Pattern Recognition Method of English Distance Online Education Based on Big Data Algorithm

Xiao-xiao Duan[1(✉)] and Ping Duan[2]

[1] International School, Chongqing Vocational Institute of Engineering, Jiangjin 402260, China
[2] Big Data and Internet of Things School, Chongqing Vocational Institute of Engineering, Jiangjin 402260, China

Abstract. In order to extract the features of English distance network education pattern, traditional pattern recognition methods of English distance network education result in low accuracy and large standard deviation of education pattern recognition. This paper proposes a pattern recognition method of English distance network education based on big data algorithm. Using big data technology, the digital English distance teaching resource database is established to avoid the duplication of acquired resources, the random forest algorithm of pattern big data is introduced, the single classifier of decision tree is used to distinguish characteristic data, the number of decision trees in the forest is adjusted, and the self construction process of random forest algorithm is optimized. In the process of pattern recognition, feature level state fusion and support vector machine are used to complete pattern recognition of English distance online education. The experimental results show that compared with the traditional algorithm, the standard deviation of the proposed algorithm is smaller, which can effectively improve the recognition accuracy.

Keywords: Big data algorithm · English distance online education · Pattern recognition · Random forest algorithm

1 Introduction

With the development of network technology, distance education system has gradually become a research hotspot. The purpose of distance education system is to achieve personalized, individualized and efficient teaching mode, which is a kind of broadening and extension of traditional teaching mode. The distance education system breaks through the limitation of time, region and resource sharing for knowledge seekers [1, 2]. However, most researchers focus on how to design personalized learning resources, provide navigation mechanism, diagnose learners' knowledge and skills, and develop effective learning programs. All education models are based on the interaction between learners and distance education system and the data reflected in the system for reasoning and judgment. If learners have correct learning attitude and pay attention in the learning process, then these data can fully reflect the education model of distance network, and

W. Fu et al. (Eds.): eLEOT 2021, LNICST 389, pp. 278–288, 2021.
https://doi.org/10.1007/978-3-030-84383-0_24

the education model recognition results obtained by the learning system based on these data the results are also true and reliable. If the attention is not focused, the result will have a certain deviation. However, the traditional pattern recognition method of English distance online education ignores the judgment of attention, which leads to low recognition accuracy. Therefore, this paper proposes a pattern recognition method of English distance online education based on big data algorithm. Using big data technology, the paper constructs the digital English distance teaching resource base, introduces the random forest algorithm of big data model, uses the single classifier of decision tree to distinguish the characteristic data, adjusts the number of decision trees in the forest, and optimizes the self-construction process of the random forest algorithm. In the process of pattern recognition, the feature level state fusion and support vector machine are used to realize the pattern recognition of English distance online education.

2 Pattern Recognition Method of English Distance Online Education Based on Big Data Algorithm

2.1 Establish Digital English Distance Teaching Resource Database

Big data technology can accurately capture business English reading resources on the Internet, and provide more diverse and comprehensive teaching resources for digital business English reading teaching. Therefore, this paper will use big data technology to establish digital English reading teaching resource database. Using web crawler technology, according to the key words of business English teaching, using breadth first traversal strategy to crawl resources from the Internet. After crawling, there are many kinds of resources and a large number of resources, and most of them will spread in multiple channels, resulting in the duplication of resources. Therefore, the collected teaching resources need to be processed [3, 4].

Extract metadata from resources for processing. Input the metadata of resources in the storage database, and process the resources according to the metadata. Audit whether the collection resources meet the collection requirements, whether the collection resources are duplicated, and whether the collection resources are missing. Using web resource filtering technology, the resources that do not match the keywords in the collected data are filtered, and the resources that do not match will be deleted directly. For the initial judgment of duplicate teaching resources, identify the resource information in the metadata, and calculate the similarity according to the following formula.

$$p(x,y) = \frac{\sum x_i y_i - n\overline{xy}}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}} \tag{1}$$

