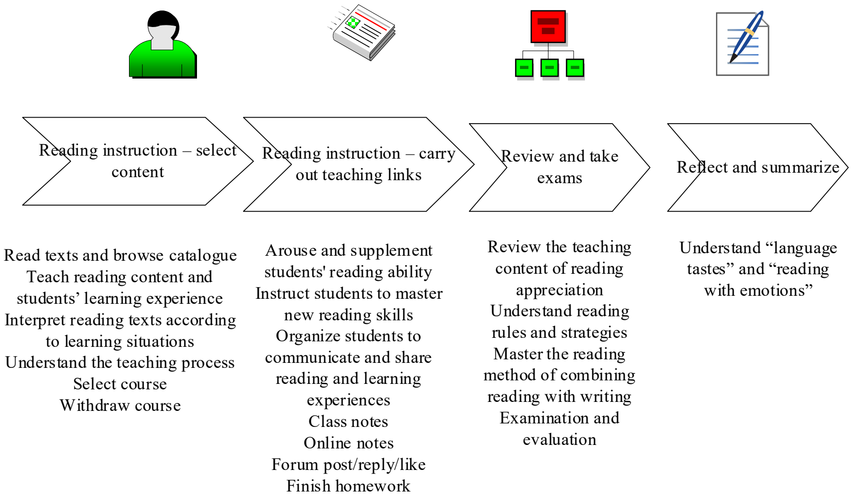

Fig. 1. Learning behavior on distance reading instruction platforms

This paper discussed the driving force of autonomous learning behavior analysis from three angles: analysis method, teaching subjects of distance reading instruction, and autonomous learning behavior and resources, as shown in Figure 2. Figure 3 shows the specific research contents and methods of learners' autonomous learning behavior in distance reading based on big data, see below.

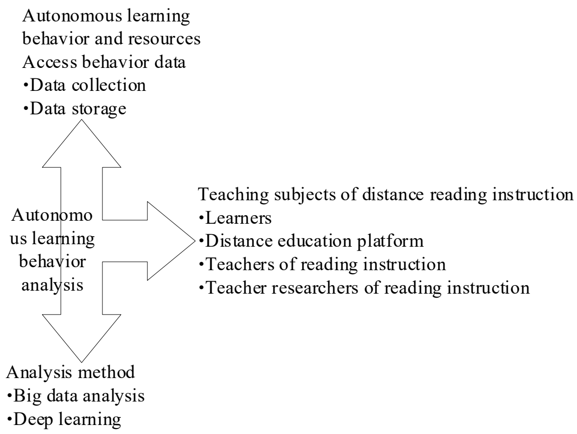

Fig. 2. The driving force of autonomous learning behavior analysis of distance reading learners

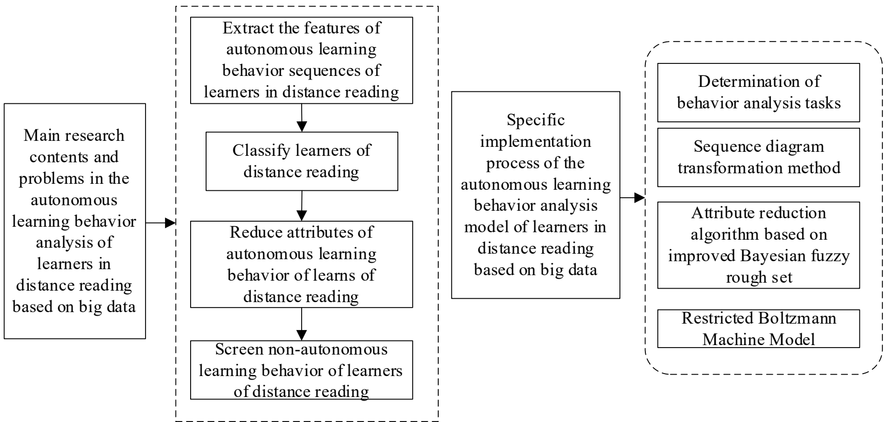

Fig. 3. Research contents and methods

3 Classification of distance reading learners

After obtaining the data of the autonomous learning behavior and resource access behavior of distance reading learners in a certain learning stage, the learners' learning styles, behavior features, group behavior similarity, and other behavior patterns could be summarized and analyzed preliminarily, however, the analysis results couldn't support real-time online processing of distance education, and the behavior distribution of similar groups couldn't be mastered quickly. The objective of this study is to reasonably classify distance reading learners with different autonomous learning behavior sequence features. Since there're great differences in the length of the behavior sequences of learners, in order to meet the requirements of fast response for unsupervised tasks, this paper introduced a sequence diagram transformation method that is self-adaptive to long and short sequences to process the behavior data of learners.

