IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 24, 2020, accepted May 31, 2020, date of publication June 11, 2020, date of current version June 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3001749

An Optimized Mining Algorithm for Analyzing Students' Learning Degree Based on Dynamic Data

ZENGZHEN SHAO[1,2], HONGXU SUN[1], XIAO WANG[1], AND ZHONGZHI SUN[1]
[1]School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China
[2]School of Data and Computer Science, Shandong Women's University, Jinan 250002, China

Corresponding author: Zengzhen Shao (shaozengzhen00@sina.com)

This work was supported in part by the China Postdoctoral Science Foundation under Grant 2016 M592697, in part by the Shandong Women's University High Level Cultivation Fund under Grant 2018RCYJ04, and in part by the Discipline Talent Team Cultivation Program of Shandong Women's University under Grant 1904.

ABSTRACT With the rapid development of educational informatization, it has enabled education to enter the era of big data. How to extract effective information from educational big data and realize adaptive personalized learning goals have become the current research hotspot. The traditional static data only analyzes the students' learning degree based on the students' final answer, but ignores the dynamic data in the process of answering questions, such as the modification and the time it answered on the question, which makes it difficult to fully and accurately mine the correlation between the massive data, so it turns from static data mining to dynamic data mining. The paper proposes an optimized mining algorithm for analyzing students' learning degree based on dynamic data. The algorithm first uses the optimized text classification technology to match the question texts to the knowledge points automatically, so as to improves the efficiency and quality. Then, it uses the subjective weighting method combined with the expert experience to generate the learning degree matrix of students on knowledge points based on dynamic data of the students' records. Finally, the DBSCAN clustering algorithm is used to cluster the personalized learning characteristics of students according to the learning degree matrix. The experimental result shows that the algorithm can deal with massive data automatically and effectively, and analyze the students' learning degree on knowledge points comprehensively and accurately, so as to classify students and realize personalized teaching.

INDEX TERMS Data mining, dynamic data, students' learning degree, subjective weighting method, clustering algorithm.

I. INTRODUCTION

Education data mining [1] is an important branch of contemporary data mining, specifically referring to data mining in the field of education. It is a process of using data mining techniques and methods to extract the valuable and meaningful information from a large number of educational data, so that relevant personnel can better serve education and teaching with the extracted information [2]. More and more researchers are applying data mining technology to all aspects of school education, mining students' learning level deeply and formulating corresponding measures to achieve the ultimate goal of improving students' education and achievement.

The associate editor coordinating the review of this manuscript and approving it for publication was Massimo Cafaro.

Learning degree refers to the students' ability to grasp and understand on specific knowledge points, that is, the mastery on knowledge points. The researcher maps out the knowledge points or concepts of the test questions, and judges the students' learning degree on the knowledge points through the students' answers to the test questions [3], [4].

In recent years, researchers have done a lot of research in mining students' learning degrees and the factors that affecting students' academic performance. Shao *et al.* [5] proposed a TA-ARM algorithm for automatic generation of concept maps based on text analysis and association rule mining. This algorithm only used students' right and wrong answers to get the students' learning degree of concepts and the degree of correlation between concepts. Duhayyim and Newbury [6] used fuzzy logic to generate color concept maps for

evaluating students' learning degree of each knowledge point; Chen *et al.* [7] analyzed the data produced in the teaching process, used the neural sequence marking algorithm to generate concepts suitable for students' learning situation and mined the association rules to get the practical significance of education. Macfadyen and Dawson [8] identified 15 variables by analyzing the LMS trajectory data in the course supported through blackboard vision. Variables have a significant and simple correlation with the students' final grades to investigate effectively which students' activities online predict students' learning degree and academic performance accurately. Kaur *et al.* [9] focused on identifying the students with slow learning speed and poor learning degree on the knowledge points, and displayed it by a predictive data mining model using classification-based algorithms.

In short, in recent years researchers have made great achievements in mining students' learning degree and other aspects, but there are great limitations in data processing, such as traditional education by data mining is a comprehensive analysis of students' learning level based on the overall answer records of all students. Due to the large amount of data analyzed in general, ignoring the internal attributes of its data and the relationship between dynamic data and student learning levels and the impact of students' personalized learning characteristics, which makes it difficult to understand student's learning degree comprehensively and accurately. In this paper, students' test data are taken as an example, and a data-driven analysis method for students' learning degree is proposed—anoptimized algorithm for mining dynamic data. Aiming at the above limitations, the algorithm mines the correlation between the knowledge points in the test questions and the dynamic data captured in the records of students answered, so that the scientific, comprehensive and in-depth analysis and research of the student's learning degree is supported by the data.

The main contributions of this study are stated as follows:

(i) The first is that the optimized classification model (i.e., SVM-KNN) not only saves the time to classify the test questions into knowledge points, but also improves the accuracy of classified.

(ii) The second is to propose an optimized mining algorithm for students' learning degree mining which combines the subjective weighting method and the answer records. The algorithm solves the limitation of analyzing the traditional static educational data in the past, and through mining and analyzing the dynamic data in the educational data captured, it makes a more comprehensive and accurate analysis of the students' learning degree on knowledge points. This is a more prominent innovation of the manuscript. The combination of the subjective weighting method and dynamic data records of students answered can be more comprehensively analyzed and researched the students' learning degree on knowledge points and can be obtained for adaptive learning. At the same time, it can assist educators in scientific teaching.

The remainder of this paper is organized as follows. Related literature is reviewed in Section 2. The explanation of the algorithm proposed are discussed in Sections 3. Section 4 conducts the computational experiments and analysis. Finally, we conclude our results and point out the future research directions in Section 5.

II. RELATED WORK

As an information processing method, data mining [10] was born with the existence and application of massive data in the contemporary, and has been the successfully applied in many fields, such as finance [11], medicine [12], biology [13] and other fields. In the field of education, data mining has also started slowly. Educational data mining (EDM) refers to the process of applying data mining methods to extract the meaningful information from data of the education system, which can better provide the more valuable services for educators and learners [1].

In recent years, researchers have done a lot of researches on education data mining. Shahirip *et al.* [14] used the predictive algorithm to identify the most important attributes that affect students' achievements data, and to improve students' achievements through the data mining techniques. First, they used cumulative average performance (CGPA) and internal assessment to predict students' learning performance, and then used several classical classification models to predict important students' attributes, performance and learning degree in students' achievements, and finally guided students to improve achievements. Chen *et al.* [7] analyzed the data generated during the teaching process, used neural sequence labeling algorithms to generate concepts suitable for students' learning degree, and mined association rules to derive the practical significance of education. The algorithm can generate concept maps based on the records of students answered automatically. Shao *et al.* [5] proposed a TA-ARM algorithm for automatic concept map generation based on text analysis and association rule mining. This algorithm used the right and wrong answers in the records of students answered to mine the concepts on students' learning degree and test the degree of correlation between concepts of test questions, which was the concept map. Åžen *et al.* [15] used a large and feature-rich data set to develop a predicted model, mined important prediction indicators related to students' learning degree and academic achievements, and performed sensitivity analysis on the predictive model. Finally, it was determined that the C5 decision tree algorithm was the optimal predictor, and the accuracy of the retained sample was 95%, followed by the support vector machine (91% accuracy) and artificial neural network (89% accuracy), which was determined by forecast indicators of the predictive model.