In formula (1), n is expressed as a hash function, $\overline{xy}$ represents the average value of resource data, x_i, y_i is the resource data that is preliminarily judged to be repeated, and the resources with high similarity are divided into the same cluster. Use the Bloom filter to sort the metadata of all crawled data, create a m bit array according to the amount of data, and set all the position values in the array to 0. k mutually independent hash functions are respectively mapped to the data in the metadata collection [5, 6]. For any

element in the set, the function value of the mapping position corresponding to the hash function will be set to 1. If the hash function values of all positions of the data array in the metadata collection are 1, the data is not duplicated. After filtering the repeated resource data by Bloom algorithm, the text data is normalized according to formula (2) to make the data in the same dimension.

$$x' = \frac{x - \mu}{\sigma} \tag{2}$$

In formula (2), x is the normalization function, μ is the mean value of all text data, and σ is the standard deviation of all data. After the establishment of a digital English reading teaching resource database, according to the requirements of business English teaching, integrated digital teaching methods, and research business English teaching models.

2.2 Introduction of Random Forest Algorithm of Big Data Model

Random forest is a combination classifier of multiple decision trees, which can play a better performance and recognition effect in the resource classification of English distance online education. The classifier provides some features of the decision tree, which is the basis of the decision system. The decision tree algorithm mainly produces rules by training classifiers [7, 8]. It is expected that the algorithm can realize a simple data cleaning process in the actual working process of the system, and classify the data sets with different location categories to further realize data mining. Determining the split attributes is the key step of constructing decision tree. Information gain should be taken as the measure in the process of attribute selection. The algorithm selects the attribute with the largest information gain after splitting to complete the main tree splitting.

Random characteristic variables are selected, and some attributes are randomly selected according to a certain probability distribution to participate in the node splitting process. The process of selecting split feature of random forest subtree is as Fig. 1.

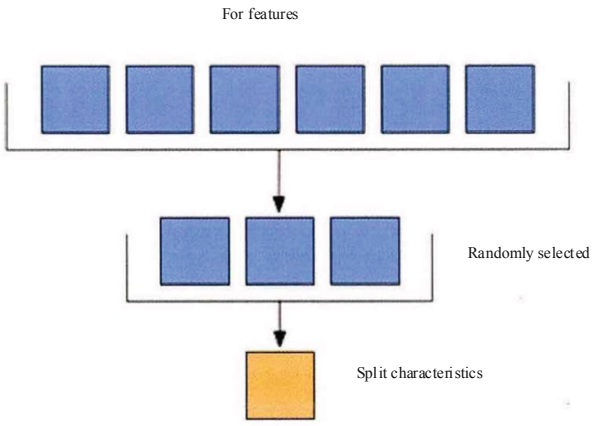

Fig. 1. Split feature selection process of random forest subtree

The change of information entropy mainly depends on the information gain to complete the appearance reflection [9], the definition formula of information entropy is:

$$E(X) = P(u_1)I(u_1) + P(u_1)I(u_2) + ... + P(u_r)I(u_r) \quad (3)$$

In formula (3), information entropy represents the mathematical expectation of the amount of information contained in the message u after it occurs. I represents the self-information amount of message u, and r represents the number of messages, P is the probability of the amount of information. When using the above formula to define information entropy to generate decision tree, the decision tree with smaller depth will be generated, so there will be no bias due to quantity. Under the support of the above algorithm, we need to build a random forest and complete the sampling in the data set. In this process, the random tree generation algorithm can be used to generate different branches of the forest decision system. The pattern recognition classification algorithm and random forest algorithm are combined to adjust the number of decision trees in the forest. By analyzing the construction process of random forest algorithm, the optimization process is completed, the classification accuracy and execution efficiency of different algorithms are verified, and the optimal algorithm is selected.