Assuming: E represents a dataset of learner behavior sequences; in the dataset, e represents a random behavior sequence composed of letter set u, and it satisfies e∈E; behavior sequence e may contain one or more letters in u, for example, E is composed of elements in Φ={A, B, C, D, E}; wherein the length of e is denoted as K(e), and the size of K(e) is the number of learner behavior events in e; in e, the value of the letter at position k in the sequence is denoted as ek, wherein k=1,...,K(e) and ek∈Φ.

In order to extract the information of the relation between letter elements, this paper adopted the sequence diagram transformation method to extract the relative position information of the elements. Assuming: ξ(e)vu represents the relationship between elements, the elements that need to be extracted satisfy v∈Φ and u∈Φ; η is a hyper-parameter with element distance λ as input and ρ as adjuster, which can be denoted as ηρa(λ), and it can measure the size of the impact of two letter element positions; ξ is a function of η. Taking η(λ(k,n)) as an example, k and n are the positions of learner behavior events, and λ(k,n) is the distance measure between positions k and n. Figure 4 shows the geometric meaning of function η(λ(k,n)).

In order to apply the sequence diagram transformation algorithm, function η needs to satisfy three conditions: $\eta_\rho a(\lambda)>0$, $\partial\eta_\rho a(\lambda)/\partial\lambda<0$, $\partial\eta_\rho a(\lambda)/\partial\rho<0$. This paper selected the exponential function as the expression form of function η, that is:

$$\eta_\rho\left(\lambda(k,n)\right)=\exp\left(-\rho|n-k|\right),\forall\rho>0,\lambda>0 \tag{1}$$

During the execution of the sequence diagram transformation algorithm, the number and position of the paths between two pending learner behavior events were counted first and then stored. The storage space was an asymmetric matrix Γ with a size of $|\Phi|\times|\Phi|$. The relative position information of all five (v, u) pairs was stored in Γ_{vu}, in all these stored position relationships, v and u were always in this sequence that u is behind v, then there is:

$$\Gamma_{vu}(e)=\left\{(k,n):e_k=v,e_n=u,k<n,(k,n)\in 1,...,K^{(e)}\right\} \tag{2}$$

Figure 5 shows the visualization of the relative positions of function $\eta(\lambda(k, n))$. In Γ_{vu}, the relative position information of each (v, u) pair would affect relationship ξ between elements v and u. When applying the sequence diagram transformation algorithm, in order to obtain the ξ function of the feed-forward behavior sequence e, this paper clustered and standardized the values of the distance measure function of each pair of letter element positions. Assuming $\xi(e)=[\xi_{vu}(e)]$ represents the applied sequence diagram transformation features of the sequence, and it satisfies v, $u \in \Phi$, then, if the feature extraction result is sensitive to the length of the autonomous learning behavior sequences of learners, then there is:

$$\xi_{vu}(e) = \frac{\sum_{\forall(k,n)\in\Gamma_{vu}(e)} \exp(-\rho|n-k|)}{|\Gamma_{vu}(e)|} \tag{3}$$

If the feature extraction result is not sensitive to the length of the autonomous learning behavior sequences of learners, then there is:

$$\xi_{vu}(e) = \frac{\sum_{\forall(k,n)\in\Gamma_{vu}(e)} \exp(-\rho|n-k|)}{|\Gamma_{vu}(e)|/K^{(e)}} \tag{4}$$

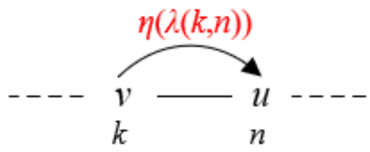

Fig. 4. The geometric meaning of function $\eta(\lambda(k,n))$

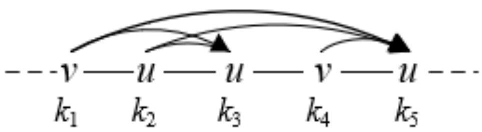

Fig. 5. Visualization of the relative positions of function $\eta(\lambda(k,n))$

4 Reduction of the autonomous learning behavior attributes of distance reading learners

In the process of distance reading instruction, since the learning behaviors of learners have the features of independent, objective, and ambiguous, and there is no obvious connection between the autonomous learning behavior and the resource access behavior, it's necessary to use fuzzy rough sets to reduce the attributes of this kind of learning behavior of learners. Similar to the traditional rough set models, fuzzy rough sets are also easily affected by noise data. In order to improve the anti-interference ability of fuzzy rough sets, this paper adopted an attribute reduction algorithm based on improved Bayesian fuzzy rough sets to reduce the attributes of the autonomous learning behavior indexes of distance reading learners, and the correlation of the behavior indexes that are closely related to the autonomous learning effect of learners was analyzed.

The inclusion degree of classical sets $|ID(X\cap Y)=|X\cap Y/|X||$, $|X|\neq Ø$ describes the proportion of elements contained both in sets X and Y in set X.