In addition, there are many studies on comparison and improvement using the text analysis methods. Wang *et al.* [16] studied the effectiveness of three neural networks in competition, back propagation (BP) and radial basis function (RBF) in text classification and their effect on text classification. The experimental results show that the supervised training model has higher accuracy and recall rate in the classification model than the unsupervised training

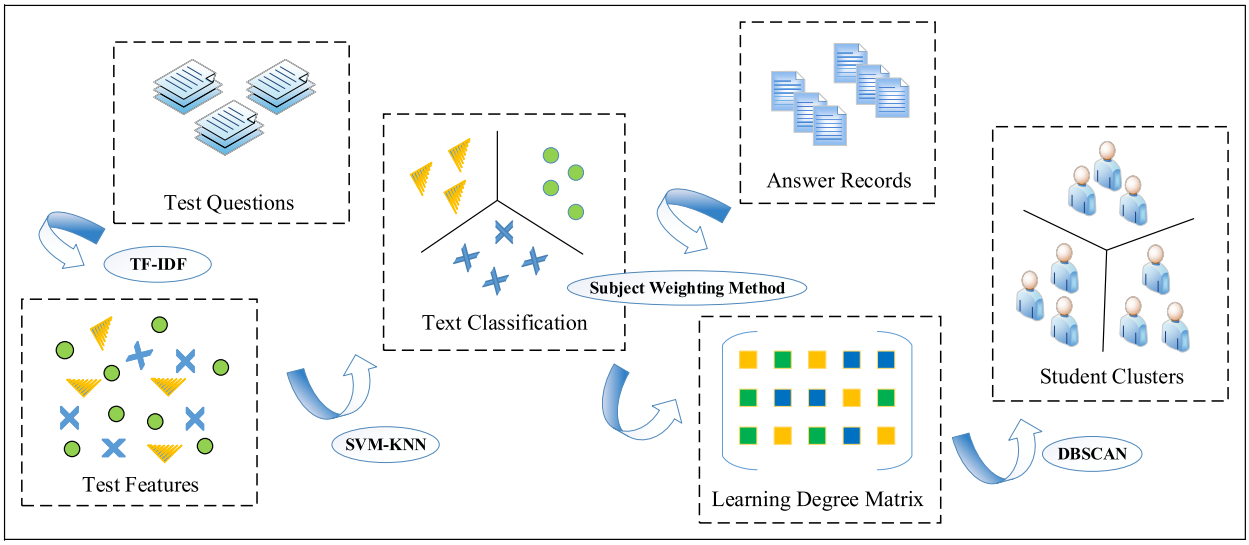

FIGURE 1. Overall flow chart of the proposed algorithm.

model. Therefore, supervised classification training models have more concerned. Cao and Chen [17] proposed a new web text classification algorithm based on the application of basic support vector machine algorithm. Combining the SVM algorithm and the KNN algorithm, a web text classification algorithm based on the SVM-KNN is proposed. The KNN algorithm is used to make up for the shortcomings of the traditional SVM algorithm. The traditional SVM algorithm is effectively improved with simple ideas and a small implementation cost. At last, it received a good classification effect. Liu and Wu [18] proposed a new classification algorithm formed by combining one class SVM and KNN: One Class SVM-KNN. As well known, KNN is one of the important classification methods in the field of data mining, and KNN training is low in cost, easy to deal with sample sets with overlapping or overlapping class domains. SVM has a good effect in solving a single class classification problem, so the new classification algorithm is combined by taking advantage of both. Through the above experimental analysis, we can know that the one class SVM-KNN method can well solve the class tilt, storage and calculation of the traditional KNN method. The disadvantages such as large overhead and obvious improvement in classification result are a feasible method. This paper is inspired by the above literatures, through establishing a training model for processing test data, the association between test questions and knowledge points is obtained, and the proposed optimized algorithm for mining dynamic data is combined to obtain the data-driven students' learning degree.

III. AN ALGORITHM FOR MINING STUDENTS' LEARNING DEGREE

A. BASIC IDEAS OF THE ALGORITHM

The algorithm consists of three parts: Firstly, a training model for processing test data is established. The TF-IDF method is used to pre-process the test questions and extract the text features. The SVM-KNN classification model is established for automatic training, which is replaced classifying problems into knowledge points manually. At this stage, the correlation between the test questions and the knowledge points can be obtained. Secondly, an optimized algorithm for mining students' learning degree using subjective weighting method combined with answer records is proposed. At this stage, the dynamic characteristic data of important influencing factors related to the student's learning degree are first mined from the students' answer records of the examination, such as the length of the answer path selected and the proportion of the number of correct options in the answer path selected. The subjective weighting method is used in combination with authoritative experts in related fields to give weight coefficients of various influencing factors to obtain the students' learning degree on the knowledge points corresponding to each test question. Finally, the clustering algorithm is used to classify the students' learning degree based on their learning level on the test questions. In this process, we use the DBSCAN clustering algorithm to classify the students' learning degree on the test questions, and obtain the students' learning degree curve trajectory with significant features on the knowledge points. According to the trajectory in-depth research and analysis, the teaching mode is changed from ''teacher-centered'' to ''student-centered'', which provides the direction and teaching guidance scheme for the personalized teaching. The algorithm is shown in Fig.1. In addition, we use the notations in Table 1 throughout the paper.

B. TEXT ANALYSIS OF THE TEST QUESTIONS

The test questions participating in the analysis are classified into knowledge points in text analysis [19] by the text classification technology [20], [21], and each test question belongs to one knowledge point. For the time being, this paper only

TABLE 1. Notations.

Symbol	Meaning	Illustration
Q	Test questions after word segmentation and stop words filtering	$Q = \{Q_1, Q_2, \ldots, Q_i, \ldots, Q_n\}$
Q_i	The i-th test question	
n	The number of test questions	
s_t	The t-th student	
C	Text features extracted by TF-IDF	$C = C_1, C_2, \ldots, C_i, \ldots, C_n$
C_i	The i-th text feature of test questions	$C_i = (C_{i1}, \cdots, C_{ij}, \cdots, C_{ir})$
r	Dimension of the text feature	
K	The knowledge points (that is, class labels)	$K = (K_1, K_2, \ldots, K_i, \ldots, K_p)$
p	The number of knowledge points.	
QK	Questions-Knowledge points matrix	The results by the SVM-KNN classification model
SQ	Student' learning degree on each question matrix	The student's answer records
sq_{tj}	Student' learning degree	$sq_{tj} \in \{0,1\}$
m	The total number of students	
β_u	The weight coefficient of each component in the proportion	$u \in \{1,2,3\}$
θ_1	The right or wrong answer	
θ_2	The students' answer path length selected	
θ_3	The proportion of the correct choice times in the length of the students' answer path length selected.	
F_e	The right answers to the questions q_j	$F_e \in [A, B, C, D]$
F_a	The final choice selected by student s_t	$F_a = 0$ or $F_a = 1$
L	Students' answer path length selected	
P	The students' answer path length selected to choose answers on the test questions as $P = < p_1, p_2, \ldots, p_h >$	$h = 1,2, \ldots, 6$
w	The number of correct answers in the students' answer path length selected	$w = 0,1,2, \ldots, 6$
SK	The learning degree matrix of students on each knowledge point	
sk_{tx}	The learning degree of the student s_t on knowledge point k_x	$sk_{tx} \in \{0,1\}$
sq_{ta}	The average value of student s_t on the questions q_a	

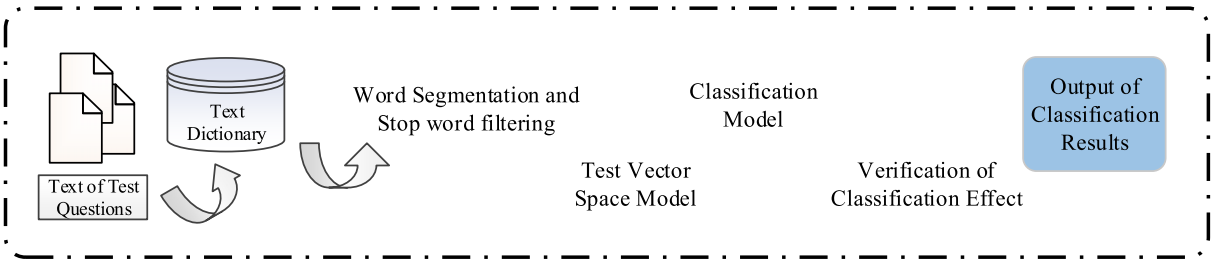

FIGURE 2. The process of the text classification.

considers the situation where one question corresponds to one knowledge point, and later will consider the situation where one question corresponds to multiple knowledge points. The data in the process of answering questions are analyzed and mined in depth to obtain the students' learning degree on knowledge points. First, the single choice test questions and questions are combined into a single record for textual analysis. If the test question belongs to a certain knowledge point, then establishes the corresponding link, otherwise will not establish the link. Specifically, we mark the knowledge point matrix of test questions as QK.

$$\mathrm{QK} = \begin{bmatrix} qk_{11} & qk_{12} & \cdots & qk_{1p} \\ qk_{21} & qk_{22} & \cdots & qk_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ qk_{n1} & qk_{n2} & \cdots & qk_{np} \end{bmatrix}$$

This paper takes the test Q_i as an example, where qk_{ij} indicates whether the test question Q_i belongs to the knowledge point K_j, $qk_{ij} \in \{0, 1\}$, n represents the number of the test questions, and p represents the number of the knowledge points. When $qk_{ij} = 1$ indicates that Q_i belongs to the knowledge point K_j; when $qk_{ij} = 0$, it means that Q_i does not belong to the knowledge point K_j.