2.3 Extract the Characteristics of English Distance Network Education Mode

In the process of pattern recognition, some big data information needs to be fused. Data level fusion is usually used for multi-source pattern composition, pattern analysis and understanding. In the process of pattern acquisition, multi-source pattern composition is the same period of education mode acquired by different sensors, which has the characteristics of registration, resampling and synthesis. After that, the technology of education pattern recognition results is obtained, simultaneous interpreting the limitations and differences of single sensor education mode in resources, modes and spatial resolution, and improving the recognition accuracy. In the aspect of pattern analysis and understanding, it mainly studies how to deduce the three-dimensional model of the observed scene by using the output of high-resolution scanning collector [10–13]. The fusion technology of data layer includes classical detection and estimation methods. Target state fusion is mainly used in the field of multi-sensor target tracking. Many methods in the field of target tracking can be modified to multi-sensor target tracking. The content of target state information fusion in feature layer is as Fig. 2.

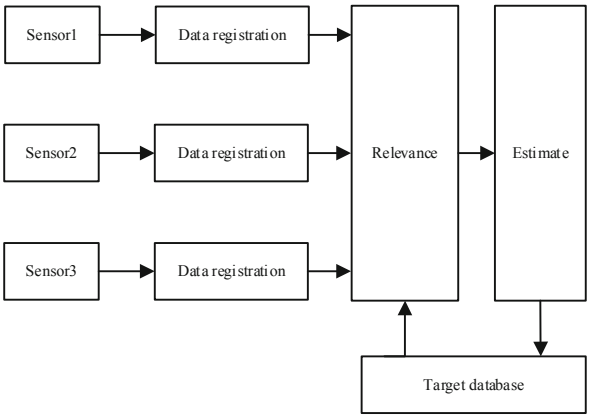

Fig. 2. Feature layer state fusion

Pattern recognition module is the core and the most important part of the system. The quality of this part directly affects whether the attention analysis module can be realized, and also directly affects the function of the system. This paper analyzes the characteristics and contents of English distance teaching mode, and under the guidance of Poa theory system, constructs digital English teaching mode by mixing flipped classroom teaching mode. The teaching mode of digital English course constructed in this paper is as Fig. 3.

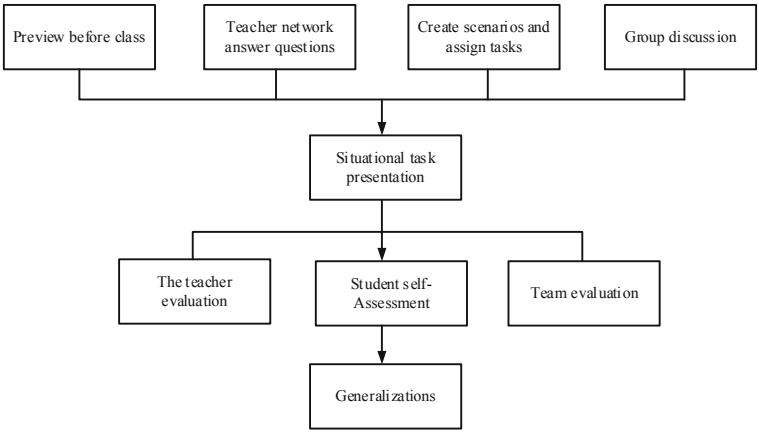

Fig. 3. English distance teaching mode

In terms of teaching content, we should flexibly select the real corpus, take the interest as the guidance, select new materials with certain gradients, pay attention to the explicit and implicit embodiment of western culture in English distance education, and update the content module of English.

In the pre class preparation stage, when teachers record English reading online courses, they should classify them reasonably according to the specific content of English

reading courses, and each category is a unit. In each unit, it can be divided into 3–4 sections according to the theme, and each section records 8–10 min of relevant videos, text, questions and answers, etc.