Let $Ω$ be a finite non-empty set, Q=[0,1], ζ is any implication, the mapping $W:G(Ω)\times G(Ω)\rightarrow Q$ is called the fuzzy inclusion degree, wherein $Q\forall X,Y\in G(Ω)$:

$$W(X,Y)=\begin{cases}\dfrac{\sum\limits_{a\in E(X)}\zeta\left(X(a),Y(a)\right)}{\left|E(X)\right|}, X\neq\varphi \\ 1, X=\varphi\end{cases} \tag{5}$$

$W(X,Y)$ is called the fuzzy inclusion degree (ζ) of X with respect to Y, wherein for ζ, it's selected that $\zeta(X(a),Y(a))=1\wedge(1\text{-}X(a)+Y(a))$, and $E(X)=\{a\in Ω|X(a)>0\}$ is a support set of fuzzy set X.

Let $Ω$ be a finite non-empty universe of discourse, $\chi=\{D_1,D_2,\ldots,D_l\}$ is a fuzzy cover of $Ω$, and it is also a division of $Ω$ about $D_i(i=1,2,...,l)$; W is the fuzzy inclusion degree, $R=(G(Ω), \chi, W)$ is called a fuzzy inclusion approximation space. $\forall A\in G(Ω)$ takes $O(\chi_1)=|\chi_1|/|Ω|$ and $O(A)=max_i(O(\chi_1))$, wherein i=1, 2,...,s and $O(A)\in(0.5,1)$, then, the lower approximation $R_{O(A)}(A)$ is $R_{O(A)}(A)=\cup\{D_i\in\chi|W(D_i,A)\geq O(A)\}$, and the upper approximation $R_{1\text{-}O(A)}(A)$ is $R_{1\text{-}O(A)}(A)=\cup\{D_i\in\chi|W(D_i,A)>1\text{-}O(A)\}$.

The positive domain of the improved Bayesian fuzzy rough set model is defined as $AR^{+}(A)=R_{O(A)}(A)=\cup\{D_i\in\chi|W(D_i,A)\geq O(A)\}$; the boundary domain is defined as $AR^{BO}(A)=\cup\{D_i\in\chi|1\text{-}O(A)<W(D_i,A)<O(A)\}$, and the negative domain is defined as $AR^{-}(A)=\cup\{D_i\in\chi|W(D_i,A)\leq 1\text{-}O(A)\}$.

Let $FD=(Ω,X,P,g)$ be a fuzzy information decision-making system, wherein $Ω=\{a_1,a_2,...,a_m\}$. Fuzzy attribute set $Y=\{y_1,y_2,....,y_n\}$, $Y\subseteq X$, then the similarity of elements a_i and a_j with respective to attribute set Y is $S_Y=(a_i,a_j)=1\text{-}(1/n\sum^{n}_{l=1}(a_{il},a_{jl})^2)^{1/2}$, where a_i and a_j are regarded as fuzzy sets on attribute set Y; a_{il} represents the membership degree of element a_i to fuzzy attribute y_l, S_Y is called the fuzzy similar relationship of Y, the similarity class $[a_i]_{SY}$ of a_i can be denoted as $[a_i]_Y$, wherein i=1,2,...,m, it is called the fuzzy information grain induced by fuzzy similar relationship SY or the attribute set Y.

Let $FD=(Ω,X,P,g)$ be a fuzzy information decision-making system, $Ω=\{a_1,a_2,...,a_m\}$, $X=D\cup W$; D is a conditional attribute set; W is a decision attribute set; $Y\subseteq D$; S_Y and S_W are respectively the fuzzy similar relationships of Y and W; there are $Ω/Y=\{Y_1,Y_2,...,Y_m\}$ and $Ω/W=\{W_1,W_2,...,W_m\}$, wherein $Y_i=[a_i]_Y$, $W_j=[a_j]_W$, and i,j=1,2,...,m. $\chi=\{D_1,D_2,\ldots,D_l\}$ is a fuzzy cover of $Ω$; W^G is the fuzzy set inclusion degree on $Ω$, and $R=(G(Ω),\chi,W)$ is called the fuzzy inclusion approximation space, $\forall A\in G(Ω)$ takes $O(A)=max_i(O(\chi_i))$, wherein i=1,2,...,m and $O(A_l)\in(0.5,1)$; then, the $O(A)$ lower approximate distribution $Y_{O(A)}(W)$ and the 1-$O(A)$ upper approximate distribution $Y_{1\text{-}O(A)}(W)$ of decision attribute set W with respect to the conditional attribute set Y are respectively $Y_{O(A)}(W)=\{Y_{O(A)}(W_1),Y_{O(A)}(W_2),...,Y_{O(A)}(W_m)\}$ and $Y_{1\text{-}O(A)}(W)=\{Y_{1\text{-}O(A)}(W_1),Y_{1\text{-}O(A)}(W_2),...,Y_{1\text{-}O(A)}(W_m)\}$.