In this paper, we use the general text classification technology to mine the test questions and classify them to the knowledge points automatically. The text classification process is shown in Fig.2.

1) WORD SEGMENTATION AND STOP WORDS FILTERING

First of all, we need to pretreat the test questions. The differences between Chinese and English test questions lies in that English is word-based, words and words are separated by spaces, while Chinese is word-based, and all the words in a sentence can be connected to describe a meaning, so Chinese character order is needed. The column is divided into meaningful words, that is, Chinese word segmentation [22].

The text after word segmentation contains many meaningless words. Therefore, we will filter these meaningless words next, that is, stop word filtering.

In order to facilitate the expression in the following steps after word segmentation and stop words filtering in the technical expertise, and we use $Q = \{Q_1, Q_2, \ldots, Q_i, \ldots, Q_n\}$ to express the test question, where Q_i is the i-th test question and n represents the number of the test questions.

2) TEXT FEATURES EXTRACTION

TF-IDF (Term Frequency–Inverse Document Frequency) [23] is a commonly weighted technique to use for information retrieval and data mining, which can evaluate the importance of a word to a file in a file set or a corpus. The more times a word appears in an article, the less times it appears in all articles, and the more it can represent the central meaning of the article. Models represented in text are usually regarded as semi-structured or unstructured. In order to adapt to the computer processing, it must be converted into a format that can be recognized by the machine, while preserves the original semantic information of the text as much as possible. At present, common classification models include Rocchio [24], Logistic Regression [25], Naive Bayes [26], KNN (K-Nearest Neighbors) [27] and SVM (Support Vector Machine) [28]. In this paper, TF-IDF method is used to extract text features of test questions, and Q is transformed into spatial vector model VSM [29].

The text features extracted by TF-IDF are expressed as $C = C_1, C_2, \ldots, C_i, \ldots, C_n$ corresponding with Q, where C_i is the i-th text feature. Similarly, C_i can also be expressed as $C_i = (C_{i1}, C_{i2}, \cdots, C_{ij}, \cdots, C_{ir})$, where C_{ij} is the weight of the feature item, j is the i-th test question, and r is the dimension of the text feature. The formula for calculating the weight of a feature item C_{ij} is as follows:

$$C_{ij} = TF_{i,j} \times IDF_j \tag{1}$$

The $TF_{i,j}$ represents the word frequency of text feature item j in the text of the i-th test question, and IDF_j represents the number of feature items j appearing in the whole text data set, which is the reverse document frequency. Because TF-IDF is a weighting function, which its value depends on the word frequency and reverse document frequency in a given document, the word frequency is proportional to the weight of feature items and reverse document frequency is inversely proportional to the weight of feature items.

3) MODEL CLASSIFICATION AND EVALUATION

In this step, it needs to perform classification model processing on the text classification feature C extracted in the previous step, and uses the classification model to classify and label the test questions automatically. In this paper, the KNN classification model is selected for the first time. KNN is a classification model with good classification effect and easy to implement relatively. Its time complexity is proportional to the number of test questions, which meets the needs of the algorithm in this paper. However, there are some defects such as class imbalance problem, storage and calculation overhead. Therefore, this paper explores a new optimized model in this stage, and finally chooses a new classification model which combines KNN and SVM algorithm.

The main evaluation index of classification model is accuracy, that is, the higher the accuracy of automatic classification, the closer result of the expert manual classification it is to. Fig.3 is shown a comparison of the accuracy of the classification model used in this paper and several classical classification models.

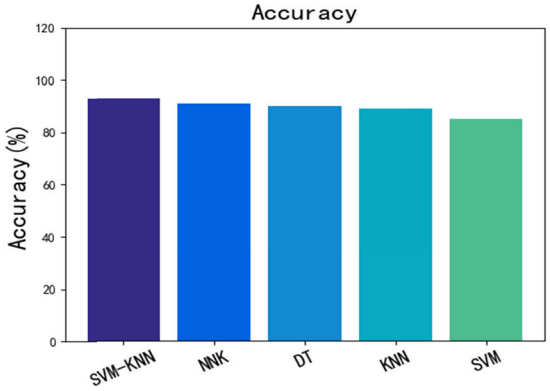

FIGURE 3. The comparison of the accuracy of classification model.

As shown in Figure 3, compared with neural networks and decision trees, the optimized SVM-KNN classification model has the highest accuracy. Therefore, the highest classification model (SVM-KNN) is selected for the experiment in the manuscript. Single-class SVM has a good effect on solving single-class problems, but it is not suitable for multi-class classification problems. The combination of the two classification models solves the class imbalance problem caused by uneven class distribution on the KNN classification model in some extent, which improves not only the classified effect but also the classified speed greatly.

We first use a single-class SVM to generate a classifier for each group in the training set, and then the whole training set as the test sample set is used to test. At last, we use the KNN algorithm as a new training set classifier for a second correction, and achieve the final classification results.

First, the text feature C obtained in the previous step is divided into two parts: a training sample C_{train} and a sample C_{test} with a test classification label. Each classification label represents a knowledge point. Among them, C_{test} is a test that uses expert experience to divide knowledge points manually. In order to facilitate the next operation, the knowledge points (i.e. classification tag) are represented as $K = (K_1, K_2, \ldots, K_i, \ldots, K_p)$, where i represents the i-th knowledge point and p represents the number of knowledge points.

4) CLASSIFICATION RESULT OUTPUT

The training steps of the SVM-KNN model include storing test text features of the training sample C_{train} and corresponding classification labels. The trained SVM-KNN

model can be used to classify the C_{test} samples to be classified. In order to facilitate the next stage of calculation, the classified results of the SVM-KNN model are transformed into a questions-knowledge points matrix QK. The questions-knowledge points matrix QK is obtained by SVM-KNN model classified, and then the students feature knowledge points matrix generated automatically in the next stage is realized by combining subjective weighting method.

C. THE MINING OF STUDENT' LEARNING DEGREE

Students' mastery represents the students' learning degree on the knowledge points corresponding to the test questions, and its level also reflects the level of students' learning degree on the knowledge points. According to the record of students answered, it can analyze and mine the students' learning degree on each knowledge point, that is, the students' mastery on knowledge points. We use the text classification technology to convert the text data of students answered to students' learning degree on each knowledge point of the test questions. The DBSCAN clustering algorithm is used to classify students into several classes with common characteristics and get students' personalized learning characteristics. At the same time, it improves students' learning performance according to the learning suggestions gave. This paper focuses on the important influencing factors in the process of students' answers: the right or wrong answer, the length of the students' answer path length selected and the proportion of the correct choice times in the students' answer path length selected.

