Data Mining Method of English Online Learning Behavior Based on Machine Learning Technology

Yi Jiang[1](✉) and Yi-huo Jiang[2]

[1] Department of General Knowledge, Fujian Chuanzheng Communications College, Fuzhou 350202, China
[2] Fuzhou University of International Studies and Trade, Fuzhou 350202, China

Abstract. In the current application process of learning behavior data mining, the poor classification of data model training problems leads to long data mining time and affects accuracy. Therefore, a data mining method for English online learning behavior based on machine learning technology is proposed. First, set up the set of association items and establish behavior association rules. Cluster student behaviors based on association rules. And according to the clustering set, construct a learning object model, and use machine learning technology to train the model. After training, decision tree is used to mine data. In order to verify whether the design method meets the original intention of the design, the experimental analysis is carried out. The literature method and the designed data mining method are used to mine the students' behavior data in the English online learning platform of a university. The experimental results show that the designed data mining method has shorter time-consuming and higher accuracy, and achieves the original design intention.

Keywords: Data mining · Association rules · Machine learning · Clustering set

1 Introduction

Data mining has achieved effective application in many industries and fields, and its application in education is more and more extensive, but its application in English online learning is less in China, and many researches on English online learning do not mention data mining [1–3]. With the increasing popularity of English online learning, a large number of online learning data have been accumulated. It is very necessary to apply data mining technology to online learning. Foreign online learning learners research is mainly reflected in learning support services, through providing various services to learners, can meet the needs of learners, and improve online learning quality [4–6]. These support services are provided after a detailed analysis of learners. In China, the development of online learning started late, and the quality of online learning platform is uneven. From the current situation of online learning in China, there are mainly two forms of academic education and non academic education. The theoretical research

W. Fu et al. (Eds.): eLEOT 2021, LNICST 389, pp. 122–133, 2021.
https://doi.org/10.1007/978-3-030-84383-0_11

of learning support service is paid more attention than the practical research. Many domestic scholars and researchers attach great importance to the study of learners' learning support service [7, 8]. However, due to the traditional English online learning behavior data mining method, there is a problem of mining accuracy and long mining time. This paper proposes the English online learning behavior data mining method based on machine learning technology. By setting the association entry, establish a behavioral relationship rule, Based on the association item of the students in English learning, based on clustering results, build a learning object model, using machine learning technology to train this model, combined with data mining methods to explore online online learning behavior data, implement English online learning Digging of behavioral data. Through simulation experiments, it verifies that this method is shorter than the time of English online learning behavior data mining, and the excavation accuracy is high, and it has laid the foundation for improving online learning quality.

2 Design of Data Mining Method for English Online Learning Behavior Based on Machine Learning Technology

The data mining method designed in this paper uses machine learning technology for data model training Fig. 1. The specific process is as follows:

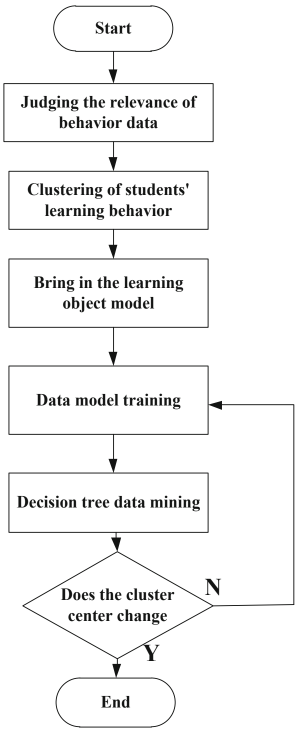

Fig. 1. Flow chart of online learning behavior data mining method

2.1 Behavioral Association Rules

Association rules can get the implicit association information between data, it can dig out the deep-level relationship between data and get the connection and law between them through analysis [9]. In the calculation of the behavior association rules, set the number of items included in the associated item set $I = \{I\ [1],\ I[2], ...I[k]\}$ as k, and call I the $k-$ item set. The association rule of English learning behavior is generally expressed as $X\ => $ Y, which reflects the concomitant relationship of item set X and item set Y, where $X \subseteq T$, $Y \subseteq T$, and $X \cap Y = \phi$, and X is used as the generation condition of the association rule, and Y represents the association rule As a result, it also reflects the rule that Y appears when itemset X appears. Usually, the support degree is used to measure the importance of association rules. The support degree of association rule X = >Y can be expressed as itemset X and itemset Y The probability of simultaneous occurrence in all itemsets, namely:

$$\sup(X\ => \ Y) = \sup(X \cup Y) = P(X \cup Y) \tag{1}$$

The formula in (1) can also be expressed as (number of transactions including X and Y / total number of transactions) x 100%. P stands for the support rate. The greater the support, the greater the frequency and importance of association rules. In practical application, a minimum support degree is defined in min_ Sup. Minimum support is a threshold defined according to the user's own requirements to measure the support. It is used to filter the generated association rules and remove some unimportant or useless rules. For a certain itemset I, if its support $\sup(I) \geq \min\_\sup$, then the itemset I is called frequent itemset. The confidence degree is usually used to describe the conditional probability of the occurrence of association rule results, denoted as $conf\,(X\ => \ Y)$, and its expression is as follows:

$$conf\,(X\ => \ Y) = \sup(X \cup Y)/\sup(X) = P(Y|X\,) \tag{2}$$

Confidence degree represents the probability of occurrence of item set Y when item set X occurs, and is generally used to measure the accuracy of association rules. In practical applications, general users will define a minimum confidence level by themselves, which is represented by min_conf, which is used to remove association rules whose accuracy is lower than the minimum confidence threshold. If the association rule $X\ => \ Y$ satisfies the conditions $\sup(X\ => \ Y) \geq \min\_\sup$ and $conf\,(X\ => \ Y) \geq \min\_conf$, then $X\ => \ Y$ is called a strong association rule. The strong association rule is generally useful information needed by users. After calculating the code rules, the cluster can be made according to the related items in English learning.

2.2 Learning Behavior Clustering

Before clustering the data set, the class is defined first. Due to the difference of sample things themselves, the representation of samples is also different, which leads to the difference in the definition of class [10]. Here are some definitions of classes. Firstly, the behavior data set is set as D, which contains sample n, S_i represents a certain sample, C and V represent the preset values of the sample. If any data S_i and $S_j \in D$ have

$d\left(S_i, S_j\right) \leq C$, then D is a class. If $S_i \in D$ in each data, then we can get the following results:

$$\frac{1}{n-1}\sum_{j=1} d\left(S_i, S_j\right) \leq C \tag{3}$$

In the formula (3), C represents the maximum number of samples in the sample set. And for any $S_i, S_j \in D$ can be obtained:

$$\frac{1}{n \times (n-1)}\sum_{i}\sum_{j} d\left(S_i, S_j\right) \leq V \tag{4}$$

If a, then D is $d\left(S_i, S_j\right) \leq V$ class. In order to make the clustering accurate, on the basis of completing the behavior association, we classify the data by similarity. In this paper, we choose the distance measurement method. If the data in the data set has t attributes, we can take each data point as a point in the t latitude space to judge the similarity of the data in the data set. Generally, we judge the similarity through the distance. If the distance between samples is small, the higher the similarity is, the smaller the difference is. On the contrary, the difference between data is larger. According to Mahalanobis formula, it can be concluded that:

$$D(X, Y) = (X - Y)^T \times \sum{}^{-1}(X - Y) \tag{5}$$

In (5), T represents the reference value of the Mahalanobis distance formula. In the teaching of English learning behaviors, the five aspects of English listening, speaking, reading, writing and overall ability can correspond to the score intervals of different levels of courses. This paper divides the clustering results into:

Table 1. Classification of cluster learning scores

Category	Preliminary	Beginner level	Qualified level	Master level	Proficiency	Proficiency
Listen	[0,46]	(46,67]	(67,75]	(75,79]	(79,83]	(83,100]
Say	[0,42]	(42,67]	(67,77]	(77,79]	(79,84]	(84,100]
Read	[0,42]	(42,62]	(62,69]	(69,82]	(82,85]	(85,100]
Write	[0,55]	(55,76]	(76,77]	(77,85]	(85,86]	(86,100]
Overall	[0,45]	(45,68]	(68,72]	(72,76]	(76,83]	(83,100]

According to the above Table 1, student samples can be specifically clustered and classified by their grades.