Where, $Y_{O(A)}(W_j)=\cup\{Y_i\in Ω/Y|W^G(Y_i,W_j)\geq O(A)\}$, and $Y_{1\text{-}O(A)}(W_j)=\cup\{Y_i\in Ω/Y|W^G(Y_i,W_j)\geq 1\text{-}O(A)\}$. $AR^{BO\text{-}O(A)}{}_Y=Ω_jY_{O(A)}(W_j)$ is called the $O(A)$ positive domain of W with respect to Y, $\alpha^{O(A)}{}_Y(W)=SR(AR^{BO\text{-}O(A)}{}_Y(W))/|\Omega$ is called the $O(A)$ dependency degree of $|W|$ on Y, or it can be called the $O(A)$ support degree of Y to W.

Let Ω be a non-empty universe of discourse, $\Omega=\{a_1,a_2,...,a_m\}$, $R=(G(\Omega),\chi,W)$ is a fuzzy inclusion approximation space, $X,Y\in G(\Omega)$, $\omega(X,Y)=(1/m\Sigma^m{}_{i=1}(X(a_i)\text{-}Y(a_j))^2)^{1/2}$ is the degree of separation between X and Y, and $0\leq\omega(X,Y)\leq 1$.

The relative importance degree is defined below using the degree of separation. Let $FD=(\Omega,X,P,g)$ be a fuzzy information decision-making system, $X=D\cup W$, D is a conditional attribute set, W is a decision attribute set, $Y\subseteq D$, ω is the degree of separation, $O(A)\in(0.5,1)$, then the $O(A)$ importance degree of $y\in Y$ relative to W in Y is $IMP_1{}^{O(A)}$ $(y,Y,W)=imp_j\{\omega(Y_{O(A)}(W_j)\},Y\text{-}y_{O(A)}(W_j))\}$, the $O(A)$ importance of $d\in D$ relative to W in D is $IMP_1{}^{O(A)}(d,D, W)=imp_j\{\omega(D_{O(A)}(W_j)\},D\text{-}d_{O(A)}(W_j))\}$ and the $O(A)$ importance of $y\in D\text{-}Y$ relative to W with respect to Y is $IMP_2{}^{O(A)}$ $(y,Y, W)=imp_j\{\omega(Y\cup y_{O(A)}(W_j),Y_{O(A)}(W_j))\}$. The totality of $O(A)$ necessary attributes with a precision τ in D relative to W can be called the τ-precision core of D relative to W, namely $PK_{O(A)}(D,W,\tau)$.

Let $FD=(\Omega,X,P,g)$ be a fuzzy information decision-making system, $X=D\cup W$, D is a conditional attribute set, W is a decision attribute set, $Y\subseteq D$, $0<\tau<1$, $O(A)\in(0.5,1)$. If Y satisfies $imp_j\{\omega(D_{O(A)}(W_j)\}, Y_{O(A)}(W_j))\}\leq\tau$, $imp_i{}^{O(A)}(y,Y,W)=imp_j\{\omega(Y_{O(A)}(W_j),Y\text{-}y_{O(A)}(W_j))\}>\tau$, then Y is called the 2-precision approximate reduct of D relatve to W.

5 The screening of non-autonomous learning behavior of distance reading learners

The screening of non-autonomous learning behavior can help distance education platforms quickly locate the potential non-autonomous learning behavior and determine whether a learner has such behavior. If the learner does have the non-autonomous learning behavior, the teacher of distance reading can quickly detect it and help the platform to correct the learning pattern of the learner and recommend appropriate reading resources to the learner. As for the autonomous learning behavior, the reasons for the changes in the abnormal behavior of learners could be analyzed to provide support for subsequent resource recommendation and decision-making of reading instruction services. To this end, this paper set up a non-autonomous learning behavior detection scenario, aiming to filter the learners through the big data analysis of learning behavior and help distance education platforms and teachers to carry out future works. This paper proposed a non-autonomous learning behavior detection method based on multiple time scales, combining short-term autonomous learning pattern with long-term autonomous learning behavior and resource access behavior, this paper analyzed the typical autonomous learning pattern of learners through the learning of hidden layer features. In the text, the learners were divided from coarse-grained and fine-grained perspectives, and combined with the preset outlier index to screen the non-autonomous learning behavior of learners.