The right or wrong answer intuitively judges students' learning degree on knowledge points. The students' answer path length selected reflects students' comprehension of the questions and their psychological activities. The more students choose each question, the worse students' learning degree of the questions. The proportion of correct choice times in the students' answer path length selected reflects the students' learning degree of the questions. The degree of students' hesitation between correct and wrong choices. Combining the proportion of the right or wrong answer, the students' answer path length selected and the proportion of the correct choice times in the students' answer path length selected to get each student' learning degree in each question matrix SQ.

$$\mathrm{SQ} = \begin{bmatrix} sq_{11} & sq_{12} & \cdots & sq_{1n} \\ sq_{21} & sq_{22} & \cdots & sq_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ sq_{m1} & sq_{m2} & \cdots & sq_{mn} \end{bmatrix}$$

where sq_{tj} represents the learning degree of students s_t on test question q_j, $sq_{tj} \in \{0, 1\}$, m represents the number of students, and n represents the number of test questions. When the value of sq_{tj} is higher, it means that student s_t has a better command of the test question q_j.The value of sq_{tj} is determined by the data of influencing factors in the process which students answered questions. The weight coefficient of each influencing factor is given though the authoritative experts in relevant fields according to the degree whether it is direct or indirect influence on them to ensure the authority and correctness of this paper, the students' learning degree on the question q_j by s_t is expressed as follows:

$$sq_{tj} = \beta_1\theta_1 - \beta_2\theta_2 + \beta_3\theta_3 \tag{2}$$

where β_u, $u \in \{1, 2, 3\}$ represents the weight coefficient of each component in the proportion of the right or wrong answer, the length of the students' answer path length selected and the proportion of the correct choice times in the length of the students' answer path length selected. According to the experience of the authoritative experts, the right or wrong answers of students answered are directly proportional to the students' learning degree, the length of the students' answer path length selected is inversely proportional to the students' learning degree, and the proportion of the correct choice times in the length of the students' answer path length selected is proportional to the students' learning degree. Therefore, the sign of the weight coefficient β_2 in formula *(2)* is a negative sign, and the others are positive signs. θ_1 represents the right or wrong answer, θ_2 represents the students' answer path length selected, and θ_3 represents the proportion of the correct choice times in the length of the students' answered path length selected. In this paper, the right answers to the questions q_j is $F_e, F_e \in [A, B, C, D]$, θ_1 indicates whether the final choice F_a is the correct answer. When $F_a = 0$, it indicates that the final choice F_a is wrong, and $F_a = 1$ indicates that the final choice F_a is right. θ_2 indicates the length of the students' answer path length selected L, and θ_3 indicates the number of correct choices w of student s_t on question q_j accounts for the frequency of correct choices in the students' answer path length selected L.

First of all, this paper defines the students' answer path length selected to choose answers on the test questions as $P = <p_1, p_2, \ldots, p_h>$, where $h = 1, 2, \ldots, 6$. According to the statistical analysis of the actual data, the ratio of the students' answer path length selected by students in question q_j over 6 times to the total path length is very small, almost close to 0. Therefore, according to the actual situation of the data, the students' answer path length selected in this paper is set to be the maximum of six times, that is, $h = 6$.

When $F_a = F_e$, it means that the final choice F_a selected by the student s_t on question q_j is correct, which means that the student s_t answers question q_j correctly. When $F_a \neq F_e$, it means that the final choice F_a selected by the student s_t on the question q_j is wrong, which means that the student s_t answers question q_j incorrectly. The final choices of students answered and the students' answer path length selected are divided into the following four situations:

$$\begin{cases} F_a = F_e, & L = 1 \\ F_a = F_e, & L > 1 \\ F_a \neq F_e, & L = 1 \\ F_a \neq F_e, & L > 1 \end{cases} \tag{3}$$

When $F_a = F_e$, $L = 1$, it means that students select the correct answer F_e once, and the learning degree of the student s_t on question q_j is $sq_{tj} = 1$. When $F_a = F_e, L > 1$, it means that students choose the correct answer F_e several times. When $F_a \neq F_e$, $L = 1$, it means that students select once and the selected answer is wrong. When $F_a \neq F_e$, $L > 1$, it means that students choose the wrong answer for several times.

The students' answer path length selected L of the student s_t on the question q_j is, the worse the student's mastery of the knowledge point on the question q_j is. The weight coefficient β_2 in the student's learning degree component uses a negative form. Frequency θ_3 of the correct options in the students' answer path length selected is shown as follows:

$$\theta_3 = \frac{w}{L} \tag{4}$$

where θ_3 represents the correct choice frequency of the student s_t in the length of the students' answer path length selected L on the question q_j, w represents the number of correct answers in the students' answer path length selected. The more correct choices appear in the students' answer path length selected on the question q_j, the higher of the correct choice frequency of the student s_t choose on the question q_j, and the better the learning degree is. The frequency of the correct choice in the length of the students' answer path length selected objectively affects the student s_t to have a higher overall learning degree of the question q_j.

The weight coefficient in the learning degree component is given by the authoritative experts in the relevant fields, which guarantees authority and correctness. The student's learning degree scores obtained by combining the three influencing factors are as follows:

$$sq_{tj} = \begin{cases} 1, & F_a = F_e \text{ and } L = 1 \\ 0, & F_a \neq F_e \text{ and } L = 1 \\ \beta_1\theta_1 - \beta_2\theta_2 + \beta_3\theta_3, & F_a \neq F_e \text{ and } L > 1 \\ \beta_1\theta_1 - \beta_2\theta_2 + \beta_3\theta_3, & F_a \neq F_e \text{ and } L > 1 \end{cases} \tag{5}$$

D. THE CLASSIFICATION OF STUDENT' LEARNING DEGREE CURVE TRAJECTORY

Cluster analysis is an unsupervised machine learning method. This method can automatically find data features and divide data with similar features into a group. The purpose is to make the distance between samples of the same category as small as possible, while the distance between samples of different categories as large as possible. In this paper, learners with similar learning degree can be divided into the same group by using the cluster analysis, while learners with different learning degree can be separated. In this way, each group of learners can be analyzed and processed separately, so that personalized guidance can be implemented smoothly.

Compared with the classical K-means algorithm, the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithm not only does not need to know the number of cluster classes to be formed in advance, but also find clusters of arbitrary shapes. At the same time, it can identify noise points and is not sensitive to the order of samples in the database, that is, the order of input to the pattern has little effect on the results. However, that are at the boundary between clusters for samples, the assignment may fluctuate depending on which cluster is detected first. In this paper, we choose a density-based the DBSCAN clustering algorithm.

According to the degree of each student's learning degree on each test question, the test questions are classified into the knowledge points after text classification, and the learning degree matrix SK of each student on each knowledge point is obtained.

$$\mathrm{SK} = \begin{bmatrix} sk_{11} & sk_{12} & \cdots & sk_{1p} \\ sk_{21} & sk_{22} & \cdots & sk_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ sk_{m1} & sk_{m2} & \cdots & sk_{mp} \end{bmatrix}$$

where sk_{tx} indicates the learning degree of the student s_t on the knowledge point k_x, $sk_{tx} \in \{0, 1\}$, m indicates the number of students, p indicates the number of knowledge points. The larger the value of $\mathrm{sk_{tx}}$, the higher the learning degree of the student s_t on the knowledge point k_x. The value of sk_{tx} is determined by the average value of sq_{ta}, $sq_{tb}, \ldots\ldots, sq_{to}$, which the student s_t answered on the question $q_a, q_b, \ldots\ldots, q_o$ (a total of v questions) that belong to the same knowledge point k_x.

$$sk_{tx} = (sq_{ta} + sq_{tb} + \ldots + sq_{to})/v \tag{6}$$

Students' learning degree on the knowledge points are manifested in three forms: good learning degree, basic learning degree, and poor learning degree. If $sk_{tx} \in (0.7, 1]$, it means that the students have a good grasp of the knowledge point k_x; If $sk_{tx} \in (0.3, 0.7]$, it means that students have a general grasp of the knowledge point k_x, which belongs to the basic grasp; If $sk_{tx} \in (0, 0.3]$, it means that the students have a bad learning degree on the knowledge point k_x, which is mastered poorly.

The students are clustered according to the learning degree of each student on each knowledge point, and the clustering effect is verified by the average of each group of students on each knowledge point.

The pseudo code of the algorithm proposed in this paper is as follows.