In the stage of internalization and absorption of classroom knowledge, we need to adopt task driven teaching method to teach the classroom content, improve students' English reading comprehension ability and the ability to solve the problems they may encounter in their career. After the end of classroom teaching, the teacher counts the students' reading course summary and other information, summarizes the knowledge points and abilities that students should improve, and puts forward corresponding suggestions, so that students can review in time. In all teaching practice, it is advocated to take the student as the center, based on the big data platform of classroom teaching, teachers and students are not only the sender and receiver of information, but also the disseminator of information. Teachers can obtain the first-hand teaching data through the teaching platform, which is not only conducive to students' preview before class, review after class, and be familiar with the key and difficult points in each teaching link, but also conducive to teachers' overall grasp of teaching materials and students, so as to achieve targeted, individualized teaching, and enhance the interest and intelligence of classroom teaching.

2.4 Pattern Recognition of English Distance Online Education

For pattern recognition, this paper describes the basic idea of support vector machine. Support vector machine is essentially a non negative quadratic optimization problem, which can obtain the global optimal analytical solution in theory. Support vector machine is a machine learning method based on statistical learning theory. It adopts the principle of structural risk minimization, and has the advantages of small sample, non-linear and "avoiding digit disaster". Support vector machine can be used to solve linear and non-linear problems, and it has good performance in pattern recognition. Its main principle is to find the optimal classification surface to get the results of pattern recognition.

For nonlinear problems, we can also transform them into high-dimensional space by introducing kernel function. The linear support vector machine (LSVM) used in this paper is based on the maximum interval method. The maximum interval method transforms the problem of finding the optimal classification surface into the problem of finding the maximum classification interval. By Lagrange multiplication method and introducing dual function, the optimization problem is transformed into a quadratic linear programming problem, and the features of training samples are extracted. The extracted features are used as the feature vectors of training samples for SVM model training, so as to obtain the training model, namely classification the device. According to the statistical learning theory, if the training sample set is not wrongly separated by the hyperplane, and the distance between the nearest sample data and the hyperplane is the largest, then the hyperplane is the optimal classification hyperplane as Fig. 4.

First determine the number of classifiers, using the SVM algorithm, the goal is to divide the sample into 2 categories, this experiment only needs to train one classifier to determine the feature vector of the training sample. After the feature extraction of the above education mode, the training samples are obtained, and the feature $C_{i,j}$ and resource feature $T_{i,j}$ of the education mode are obtained, as the feature vector $F\left(C_{i,j}, T_{i,j}\right)$

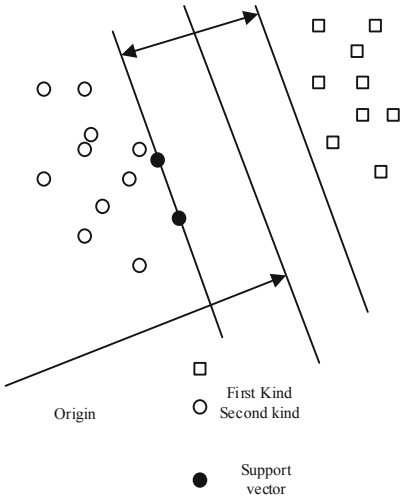

Fig. 4. The optimal classification hyperplane

of the training sample, j represents the category, and i represents the j feature point in the i category. The features of each feature point in R, G and B channels are extracted as follows:

$$C_{i,j} = \left(C_{i,j}^{R}, C_{i,j}^{G}, C_{i,j}^{B}\right) \tag{4}$$

The resource characteristics of pixels in energy $E(n)$, entropy $H(n)$, moment of inertia $I(n)$ and correlation $C(n)$ are:

$$T_{i,j} = \left(T_{i,j}^{E}, T_{i,j}^{H}, T_{i,j}^{I}, T_{i,j}^{C}\right) \tag{5}$$

In the training process, use the SVM toolkit that comes with Matlab, and use the following model model = svmtrain (TrainLabel, TrainData) to complete the training. Among them: TrainLabel is the category label, TrainData is the training sample data, and the extracted feature vector $F\left(C_{i,j}, T_{i,j}\right)$, import and perform model training, that is, build a classifier. The penalty parameter C is introduced to implement the penalty for misclassification. In practical applications, some important samples have high requirements for correct classification, and some samples have low requirements for correct classification. Therefore, in the description of the optimization problem, a different penalty coefficient is used for each sampling point data to obtain a more accurate classification. This kind of support vector machine is called a weighted support vector machine, which can be expressed as:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{i} S_i\xi_i \tag{6}$$