The monthly data of learners' autonomous learning behavior were subjected to short time scale analysis, the semester data of learners' autonomous learning behavior were combined with the results of short time scale and then subjected to the long time scale analysis.

$$ET = \{ET_1, ET_2, ET_3, ..., ET_m\} \tag{6}$$

$$ET_i = \{BR_1, BR_2, BR_3, ..., BR_{30}\} \tag{7}$$

$$BR_j = \{AU_1, AU_2, AU_3, ..., AU_{120}\} \tag{8}$$

In the short time scale analysis, this paper adopted the restricted Boltzmann machine for hidden variable learning, so as to obtain the features of the non-autonomous learning behavior of learners which are usually difficult to find. The training input was a 120-dimensional vector of the daily data of learners' autonomous learning, and the short time scale analysis was adopted further for the cluster analysis based on the n-dimensional hidden features output by the restricted Boltzmann machine.

$$F_i\left(BR_j\right) = \{F_1, F_2, F_3, ..., F_n\} \tag{9}$$

Assuming: r and f respectively represent the state vectors of the visible layer and the hidden layer of the restricted Boltzmann machine model; ES represents all training samples; the energy model RM was selected and the maximum KE was solved. To get the probability distribution of variables, suitable energy function could be selected based on the restricted Boltzmann machine model, and then the maximum likelihood estimation could be solved:

$$R_K\left(u, f\right) = -\sum_{i=1}^{m_u} x_i u_i - \sum_{j=1}^{m_f} y_j f_j - \sum_{i=1}^{m_u} \sum_{j=1}^{m_f} f_j q_{ij} u_i \tag{10}$$

$$\ln K_e = ln \prod_{i=1}^{m_e} O\left(u^i\right) = \sum_{i=1}^{m_e} \ln O\left(u^i\right) \tag{11}$$

This paper used the restricted Boltzmann machine model to learn the data of short time autonomous learning behavior and extract the learner's autonomous learning pattern from more angles. After the categories of autonomous learning behavior had been attained, the trained restricted Boltzmann machine model was reversely called for the features of category cluster centers, and the monthly data curve of the learners' learning behavior were learned through the reconstruction of hidden layer features, and the typical pattern of learners' autonomous learning behavior for one week was obtained.

Assuming: AV_{mi} represents the average distance from a normal point of autonomous learning behavior to the data points in a certain category; $AV_{Y\text{-}min}$ represents the minimum value of the average distance from a normal point of autonomous learning behavior to the data points in other categories; and $AV_{N\text{-}min}$ represents the minimum value of the average distance from an abnormal point of non-autonomous learning behavior and all data points in different categories, in order to reduce the influence of the abnormal outlier points of non-autonomous learning behavior on the clustering results, this paper defined an outlier index as shown in Formula 12:

$$QU = \frac{AV_{N-\min} + AV_{Y-\min} - AV_{mi}}{Max\left(AV_{N-\min}, AV_{Y-\min}, AV_{mi}\right)} \tag{12}$$

The larger the $AV_{N\text{-}min}$ and $AV_{Y\text{-}min}$, the smaller the AV_{mi}, and the larger the outlier index of the learning behavior data of learners, which means that the distance from abnormal point of non-autonomous learning behavior and the category cluster is farther, and the constructed model showed a better effect in screening non-autonomous learning behavior.

6 Experimental results and analysis

Figure 6 gives a visual comparison of the clustering results of distance reading learners. Under the condition of the same cluster number, the sequence diagram transformation method was used to extract the position information of the elements. Compared with other algorithms, the sequence diagram transformation method can better distinguish and express the features when being applied to the feature extraction of the autonomous learning behavior sequences of learners, and it can better adapt to the learning behavior sequences of different lengths. Therefore, with the help of this method, we can better classify the unlabeled learning behavior sequence data, and find more valuable types in behavior patterns such as learning style, behavior feature, and group behavior similarity.

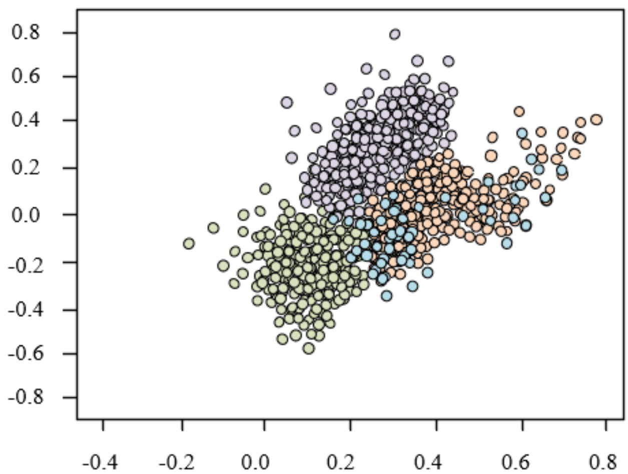

Fig. 6. Visual comparison of the clustering results of distance reading learners

Then, this paper compared the restricted Boltzmann machine model with other clustering methods, including K-Means clustering, mean-shift clustering, and

DBSCAN clustering. Table 1 shows the performance comparison of these clustering algorithms. As can be seen from the data, the restricted Boltzmann machine model proposed in this paper outperformed other methods in terms of *V_measure*, AMI, ARI, and outlier index. The algorithm proposed in this paper was implemented based on hierarchical clustering. For the autonomous learning behavior of distance reading learners, the effect was relatively ideal in single classification scenario, but when the autonomous learning behavior was sub-divided further, the restricted Boltzmann machine model can cluster the outlier points of non-autonomous learning behavior to another cluster, therefore, if the influence of the outlier values of non-autonomous learning behavior is taken into consideration, the restricted Boltzmann machine model shows an obvious advantage in terms of the score of the outlier index.