IV. EXPERIMENT

A. DATA SOURCES AND EXPERIMENTAL ENVIRONMENT

In order to verify the feasibility and validity of the algorithm, this paper chooses about 180,0000 records of about 3,0000 answered in a large-scale examination of Computer Culture Foundation as the experimental data set, including 50 questions covering 9 knowledge points, each with a score of 2 and a total score of 100 points. The students' answers to each question are collected and recorded by the examination system online. The data set is collected in January 2019 for

Pseudocode of the Proposed Algorithm

Begin:
1. **Input:** test questions and the records of students answered
2. Word Segmentation and stop word filtering on the test questions
3. Extract text features with (1)
4. QK ← classify test text features into knowledge points with the SVM-KNN model
5. Extract the right or wrong answer, the length of the students' answer path length selected and the proportion of the correct choice times in the students' answer path length selected based on the records of students answered
6. For each data:
7. Preprocess the extracted data with (3) and (4)
8. Enter the values of β_1, β_2 and β_3 with (5)
9. For each student:
10. sq_{tj} ← Calculate the learning degree of each student on each test question
11. End for
12. SQ ← establish the matrix of students' learning degree on each test question with (2)
13. End for
14. Combining matrix QK and matrix SQ
15. SK ← turn into the master degree matrix SK of each student on each knowledge point
16. Cluster students' learning degree on each test question using the DBSCAN algorithm
17. **Output:** generate student clusters with the characteristics of students' learning degree

End

testing online in a province of China. In addition, 2608 questions with knowledge points labeled are collected from other examinations related to the course as training samples. The corresponding distribution of test questions and knowledge points in training samples is shown in Fig.4.

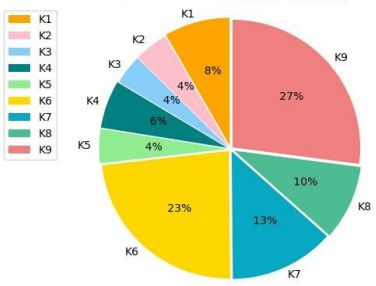

FIGURE 4. The distribution between test questions and knowledge points.

The experimental running environment is Windows 7 Operating System, the programming language is Python 3.6, and the software development environment is PyCharm Community Edition 2018 and SQL Server 2008.

B. PARAMETER SETTINGS

The value of sq_{tj} is determined by the data of the influencing factors of the proportion of the right and wrong answers, the students' answer path length selected and the number of correct choices in the students' answer path length selected. The weight coefficients of each influencing factor are determined by the authoritative experts in the relevant fields according to whether the influence is direct or indirect. It guarantees the authority and correctness of this study.

The influencing factors of student s_t mastery on the question q_j are shown in Table 2.

TABLE 2. Values of influencing factors.

	$L = 1$	$L > 1$
$F_a = F_e$	$sq_{tj} = 1$	$sq_{tj} = 0.8 * 1 - 0.1 * L + 0.1 * \theta_3$
$F_a \neq F_e$	$sq_{tj} = 0$	$sq_{tj} = 0.8 * 0.9 - 0.1 * L + 0.1 * \theta_3$

When $F_a = F_e, L = 1$, it means that the student s_t chooses the correct answer F_e at one time, and then the learning degree sq_{tj} on the question q_j is 1; When $F_a \neq F_e, L = 1$, it means that the student s_t chooses the wrong answer at one time, the learning degree sq_{tj} on the question q_j is 0. β_1 indicates whether the students finally answer the right questions or not, which has the greatest influence on the students' learning degree on the knowledge point, and the value is 0.8. β_2 is the weight coefficient of the path length selected θ_2. The longer the path length is, the worse the students grasp the test questions. θ_2 has little influence on the students' learning degree on the questions, so the value of β_2 is -0.1. β_3 is the weight coefficient of θ_3, which is the proportion of students' correct answers in the path length selected, and θ_3 has little influence on students' learning degree on the question q_j, so the value of β_3 is 0.1. When $F_a = F_e, L > 1$, it means that the correct answer F_e is the answer selected by the students many times and the final answer is the correct answer F_e, and the value of θ_1 is 1; When $F_a \neq F_e, L > 1$, it represents that the answer selected by the students many times and the final choice is the wrong answer, and the value of θ_1 is 0.9.

C. EXPERIMENTS ON TEXT ANALYSIS OF TEST QUESTIONS

In order to judge and process the experimental data objectively and accurately, there are only single choice questions

in the test questions used in the experiment. Before the text analysis stage, we need to deal with the test questions, and take the options and the stem as a complete text. Each question is accompanied by a knowledge point label which the question belongs to. The actual content of the knowledge points corresponding to the label on the knowledge points is shown in Table 3. In the next steps, this paper first performs data processing on multiple-choice questions in the test, and treats the options and questions together as a complete question text, then uses the processed question text to perform knowledge point label matching experiments. All knowledge points are represented by the label on knowledge points.

TABLE 3. Actual content of knowledge points corresponding to the labels of knowledge points.

Knowledge Point Labels	Actual Content
K1	Data communication
K2	Computer network composition, function and architecture
K3	Computer network classification
K4	Computer network basics
K5	Internet IP address and domain name system
K6	The application of the Internet
K7	Protocols common on the Internet
K8	Internet basics
K9	Web page and making

There are no obvious delimiters in the Chinese test, so it is necessary to do the word segmentation and stop word filtering for the test questions. In the word segmentation step, the open source word segmentation tool Jieba is selected; In the stop word filtering step, the mainstream Chinese stop word list is used to stop the vocabulary.

The text feature C is extracted using the formula (1), and the dimension of each text feature C_i is 3367 dimensions. Before the text classification, the text feature C is divided into a training sample C_{train} and a sample to be classified C_{test}. The number of samples in the training sample C_{train} is 2608, and the number of samples in the sample to be classified C_{test} is 50. Next, we used the training sample C_{train} to train the classification model SVM-KNN. That is, the SVM-KNN stores all the feature vectors and the knowledge point tags in the training sample C_{train}.

After the SVM-KNN classification model stores C_{train}, the SVM-KNN model is used to classify the classification sample C_{test}. The model selects the 5 nearest neighbors when classifying, which is also the default value set by the toolkit we use. For the selection of the k value, choosing a smaller k value generally according to the distribution of the samples is equivalent to predicting with training examples in a smaller field. The training error will be reduced, and the algorithm is susceptible to noise, which makes the classification results unstable; If a larger value of k is selected, the model tends to classify the prediction object into a class with a large number of classes. At the same time, too large value also increases the time complexity of the algorithm. In the practical application of this paper, the optimal value of k is selected by using the cross-validation method, and its value is 5. The classification result of SVM-KNN is converted into a test knowledge point matrix QK. Among them, the abscissa represents the test question, and the ordinate represents the knowledge point. The form is as follows:

$$QK = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

In order to observe the classified effect of the sample C_{test} to be classified, we also added the knowledge point label to the 90 questions to be classified. This paper calculates the classified report of SVM-KNN model for sample C_{test} and the Macro Average F1_macro is 0.922222. Finally, the distribution of the nine knowledge points is shown in Fig.5, and the corresponding distribution between test questions and knowledge points is shown in Fig.6.

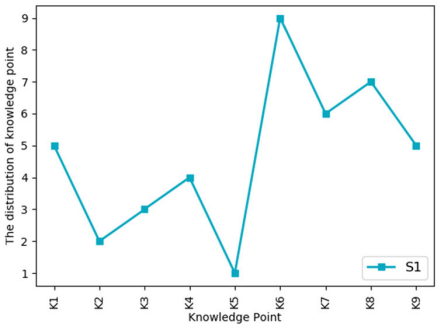

FIGURE 5. The distribution of the nine knowledge points.

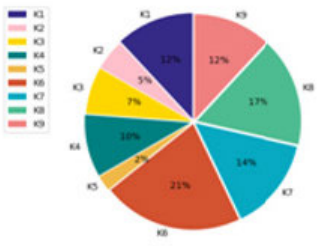

FIGURE 6. The corresponding distribution between test questions and knowledge points.