In formula (6), w represents the weight, ξ_i represents the number of samples, S_i represents the weighting coefficient, and the weighted support vector machine needs to

choose a reasonable S_i. The first sample in the sample set has the lowest importance, S_i is set to a value less than 1, and the last sample has the highest importance, S_i is set to 1, and linear interpolation is used to obtain the weighting coefficients of other sample points. S_i has only 2 values, which is determined by the number of samples in each category in the second category. The obtained support vector machine avoids the problem that the classification result of the conventional weighted support vector machine will be biased towards the larger number when the number of categories is not balanced. The above is the recognition of education mode under the background of big data technology studied in this article.

3 Simulation Experiment

3.1 Design Simulation Experiment

Put the pattern recognition algorithm designed in this paper and the traditional recognition algorithm in the same experimental environment for simulation, and compare the results to verify the reliability of this algorithm. In the experimental platform of this article, the CPU is the Inter(R) Q4800 model, the frequency is 2.66 GHz, the computer memory is 512 GB, the simulation programming environment is Matlab 2016 under Windows10, the classifier uses LibSVM, and the grid search is used. Method for parameter optimization, where the kernel function used is Gaussian kernel:

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right) \tag{7}$$

In formula (7), σ^2 represents the width of the function. According to the actual situation of this experiment, adjust the value of σ^2 to 0.64. Establishing distributed multi-sensor detection in the above simulation environment, the state of these sensors sent to the fusion center can be expressed as:

$$\hat{X}_j^i\left(i = 1, 2, ..., N; j = 1, 2, ..., n_j\right) \tag{8}$$

In the experimental simulation platform, set the initial parameters of the sensor: In sensor 1, its position coordinate is $(X, Y) = (120\text{ km}, 120\text{ km})$, the observation radius r_1 is 100 km, the ranging error σ_{l_1} is 150 m, the angle measurement error σ_{θ_1} is 1°, and in sensor 2, its position coordinate is $(X, Y) = (140\text{ km}, 140\text{ km})$, the observation radius r_2 is 80 km, the ranging error σ_{l_2} is 120 m, and the angle measurement error σ_{θ_1} is 1°. In the simulation experiment, five datasets of wine, forest, glass, iris and segmentation are used for recognition in UCI standard database. The basic information of datasets is as Table 1.

Table 1. Basic information of the data set

Data set	Sample input dimension	Number of training samples	Number of test samples	Number of identifications
Wine	24	133	133	4
Forest	47	247	317	7
Glass	18	159	114	2
Iris	11	121	84	3
Segmentation	37	331	3310	5

In the above five authoritative data sets, the algorithm designed in this paper and the traditional algorithm are respectively used to run 50 times, and the mean and standard deviation of the two algorithms are calculated.

3.2 Experimental Results and Comparison

In the above experimental process, the recognition accuracy of the two algorithms is counted as Table 2.

Table 2. Comparison of recognition accuracy of two algorithms

Recognition pattern classification based on different algorithms		Number of correct samples	Total samples	Recognition accuracy
Traditional algorithm	Wine	107	133	80.4%
	Forest	271	317	85.5%
	Glass	87	114	76.3%
	Iris	62	84	73.8%
	Segmentation	2417	3310	73.0%
The algorithm in this paper	Wine	128	133	96.2%
	Forest	287	317	90.5%
	Glass	107	114	93.9%
	Iris	77	84	91.7%
	Segmentation	3064	3310	92.6%

The statistical results of the standard deviation of the five data sets of the two algorithms in the experiment are as Table 3.