Table 1. Performance comparison of different clustering algorithms

Algorithm	Mean Shift Clustering	K-MEANS Clustering	DBSCAN Clustering	The proposed method
Cluster number	15	13	11	12
V_measure	0.7485	0.7628	0.8692	0.8152
AMI	0.7629	0.7184	0.6295	0.7852
ARI	0.3268	0.3928	0.6174	0.6283
Outlier index	0.1526	0.4157	0.3292	0.8475

When reclassifying the coarse-grained clustering results of autonomous learning behavior of distance reading learners based on the restricted Boltzmann machine model, the outlier points that are far from the cluster centers could also be screened according to the density of data points, and the outlier values could be defined as well. Table 2 summarizes the abnormal points of non-autonomous learning behavior detected in each category cluster and the information of possible causes. A total of 112 abnormal learners had been filtered out by the restricted Boltzmann machine model, accounting for 9.51% of all learners. In all distance reading learner clusters, more or less, there were abnormal learners, and the main causes included learning motivation, self-efficacy, and distance education mode.

Table 2. Proportions of the causes of abnormal points of non-autonomous learning behavior

Abnormal result	Abnormal point	Learning motivation	Self-efficacy	Distance education mode
Cluster 1	2	1	3	2
Cluster 2	5	3	1	4
Cluster 3	9	7	5	2
Cluster 4	3	1	5	8
Cluster 5	22	8	2	4
Cluster 6	16	12	9	5
Cluster 7	19	6	2	8
Cluster 8	13	9	5	1
Cluster 9	18	14	1	3
Cluster 10	5	8	1	1
Total (%)	100%	63.5%	22.4%	14.1%

After abnormal learners were screened out, for different causes of abnormities, the corresponding intervention measures had been taken, and Table 3 compares the score of autonomous learning of learners before and after the intervention. As shown in above table, before intervention, for the question of whether the learners have made their own reading plans, the average score was 2.14, after the intervention, this score was 3.62. It can be seen that in terms of autonomously supplementing reading plans, the learners showed obvious improvement, the t-value was -12.47, the p-value was 0.000, which was less than 0.01, indicating that the statistical results were significantly different. Before and after the intervention, for the question of whether the learners have set reading goals based on their own situations, the score changed from 2.95 to 3.74, the t-value was -9.638, and the p-value was 0.000, the difference was also significant. Before and after the intervention, for the question of whether the learners have adjusted their reading plans, the score changed from 2.71 to 4.85, the t-value was -1.36, and the p-value was 0.001, the difference was also statistically significant. As intervention measures had been implemented, the learners' score of reading time planning reached 3.2 points, indicating that their learning efficiency in reading had been improved obviously, the t-value was -12.629, the p-value was 0.001, indicating that the difference was statistically significant, too. In terms of the question of whether the learners have set learning goals based on reading proficiency improvement, the average score had increased by 1.37 points, the t-value was -11.482, the p-value was 0.000, the difference was significant as well.

Table 3. Comparison of the scores of autonomous learning of learners before and after intervention

Question	**Do you have your own reading plan?**		**Have you set reading goals based on your own situation?**		**Have you adjusted your reading plan?**		**Have you planned reading time for yourself?**		**Have you set learning goals based on reading proficiency improvement?**	
Item	Pre-test	Post-test	Pre-test	Post-test	Pre-test	Post-test	Pre-test	Post-test	Pre-test	Post-test
Mean	2.14	3.62	2.95	3.74	2.71	4.85	2.63	3.2	2.58	3.95
Standard deviation	0.63	1.25	0.95	1.48	0.75	1.37	0.51	1.84	0.69	1.38
Average difference	-1.58		-0.85		-1.36		-1.74		-1.86	
T	-12.47		-9.638		-13.847		-12.629		-11.482	
P	0.000		0.000		0.001		0.001		0.000	

7 Conclusion

This paper analyzed the autonomous learning behavior of distance reading learners based on big data, gave the analysis contents and methods, introduced a sequence diagram transformation method to process learner behavior data, and reasonably classified the learners with different autonomous learning behavior sequence features. In the text, an attribute reduction algorithm based on improved Bayesian fuzzy rough sets was adopted for attribute reduction, and the behavior indexes were selected for correlation analysis. Then, this paper proposed a method for detecting non-autonomous learning behavior based on multiple time scales, and analyzed the typical autonomous learning pattern of learners through the learning of hidden layer features. In the experimental results, this paper compared the clustering results of distance reading learners, and the performance of different clustering algorithms, and data proved that the proposed model outperformed other methods in terms of *V_measure*, AMI, ARI, and outlier index. After that, the causes of the abnormal points of non-autonomous learning behavior and their proportions were given in a table; a total of 112 abnormal learners had been filtered out, and the main causes included learning motivation, self-efficacy, and distance education mode. The scores of the learners' autonomous learning before and after the intervention were given and subjected to significance test, and the results demonstrated that the differences in learners in 5 aspects before and after intervention were statistically significant.