D. MULTI-FEATURE EXTRACTION OF STUDENTS BASED ON TEST DATA

After processing the data in the student's answering process, the student's learning degree matrix SQ on each knowledge point is expressed as follows according to formula (5). The abscissa represents the test question, and the ordinate represents the students SQ, as shown at the bottom of the next page.

E. THE ANALYSIS OF STUDENTS' MULTI-FEATURE LEARNING DEGREE COMPONENTS

Combines the correct and wrong answer θ_1, the students' answer path length selected L and the correct choice times in the proportion of the students' answer path length selected θ_3 obtained the learning degree sq_{tj} of the student s_t on the question q_j.

Before calculating the learning degree sq_{tj} of the students s_t on the question q_j, the record of the student answered is pretreated as the right and wrong matrix θ_1, the abscissa represents the test question, and the ordinate represents the student. It is as follows.

$$\theta_1 = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & \dots & 1 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & \dots & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & \dots & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & \dots & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & \dots & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & \dots & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & \dots & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & \dots & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & \dots & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & \dots & 0 \end{bmatrix}$$

$$\mathrm{w} = \begin{bmatrix} 2 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & \dots & 1 \\ 0 & 1 & 1 & 2 & 2 & 1 & 0 & 1 & 1 & 3 & 0 & 1 & \dots & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & \dots & 2 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 3 & 1 & \dots & 1 \\ 0 & 1 & 0 & 2 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & \dots & 3 \\ 1 & 0 & 1 & 1 & 2 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & \dots & 1 \\ 1 & 2 & 1 & 1 & 0 & 1 & 2 & 1 & 1 & 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & \dots & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & \dots & 1 \\ 3 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 3 & 0 & 1 & \dots & 1 \\ 0 & 3 & 1 & 1 & 2 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 2 & \dots & 0 \end{bmatrix}$$

The number of correct choices of students is preprocessed as matrix w, where the abscissa represents the test questions, the ordinate represents the students, and w represents as above.

The length of the students' answer path selected is pre-processed as matrix L, where the abscissa represents the test questions, and the ordinate represents the students. It is as shown on the right.

According to formula (4), the proportion of students' correct times in the length of answer path selected is obtained. The data is preprocessed as matrix θ_3, as shown at the bottom of the next page, which is expressed as follows, where the abscissa represents the test questions, and the ordinate represents the students.

$$\mathrm{L} = \begin{bmatrix} 2 & 1 & 1 & 2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \dots & 1 \\ 1 & 1 & 2 & 2 & 5 & 3 & 1 & 1 & 3 & 5 & 1 & 2 & \dots & 1 \\ 1 & 1 & 1 & 4 & 1 & 2 & 3 & 4 & 1 & 1 & 1 & 1 & \dots & 3 \\ 1 & 1 & 2 & 1 & 1 & 1 & 4 & 1 & 1 & 1 & 4 & 1 & \dots & 1 \\ 1 & 2 & 1 & 3 & 1 & 1 & 1 & 1 & 1 & 1 & 2 & 1 & \dots & 3 \\ 1 & 1 & 1 & 1 & 2 & 1 & 1 & 1 & 1 & 1 & 1 & 4 & \dots & 1 \\ 1 & 3 & 1 & 1 & 1 & 1 & 3 & 1 & 1 & 1 & 2 & 1 & \dots & 1 \\ 2 & 1 & 1 & 1 & 2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \dots & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 3 & 1 & 2 & 1 & 1 & \dots & 3 \\ 5 & 1 & 1 & 2 & 1 & 3 & 1 & 2 & 1 & 5 & 1 & 1 & \dots & 1 \\ 1 & 5 & 1 & 1 & 4 & 1 & 1 & 1 & 1 & 1 & 3 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 4 & \dots & 1 \end{bmatrix}$$

F. THE STUDENT' LEARNING DEGREE CURVE TRAJECTORY

Matrix QK obtains the relationship between questions and knowledge points. Matrix SQ obtains each student's learning degree on each test question. Combining matrix QK and matrix SQ using the formula (5), the master degree matrix SK, as shown at the bottom of the next page, of each student on each knowledge point is obtained. The abscissa represents the knowledge point and the ordinate represents the student.

The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm is a density-based spatial clustering algorithm. The significant advantage is that the

$$\mathrm{SQ} = \begin{bmatrix} 0.70 & 1 & 1 & 0.57 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & \dots & 1 \\ 0 & 1 & 0.65 & 0.70 & 0.34 & 0.45 & 0 & 1 & 0.53 & 0.36 & 0 & 0.65 & \dots & 1 \\ 1 & 1 & 1 & 0.43 & 0 & 0.52 & 0.53 & 1 & 1 & 1 & 0 & 1 & \dots & 0.57 \\ 1 & 1 & 0.65 & 1 & 0 & 1 & 0.43 & 0.65 & 0 & 1 & 0.54 & 1 & \dots & 1 \\ 0 & 0.65 & 0 & 0.57 & 1 & 0 & 1 & 1 & 1 & 0 & 0.70 & 1 & \dots & 0.6 \\ 1 & 0 & 1 & 1 & 0.70 & 1 & 1 & 0 & 1 & 1 & 0 & 0.43 & \dots & 1 \\ 1 & 0.49 & 1 & 1 & 0 & 1 & 0.57 & 1 & 1 & 1 & 0.57 & 0 & \dots & 0 \\ 0.65 & 1 & 0 & 1 & 0.65 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & \dots & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0.53 & 1 & 0.65 & 1 & 0 & \dots & 0.53 \\ 0.36 & 0 & 1 & 0.57 & 1 & 0.45 & 1 & 0.43 & 1 & 0.36 & 0 & 1 & \dots & 1 \\ 0 & 0.36 & 1 & 1 & 0.37 & 0 & 1 & 0 & 1 & 0 & 0.45 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0.65 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0.37 & \dots & 0 \end{bmatrix}$$

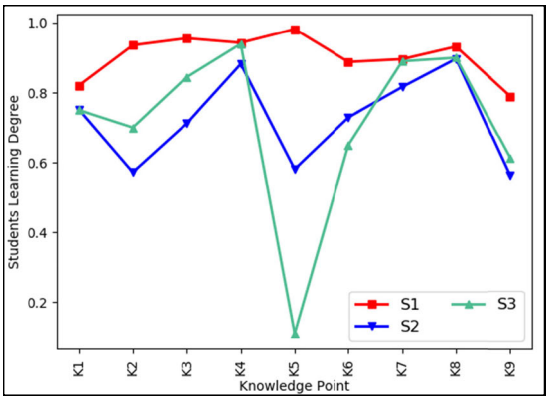

FIGURE 7. The clustered results of students' learning degree.

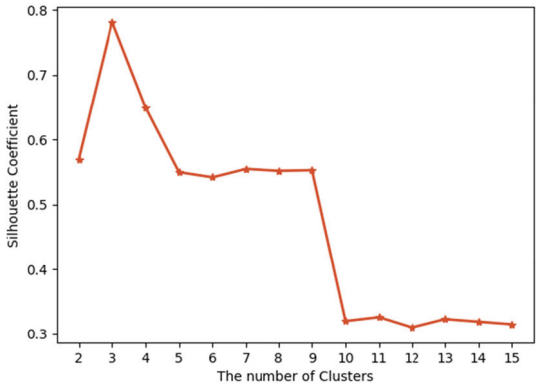

FIGURE 8. The result of the optimal silhouette coefficient.

clustering speed is fast and it can deal with noisy points effectively and find spatial clusters of arbitrary shape. Compared to the clustering algorithms like K-means, the DBSCAN algorithm can cluster dense data sets of any shapes. It can find outliers is not sensitive to outliers in the data set while clustering, and has no bias in the clustering results, which is in accordance with the needs of this paper. However, adjusting parameter is more complicated slightly compared to the traditional K-means clustering algorithm. It is mainly need to make joint adjusting for the neighborhood threshold Eps within the given object radius Eps and neighborhood sample number threshold MinPts, and the different parameter combinations have a great influence on the final clustered effect.