2.3 Construction of Learning Object Model

Personal E-learning behavior is a process of cognition and mastery of personal e-learning resources. However, due to the lack of supervision during personal online learning, the requirements for the self-consciousness and control ability of individual online learners are higher, and the quantity and variety of personal online learning resources are also a test of learners' information choice. In order to obtain higher learning efficiency, it is necessary to classify or formulate learning resources systematically. According to the degree of encapsulation, learning objects can be divided into compound type and atomic type. The reusability of learning objects is affected by the partition degree. The greater the partition degree is, the smaller the reusability will be, and the smaller the arc partition degree will be, the greater the reusability will be. It is also considered that the two types of learning objects are related and can be converted to each other under certain conditions. The atomic learning object mainly refers to the minimum segmentation unit that is achieved when the relevant knowledge is divided into the frame structure of the learning object. The division of atomic objects is based on the information block areas established according to the learning goals of different learners. Each information area contains a series of knowledge points. The atomic learning object is the most basic form of expression of the content of the smallest teaching unit. Among all learning objects, it has the best reusability. Atomic learning objects usually consist of a variety of learning materials, which have multiple formats, such as text, pictures, or other multimedia materials. When the learner is learning the course, the text description is usually used to logically explain the content of the course, and learners prefer non-text materials, such as images, videos, sounds, etc. available for personal online learning. Come for a better experience and increase the fun of learning (Fig. 2). The atomic learning object is shown below:

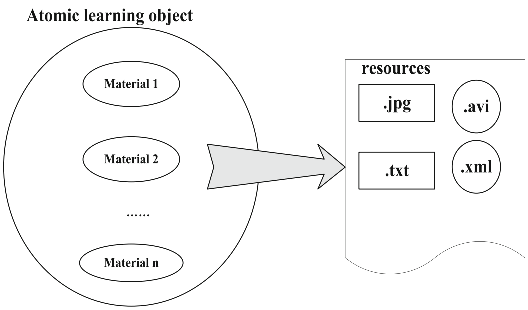

Fig. 2. Atomic learning objects

Compound learning object refers to a way of presenting the content, framework or internal relationship of specific learning resources. It can generally meet a variety of learning needs, and is usually composed of a variety of complex types of learning objects. Its main composition is shown in the Fig. 3.

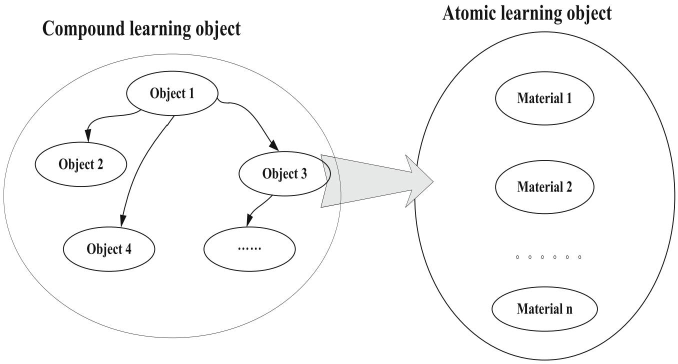

Fig. 3. Compound learning objects

According to the arrangement of the course content and the learning order of individual network learners, the relationship between the objects contained in the compound object can be divided into the following: 1. Antecedent relationship two objects. If the learning priority of one object is prior to the other, the former object is called the latter object.

The antecedent object of. 2. Similarly, for two objects, if the learning priority of one object is after the other, the latter object is said to be the successor of the previous object.