8 Acknowledgment

This paper were suppported by Higher education teaching reform research project in Chongqing Municipal Education Commission Collaborative Research on the practice teaching optimization and Evaluation Mechanism Reform of Literature Reading Appreciation Course Group from the comparison perspective of Chinese and Western (Grant No.: 213278).

9 References

[1] Haerazi, H., Irawan, L.A. (2020). The Effectiveness of ECOLA Technique to Improve Reading Comprehension in Relation to Motivation and Self-Efficacy, International Journal of Emerging Technologies in Learning, 15(1): 61-76. https://doi.org/10.3991/ijet.v15i01.11495

[2] Chen, B., Yin, H. (2018). Learning category distance metric for data clustering. Neurocomputing, 306: 160-170. https://doi.org/10.1016/j.neucom.2018.03.048

[3] Marcelino, M.J., Pessoa, T., Vieira, C., Salvador, T., Mendes, A.J. (2018). Learning computational thinking and scratch at distance. Computers in Human Behavior, 80: 470-477. https://doi.org/10.1016/j.chb.2017.09.025

[4] Pan, X.Q., Zhang, Z.X. (2020). An Empirical Study of Application of Multimodal Approach to Teaching Reading in EFL in Senior High School, International Journal of Emerging Technologies in Learning, 15(2): 98-111. https://doi.org/10.3991/ijet.v15i02.11267

[5] Fu, L., Li, Z., Ye, Q., Yin, H., Liu, Q., Chen, X., Yang, G. (2020). Learning robust discriminant subspace based on joint L2, p-and L2, s-Norm distance metrics. IEEE Transactions on Neural Networks and Learning Systems, 33(1): 130-144. https://doi.org/10.1109/TNNLS.2020.3027588

[6] Gong, A., Huang, L., Shi, J., Liu, C. (2022). Unsupervised RGB-T saliency detection by node classification distance and sparse constrained graph learning. Applied Intelligence, 52(1): 1030-1043. https://doi.org/10.1007/s10489-021-02434-y

[7] Ye, H.J., Zhan, D.C., Jiang, Y. (2019). Fast generalization rates for distance metric

learning. Machine Learning, 108(2): 267-295. https://doi.org/10.1007/s10994-018-5734-0
[8] Utomo, M.N.Y., Sudaryanto, M., Saddhono, K. (2020). Tools and strategy for distance learning to respond COVID-19 pandemic in Indonesia. Ingénierie des Systèmes d'Information, 25(3): 83-390. https://doi.org/10.18280/isi.250314
[9] Yoshida, T., Takeuchi, I., Karasuyama, M. (2018). Safe triplet screening for distance metric learning. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 653-2662. https://doi.org/10.1145/3219819.3220037
[10] Zhai, D., Liu, X., Chang, H., Zhen, Y., Chen, X., Guo, M., Gao, W. (2018). Parametric local multiview hamming distance metric learning. Pattern Recognition, 75: 250-262. https://doi.org/10.1016/j.patcog.2017.06.018
[11] Lipor, J., Wong, B.P., Scavia, D., Kerkez, B., Balzano, L. (2017). Distance-penalized active learning using quantile search. IEEE Transactions on Signal Processing, 65(20): 5453-5465. https://doi.org/10.1109/TSP.2017.2731323
[12] Sun, Y., Chai, R.Q. (2020). An early-warning model for online learners based on user portrait. Ingénierie des Systèmes d'Information, 25(4): 535-541. https://doi.org/10.18280/isi.250418
[13] Alias, U.F., Ahmad, N.B., Hasan, S. (2017). Mining of E-learning behavior using SOM clustering. In 2017 6th ICT International Student Project Conference (ICT-ISPC), 1-4. https://doi.org/10.1109/ICT-ISPC.2017.8075350
[14] Khlifi, Y., El-Sabagh, H.A. (2017). A novel authentication scheme for E-assessments based on student behavior over e-learning platform. International Journal of Emerging Technologies in Learning, 12(4): 62-89. https://doi.org/10.3991/ijet.v12i04.6478
[15] Hayashi, T., Wang, Y., Kawai, Y., Sumiya, K. (2016). An active communication mechanism for e-learning by analyzing SNS user behavior. In 2016 IEEE 5th Global Conference on Consumer Electronics, 1-3. https://doi.org/10.1109/GCCE.2016.7800405
[16] Lim, Y. M., Ayesh, A., Stacey, M. (2016). The motivation/attitude-driven behavior (MADB) model in E-learning and the effects on mouse dynamics. International Journal of Cognitive Informatics and Natural Intelligence (IJCINI), 10(3): 38-52. https://doi.org/10.4018/IJCINI.2016070104
[17] Chu, T.H., Chen, Y.Y. (2016). With good we become good: Understanding e-learning adoption by theory of planned behavior and group influences. Computers & Education, 92: 37-52. https://doi.org/10.1016/j.compedu.2015.09.013
[18] Panuš, J., Jonášová, H. (2015). Students' behavior on Internets social networks as a base for e-learning tools. In 2015 International Conference on Information and Digital Technologies, 286-290. https://doi.org/10.1109/DT.2015.7222985
[19] Shaqrah, A.A. (2015). Explain the behavior intention to use e-learning technologies: A unified theory of acceptance and use of technology perspective. International Journal of Web-Based Learning and Teaching Technologies (IJWLTT), 10(4): 19-32. https://doi.org/10.4018/IJWLTT.2015100102
[20] Chang, C.T., Hajiyev, J., Su, C.R. (2017). Examining the students' behavioral intention to use e-learning in Azerbaijan? The general extended technology acceptance model for e-learning approach. Computers & Education, 111: 128-143. https://doi.org/10.1016/j.compedu.2017.04.010
[21] Xu, W., He, Y. (2013). An Empirical study on influencing factors of E-Learning behaviour in an enterprise knowledge community. In Future Computer and Information Technology, Proceedings of the 2013 International Conference on Future Computer and Information Technology (ICFCIT 2013), 86: 83-91. https://doi.org/10.2495/ICFCIT130101
[22] Xie, T., Zheng, Q., Zhang, W. (2016). A behavioral sequence analyzing framework for