The silhouette coefficient is a method to evaluate the clustered effect. The value of the silhouette coefficient is between $[-1, 1]$, and the closer to 1 it is, the better the cohesion and separation are. When the silhouette coefficient is 0.782 in the experiment, the clustering effect is the best. Students are divided into three curve trajectories with obvious trend according to the best silhouette coefficient, and the experimental results are shown in Fig. 7 and Fig. 9. In this paper, the tuning results of the DBSCAN algorithm in the experiment are shown in Fig. 8. In this paper, the result of the optimal silhouette coefficient selected for the DBSCAN algorithm is shown in Figure 8, and the adjusting results are Eps $= 0.1$ and MinPts $= 10$.

$$\theta_3 = \begin{bmatrix} 1 & 1 & 1 & 0.5 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & \dots & 1 \\ 0 & 1 & 0.5 & 1 & 0.4 & 0.33 & 0 & 1 & 0.33 & 0.6 & 0 & 0.5 & \dots & 1 \\ 1 & 1 & 1 & 0.25 & 0 & 0 & 0.33 & 0.25 & 1 & 1 & 0 & 1 & \dots & 0.67 \\ 1 & 1 & 0.5 & 1 & 0 & 1 & 0.25 & 1 & 0 & 1 & 0.75 & 1 & \dots & 1 \\ 0 & 0.5 & 0 & 0.67 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & \dots & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0.25 & \dots & 1 \\ 1 & 0.67 & 1 & 1 & 0 & 1 & 0.67 & 1 & 1 & 1 & 0.5 & 0 & \dots & 0 \\ 0.5 & 1 & 0 & 1 & 0.5 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & \dots & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0.33 & 1 & 0.5 & 1 & 0 & \dots & 0.33 \\ 0.6 & 0 & 1 & 0.5 & 1 & 0.33 & 1 & 0.5 & 1 & 0.6 & 0 & 1 & \dots & 1 \\ 0 & 0.6 & 1 & 1 & 0.5 & 0 & 1 & 0 & 1 & 0 & 0.33 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0.5 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0.5 & \dots & 0 \end{bmatrix}$$

$$\mathrm{SK} = \begin{bmatrix} 0.8 & 0.5 & 1 & 0.74 & 0.73 & 0.67 & 0.39 & 0.91 & 0.65 \\ 0.42 & 0.71 & 0.475 & 0.57 & 0.18 & 0.94 & 0.84 & 0.68 & 0.79 \\ 0.69 & 0.77 & 0.77 & 0.91 & 0.57 & 0.83 & 0.78 & 0.39 & 0.91 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.33 & 0.73 & 0.66 & 0.90 & 0.66 & 0.67 & 0.78 & 0.69 & 0.65 \\ 0.8 & 0.72 & 1 & 0.97 & 0.70 & 0.69 & 0.96 & 0.56 & 0.91 \\ 1 & 0.79 & 0.76 & 0.72 & 0.78 & 0.89 & 0.76 & 0.59 & 0.2 \\ 0.88 & 0.83 & 0.85 & 0.82 & 0.73 & 0.86 & 0.74 & 0.82 & 0.42 \\ 0.76 & 0.82 & 0.89 & 0.74 & 0.88 & 0.83 & 0.68 & 0.51 & 0.79 \end{bmatrix}$$

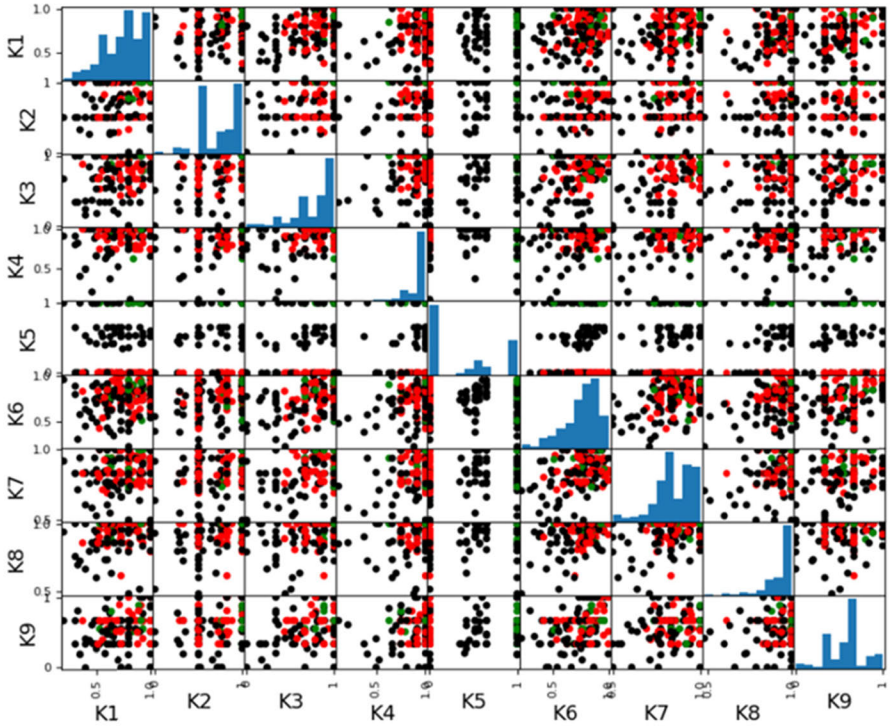

FIGURE 9. Experimental results of the DBSCAN algorithm.

The Fig.7. shows the students' learning degree on each knowledge point. The abscissa represents the knowledge point and the ordinate represents the students' learning degree. The DBSCAN clustering algorithm is used to cluster students into three curves trajectories for different situations. S1 is a stable curve trajectory, S2 is an undulating curve trajectory, and S3 is an irregular curve trajectory.

The stable curve trajectory is subdivided into two types. One is that students master knowledge points well, and the other is that students master knowledge points poorly. The difference between the two is only that students have different learning degrees on knowledge points. In both cases, the curve of students' learning degree on the knowledge points is stable and gentle, and there is little difference in the weight of mastery on each knowledge point. However, there are some points whose weights differ greatly from those of other points. We call some points whose weights differ greatly from those of other points as ''outliers'', but we do not consider ''outliers'' in clustering.

The undulating curve trajectory is curve of master degree on knowledge points. There are several or more consecutive points weight compared to other parts of the weight is lower (or higher), and other parts are similar to the stable curve trajectory S1. Therefore, we call such a curve trajectory a fluctuating one, and we suspect the continuous low (or high) of their corresponding knowledge point may have a potential connection, which requires further study and analysis.

Irregular curve trajectory does not belong to the above two cases, and may also include the above two cases, that is, it includes both partial gradual curve trajectory and partial undulating curve trajectory. The point is that irregular curve trajectory has no regularity.

Calculate the average score of all kinds of students on each knowledge point, and obtain the average score curve trajectory of all kinds of students on each knowledge point. It is as shown in Fig. 10.

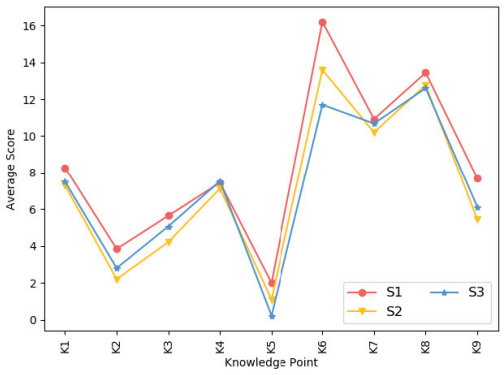

FIGURE 10. The trajectory of average score curve of all kinds of students.

As is shown in Fig.10, the average score of students on knowledge points fluctuates greatly. Due to the different number of questions corresponding to each knowledge point,

it needs to calculate the proportion of the average score on each knowledge point of each group to the total score on each knowledge point.