2.4 Machine Learning Technology to Achieve Model Training

After the learning object model is established, multiple separation problems often occur during data set training, and machine learning technology is required for model training. For multi-classification problems, that is, when the sample points in the training set to be classified belong to more than two categories, the two-class machine learning algorithm needs to be extended. The one-category reclassification method transforms the multi-classification problem into multiple two-classification problems, where each two-classification problem takes a certain class in the multi-classification problem as the positive class in the two-class classification, and all the rest in the multi-classification problem Class as the negative class in the two-classification problem, and then use the two-class support vector machine to construct multiple division functions. If a certain division function takes a positive value during classification, it means that the corresponding two-class support vector machine judges the sample The point belongs to the positive class of the classifier, that is, a certain class in the multi-classification problem. If the division function uses a negative value, it means that the corresponding two-level support vector machine determines whether the sample point belongs to the classifier, that is, it does not belong to the class in the multi-class problem. The way to solve the multi-classification problem by using the one-to-remainder method is shown in the figure below:

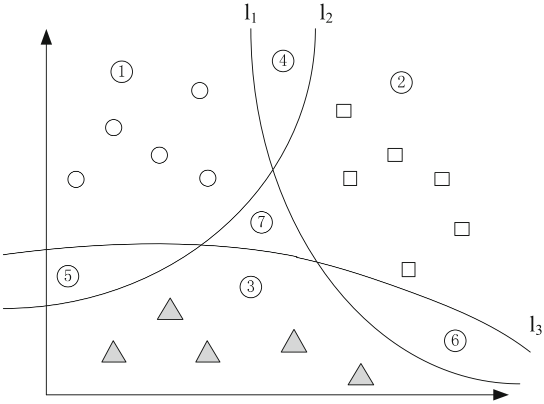

Fig. 4. Example of multi classification problem processing

In Fig. 4, the existence of rectangles, circles, and triangles represent three types of sample points, while l1, l2, and l3 respectively represent the three reticle curves obtained by solving the two-class support vector machine. When classifying, just substitute the feature vector of the sample point into the corresponding function of the three curves, and then make a decision based on the function value.

2.5 Decision Tree Data Mining

After completing the data clustering and the establishment of the learning object model, use it as a database. The database is set to $D = \{d_1, d_2, ..., d_n\}$, where $d_i =< d_{il}, ..., d_{ih} >$, the database contains the following attributes $\{A_1, A_2, ..., A_h\}$, and the point category set $L = \{L_1, ..., L_m\}$. For database D, the decision tree contains the following properties. Each internal node corresponds to an attribute Ai, each arc is used to judge the parent node, and the direction of decision-making is judged according to the attribute value corresponding to the parent node. Each leaf node is a completely divided Class, all the leaf nodes constitute all the classification of the sample data. The tree starts with a single node representing the training sample. If the sample is in the same type, the node becomes a leaf node and is labeled with this class. Otherwise, the algorithm uses the entropy measure called information gain as the heuristic information, and selects the attribute that can best classify the samples, which becomes the test or decision attribute of the node. Create a branch for each known value of the test property and divide the sample accordingly. The algorithm uses a similar method to recursively form the sample decision tree on each partition. Once an attribute appears on a node, it is not necessary to consider the attribute on the descendants of that node. On each node of the tree, information gain measure is used to select test attributes, which is also called attribute selection measure. By selecting the attribute with the highest information gain as the test attribute of the current node, the amount of information needed to classify the sample in the result division is minimized, thereby ensuring the simplest decision tree generated. Suppose B is a collection of b data samples, and there are m category attributes with different values, and m different classes $C_i (i = 1, ..., m)$ are defined at the same time. If s_i is the number of samples in class C_i, then it is necessary to classify a given sample The expected value is:

$$I(s_1, s_2, ..., s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{6}$$

In (6), p_i is the probability that any sample belongs to C_i. the logarithm function of $p_i = s_i/S$, $S = s_1 + s_2 + ... + s_m$ is based on 2, because the computer information is encoded by binary code. Let A have v different discrete attribute values $\{a_1, a_2, ...a_v\}$. The set S can be divided into v subsets $\{S_1, S_2, ...S_v\}$ by using attribute A, where S_j contains the data samples with a_j value taken by attribute A in S set. If attribute A is selected as test attribute, the current sample set is divided by attribute A. Let S_{ij} be the number of samples belonging to class C_i in subset S_j. The subset Di divided by A is obtained by the following formula:

$$E(A) = -\sum_{i=1}^{v} \frac{s_{1j} + s_{2j} + ... + s_{mj}}{s} I\left(s_{ij}, s_{2j}, ..., s_{mj}\right) \tag{7}$$

When the categories in the divided subsets are more unified, the direct value is smaller. Finally, from the formulas (6) and (7), the attribute can be used as the information gain value of test attribute A:

$$Gain(A) = I(s_1, s_2, ..., s_m) - E(A) \tag{8}$$

In the selection of attributes, the maximum information gain value is selected as the test attribute. At the same time, data mining is guaranteed by decision tree.

3 Experimental Demonstration Analysis

In order to verify the feasibility of the proposed method. In this paper, the data mining method in literature [2] and literature [10] is used for data mining of English online learning platform of a university.

3.1 Learning Platform Environment

The English online learning platform of a different grade of different grades in the experiment is the environment topology, Fig. 5 and the structure is as follows:

The operating system of the platform is Windows7, the IDE environment is MyEclipse10, the web server is JDK1.7 + Tomcat7, the programming language is JAVA + JSP, and the database is Oracle11g (Table 2). The hardware environment of the platform is as follows:

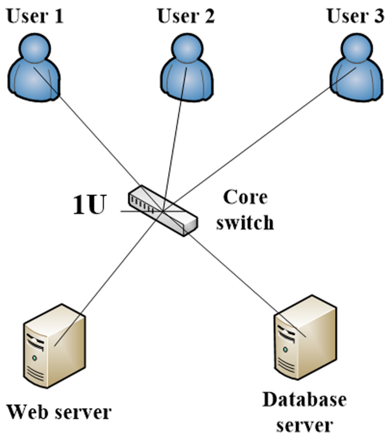

Fig. 5. Topology diagram of learning platform

Table 2. Hardware environment of experimental platform

Equipment	Model	CPU	Memory	Network bandwidth
Controller	HP ProDesk 498 G2 MT	Intel(R)Core(TM) I7–4790 CPU@3.60GHz	16 GB	100 Mbps
Calculate node	HP ProDesk 498 G2 MT	Intel(R)Core(TM) I7–4790 CPU@3.60GHz	16 GB	100 Mbps
Client	HP ProDesk 498 G2 MT	HP ProDesk 498 G2 MT	16 GB	100 Mbps

3.2 Experimental Data Situation

The data used in the experiment are the learning behavior data of the students in the above platform, and the experimental parameters are set as follows: the difference threshold is 0.01. The minimum number of samples is 2. In the experiment, the samples are divided into 10 samples, 9 of which are used as training set and the other as test set. The average accuracy was obtained by three experiments (Table 3) The sample dataset information is as follows:

Table 3. Experimental data set information

Data set	Data size	Number of attributes	Number of categories
Zoo	101	17	7
Banding	138	29	2
Monk1	124	6	2
Monk2	169	6	2
Vote	300	16	2
Crx	490	15	2
Soybean	683	35	19
Anneal	898	38	6
Hypo	2514	29	5
Letter	20000	16	26

3.3 Experimental Results

The data mining history of the experimental data set is shown in the following table:

Table 4. Data set mining operation table

Data set	Time consumption (MS)			Accuracy rate (%)		
	Method 1	Method 2	Method 3	Method 1	Method 2	Method 3
Zoo	87	89	99	100	99.4	100
Banding	138	144	165	99.6	89.4	97.3
Monk1	98	102	133	99.8	89.8	96.8
Monk2	96	118	145	97.5	83.4	88.5
Vote	111	127	156	95.4	80.2	87.5
CrX	145	162	178	94.3	79.4	84.4
Soybean	277	335	387	93.4	77.4	81.5
Anneal	343	368	388	97.5	92.1	94.5
Hypo	397	408	465	94.3	90.2	91.5
Letter	14531	15442	21813	92.1	88.4	90.1

In the above table, method one is the method used in this article, method two is the method in literature [2], and method three is the method in literature [10]. It can be seen from the above table that the data mining method designed in this paper takes less time than the other two methods when mining the data set, and has a higher accuracy rate.

In order to further prove the effectiveness of the method, an incremental operation is performed on the data set, and the experimental results obtained are as follows:

Table 5. Data set incremental mining operation table

Data set	Incremental scale	Time consumption (MS)			Accuracy rate (%)		
		Method 1	Method 2	Method 3	Method 1	Method 2	Method 3
Zoo	25%	54	69	63	99.8	97.4	93.2
Banding	35%	77	79	85	94.5	90.2	89.5
Monk1	30%	66	78	91	98.4	93.5	92.1
Monk2	40%	70	82	89	94.6	91.2	90.2
Vote	35%	61	73	82	99.6	92.1	94.5
Crx	50%	94	115	143	94.2	93.5	91.1
Soybean	30%	94	221	154	92.4	91.6	88.5
Anneal	20%	214	336	298	99.2	85.4	92.1
Hypo	30%	184	218	305	96.4	82.5	79.3
Letter	20%	5964	13541	13481	96.4	90.1	88.2

The experiments in Table 4 and Table 5 prove that the data mining method of English online learning behavior designed in this paper has higher accuracy and lower time-consuming. The design method is feasible.

4 Conclusion

This paper uses machine learning technology to train data mining methods, reducing the time-consuming and improving accuracy of data mining methods. However, there are still shortcomings in the research. The amount of data for data mining is relatively small. In the follow-up research, the selection of data sources should be expanded to further prove the research conclusions.

References

1. Zhu, J., Gong, C., Zhang, S., Zhao, M., et al.: Foundation study on wireless big data: concept, mining, learning and practices. China Commun. **15**(12), 1–15 (2018)
2. Ma, X.L., Zhai, X., Liu, J.W., et al.: Study on the biological basis of hypertension and syndrome with liver-fire hyperactivity based on data mining technology World. J. Tradit. Chin. Med. **4**(04), 176–180 (2018)
3. Cai, G., Lee, K., Lee, I.: Mining semantic trajectory patterns from geo-tagged data. J. Comput. Sci. Technol. **33**(4), 849–862 (2018)

4. Terrazas-Arellanes, F.E., Gallard, A.J.M., Strycker, L.A., et al.: Impact of interactive online units on learning science among students with learning disabilities and English learners. Int. J. Sci. Educ. **40**(5), 498–518 (2018)
5. Müller, A., Son, J.-B., Nozawa, K., et al.: Learning english idioms with a web-based educational game. J. Educ. Comput. Res. **56**(6), 848–865 (2018)
6. Liu, S., Lu, M., Li, H., et al.: Prediction of gene expression patterns with generalized linear regression model. Front. Genet. **10**, 120 (2019)
7. Liu, S., Li, Z., Zhang, Y., et al.: Introduction of key problems in long-distance learning and training. Mobile Netw. Appl. **24**(1), 1–4 (2019)
8. Liu, S., Glowatz, M., Zappatore, M., Gao, H., Jia, B., Bucciero, A. (eds.): eLEOT 2018. LNICSSITE, vol. 243. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93719-9
9. Wang, J.: A new system for distributed machine learning. Nat. Sci. Rev. **5**(03), 303–304 (2018)
10. Su, H., Li, Y., Zhang, X.: Data mining simulation of logistics frequent path based on radio frequency identification. Comput. Simul. **36**(09), 357–360–384 (2019)