grouping students in an e-learning system. Knowledge-Based Systems, 111: 36-50. https://doi.org/10.1016/j.knosys.2016.08.001
[23] Haghparast, M., Hanum, N.F., Abdullah, N. (2013). Modeling an e-learning tool to cultivate critical thinking in students based on information needs and seeking behavior. In Proceedings of 2013 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE), 521-526. https://doi.org/10.1109/TALE.2013.6654492
[24] Tjhin, V.U. (2017). Exploring effective learning resources affecting student behavior on distance education. In 2017 10th International Conference on Human System Interactions (HSI), 104-107. https://doi.org/10.1109/HSI.2017.8005007
[25] Dow, C.R., Li, Y.H., Huang, L.H., Hsuan, P. (2014). Development of activity generation and behavior observation systems for distance learning. Computer Applications in Engineering Education, 22(1): 52-62. https://doi.org/10.1002/cae.20528
[26] Chang, H.C. (2010). Developing EL-RFM model for quantification learner's learning behavior in distance learning. In 2010 2nd International Conference on Education Technology and Computer, 4, 4-452. https://doi.org/10.1109/ICETC.2010.5529644
[27] Chen, Q. (2017). Learning behavior analysis and prediction based on e-learning platform data. In 2017 IEEE 17th International Conference on Communication Technology (ICCT), 2031-2036. https://doi.org/10.1109/ICCT.2017.8359985
[28] Umbleja, K., Ichino, M. (2016). Predicting students' behavior during an E-learning course using data mining. In International Conference on Interactive Collaborative Learning, 545: 175-189. https://doi.org/10.1007/978-3-319-50340-0_14

10 Authors

Shuliang Huan (1977), PhD candidate of School of Humanities, Southwest Jiaotong University, associate professor and master supervisor of College of Arts, Chongqing Three Gorges University. His research interests include Chinese classical literature and culture, higher education and reading teaching. He has presided over 3 national Social Science Fund projects, 20 provincial and ministerial scientific research projects (including 8 education and teaching reform projects), and 18 university-level projects (including 7 higher education and teaching reform projects). He has published more than 30 academic papers, 10 books, and granted 3 patents, respectively (email: 20020011@sanxiau.edu.cn).

Cuibi Yang (1978), Master. She has been engaged in the education and teaching of cultural industry management for a long time. She has presided over and researched a number of scientific research work related to cultural industry management and undertaken 7 national and provincial projects, including 2 key projects. She has presided over and researched 6 university-level projects, including 2 key projects, and 4 horizontal projects. She has published more than 10 academic papers, including 3 papers in EI journals, 1 in SSCI journal, and 2 books (email: lisaycb1978@163.com).

Article submitted 2022-02-03. Resubmitted 2022-03-13. Final acceptance 2022-03-20. Final version published as submitted by the authors.