Table 3. Comparison of standard deviation results of two algorithms

Data set	Traditional algorithm	The algorithm in this paper
Wine	0.084	0.051
Forest	0.087	0.071
Glass	0.109	0.066
Iris	0.036	0.067
Segmentation	0.071	0.059

From the experimental results in Table 2 and Table 3 above, it can be seen that for the recognition results of the same data set, the average accuracy of the traditional algorithm is 77.8%, and the average accuracy of this algorithm is 93.0%. The average standard deviation of the algorithm in this paper is 0.063 in the experiment, while the average standard deviation of traditional algorithm in the experiment is 0.077. The recognition standard deviation of this algorithm in the experiment is also smaller than that of the traditional algorithm, and it is relatively stable. Therefore, it can be concluded that the recognition accuracy of this algorithm is higher than that of the traditional algorithm high. In conclusion, the algorithm can effectively reduce the standard deviation of English distance education pattern recognition and improve the accuracy of the pattern recognition.

4 Conclusion

Pattern recognition technology has been developed rapidly in recent years. At the same time, pattern recognition technology is also a subject with scientific research and application prospects. With the efforts of researchers, pattern recognition technology has been applied in many industries. This paper analyzes the common problems of distance education system, and proposes a pattern recognition method of English distance education based on big data algorithm. That is to say, pattern recognition technology is applied to process the learner's head image captured by the camera to identify the degree of eye deviation in the learner's head image. Fuzzy set and probability theory are used to analyze the learner's attention and give necessary tips. To a certain extent, the system makes up for the shortcomings of the existing distance teaching system. The image processing technology such as Gaussian smoothing and binary is used to preprocess, which reduces the standard deviation of pattern recognition. The recognition of educational pattern is finished by using SUSAN operator algorithm and fuzzy set theory, so as to improve the recognition accuracy.

References

1. Chen, X.C., Ding, P., Yan, N.Q., et al.: Study on discrete manufacturing quality control technology based on big data and pattern recognition. Math. Probl. Eng. **5**, 1–10 (2021)

2. Su, G.: Analysis of optimisation method for online education data mining based on big data assessment technology. Int. J. Contin. Eng. Educ. Life-Long Learn. **29**(4), 321–335 (2019)
3. Yang, H., Liu, J., Zhang, M.: Face recognition algorithm based on orthogonal gradient difference local directional pattern. Laser Optoelectron. Prog. **55**(4), 150–156 (2018)
4. Wang, S., Xue, J., Hu, H., et al.: Pattern recognition of partial discharge based on the feature parameter optimization selection and multi-algorithm combined methods. High Volt. Appar. **54**(10), 112–119 (2018)
5. Geng, C., Zhang, J., Guan, L.: A recommendation method of teaching resources based on similarity and ALS. J. Phys. Conf. Ser. **1865**(4), 042043 (8pp) (2021)
6. Xing, Z., Li, G.: Intelligent classification method of remote sensing image based on big data in spark environment. Int. J. Wirel. Inf. Netw. **26**(3), 183–192 (2019). https://doi.org/10.1007/s10776-019-00440-z
7. Yuan, S.: Research on pattern recognition of laser fluorescence spectrum data in big data background. Laser J. **39**(05), 124–127 (2018)
8. Zhao, J., Liu, Y.: Research on big data technology in computer network intrusion detection. J. Netw. New Media **7**(04), 45–49 (2018)
9. Song, M., Wang, D., Zhang, S., et al.: Flatness pattern recognition model based on recurrent neural network. Iron Steel **53**(11), 56–62 (2018)
10. Liang, Z., Lin, D., Huang, R.: New cloud security management model based on big data and cloud computing. Autom. Instrum. **2018**(07), 189–191+196
11. Fu, W., Liu, S., Srivastava, G.: Optimization of big data scheduling in social networks. Entropy **21**(9), 902 (2019)
12. Liu, S., Bai, W., Zeng, N., et al.: A fast fractal based compression for MRI images. IEEE Access **7**, 62412–62420 (2019)
13. Liu, S., Liu, D., Srivastava, G., et al.: Overview and methods of correlation filter algorithms in object tracking. Complex Intell. Syst. (3), 2580–2583 (2020). https://doi.org/10.1007/s40747-020-00161-4