As is shown in Fig.11, the curve of the average score of students in S1 on the total score of knowledge points is relatively stable. In S2, the proportion of the average score of students on knowledge points to the total score on knowledge points fluctuates greatly, and this is similar to the trend of learning degree of the three groups on each knowledge point obtained by clustering. It can be obtained that the algorithm can classify students according to their characteristics on knowledge points.

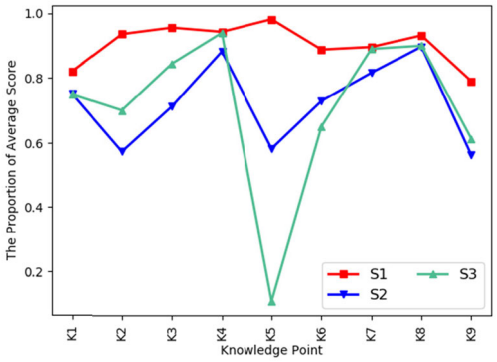

FIGURE 11. The ratio of the average score of students to the total score.

V. CONCLUSION

In view of the current analysis of educational data, it is difficult to consider the limitations of students' learning degree comprehensively and accurately. This paper proposes a data-driven students' learning degree analysis – an optimized algorithm for mining dynamic data. The algorithm uses the optimized text analysis technology to replace classifying test questions into the knowledge points manually, and through the subjective weighting method combined with the dynamic data captured to mine and analysis the students' learning degree on knowledge points. Then the students' learning degree curve trajectory and student clusters are obtained through the DBSCAN algorithm.

Experimental results show that the algorithm has the following characteristics: 1) Automatically process massive data, and using the classification model to classify the test questions into knowledge points labels can improve the classification efficiency though the usual single classification model; 2) The traditional static education data analysis and mining methods are transformed into dynamic data mining of students' individual characteristics to obtain students' learning degree. The method of mining students' learning degree has changed from the traditional static educational data analysis to dynamic data analysis, and it solves the limitation of analyzing the traditional static educational data in the past, and through mining and analyzing the dynamic data in the educational data captured, it makes a more comprehensive and accurate analysis of the students' learning degree on knowledge points.

Although the algorithm performs well in automatically classifying students' learning degree on knowledge points, it also has limitations: 1) Depend on expert experience; 2) Consider only dynamic data, such as whether the student answered the test question correctly, the student's choice of path length for the answer to the test question, and the number of correct options in the student's answer path length selected. Factors such as the time it takes to answer a test question and the difficulty of the test question are not considered. In short, in the future, we will fully consider the various factors that affect the degree of mastery on knowledge points, and obtain students' learning degree more accurately.

ACKNOWLEDGMENT

The authors are grateful to the anonymous reviewers for their constructive comments and invaluable contributions to enhance the presentation of this article.

REFERENCES

[1] R. S. Baker, "Educational data mining: An advance for intelligent systems in education," *IEEE Intell. Syst.*, vol. 29, no. 3, pp. 78–82, May 2014.
[2] C. Heiner and N. Heffernan, "Educational Data Mining," *Stud. Comput. Intell.*, vol. 1, pp. 467–474, Oct. 2014.
[3] R. Asif, "Analyzing undergraduate students " *Perform. Educ. Data Mining Comput. Edu.*, vol. 1, p. 177–194, Apr. 2017.
[4] W. Jie, "Application of educational data mining on analysis of students' online learning behavior," in *Proc. 2nd Int. Conf. Image, Vis. Comput.*, Chengdu, China, vol. 1, 2017, pp. 1011–1015.
[5] Z. Shao, "Research on a new automatic generation algorithm of concept map based on text analysis and association rules mining," *J. Ambient Intell. Hum. Comput.*, vol. 1, pp. 1–13, Oct. 2018.
[6] M. A. Duhayyim and P. Newbury, "Concept-based and fuzzy adaptive E-learning," in *Proc. 3rd Int. Conf.*, vol. 1, 2018, pp. 49–56.
[7] P. Chen, Y. L, and V. W. Zheng, "KnowEdu: A system to construct knowledge graph for education," *IEEE Access*, vol. 6, pp. 31553–31563, 2018.
[8] L. P. Macfadyen and S. Dawson, "Mining LMS data to develop an, early warning system," *Educators, Proof concept. Comput. Edu.*, vol. 54, p. 59, Oct. 2010.
[9] P. Kaur, M. Singh, and G. S. Josan, "Classification and prediction based data mining algorithms to predict slow learners in education sector," *Procedia Comput. Sci.*, vol. 57, pp. 500–508, Oct. 2015.
[10] D. E. Booth, "Data mining methods and models, by Daniel T. Larose," *Technometrics*, vol. 4, pp. 1–50, Oct. 2007.
[11] A. M. Cowan, "Data mining in finance: Advances in relational and hybrid methods," *International Journal of Forecasting*, vol. 181. B. Kovalerchuk and E. Vityaev, Eds. Norwell, MA, USA: Kluwer, 2002, pp. 155–156.
[12] L. Li, H. Tang, Z. Wu, J. Gong, M. Gruidl, J. Zou, M. Tockman, and R. A. Clark, "Data mining techniques for cancer detection using serum proteomic profiling," *Artif. Intell. Med.*, vol. 32, no. 2, pp. 71–83, Oct. 2004.
[13] L. Hirschman, J. C. Park, J. Tsujii, L. Wong, and C. H. Wu, "Accomplishments and challenges in literature data mining for biology," *Bioinformatics*, vol. 18, no. 12, pp. 1553–1561, Dec. 2002.
[14] A. M. Shahiri, W. Husain, N. A. Rashid, "A review on predicting student's performance using data mining techniques," *Procedia Comput. Sci.*, vol. 72, pp. 414–422, Dec. 2015.
[15] B. Åžen, E. Uäar, and D. Delen, "Predicting and analyzing secondary education placement-test scores: A data mining approach," *Expert Syst. Appl.*, vol. 39, no. 10, pp. 9468–9476, Aug. 2012.
[16] Z. Wang, H. He, and M. Jiang, "A comparison among three neural networks for text classification," in *Proc. IEEE Int. Conf. Signal Process.*, Nov. 2006, pp. 1–5.
[17] J. F. Cao and J. J. Chen, "An improved Web text classification algorithm based on SVM-KNN," *Appl. Mech. Mater.*, vols. 278–280, pp. 1305–1308, Jan. 2013.

[18] W. Liu and C. Wu, "A new algorithm for Chinese text classification-One Class SVM-KNN algorithm," *Comput. Technol. Develop.*, vol. 22, pp. 83–86, 2012.
[19] T. Matsumoto, "Data analysis support by combining data mining and text mining," in *Proc. 6th IIAI Int. Congr. Adv. Appl. Inform.*, vol. 1, 2017, pp. 313–318.
[20] B. Agarwal and N. Mittal, "Text classification using machine learning methods—A survey," in *Proc. 2nd Int. Conf. Soft Comput. Problem Solving*, vol. 1, 2012, pp. 701–709.
[21] A. Kurbatow, "The research of text preprocessing effect on text documents classification efficiency," in *Proc. Int. Conf. Stability Control Processes*, vol. 1, 2015, pp. 653–655.
[22] A. Islam, D. Inkpen, and I. Kiringa, "Applications of corpus-based semantic similarity and word segmentation to database schema matching," *VLDB J.*, vol. 17, no. 5, pp. 1293–1320, Aug. 2008.
[23] I. A. El-Khair, "TF*IDF," *Encyclopedia Database Syst.*, vol. 13, pp. 3085–3086, Oct. 2009.
[24] S. T. Selvi, "Text categorization using Rocchio algorithm and random forest algorithm," in *Proc. 8th Int. Conf. Adv. Comput. (ICoAC)*, vol. 1, 2017, pp. 7–12.
[25] D. Bertsimas and A. King, "Logistic Regression: From art to science," *Stat. Sci.*, vol. 32, pp. 84–367, Oct. 2017.
[26] I. Ü. Ogul, C. Özcan, Ö. Hakdagli, "Fast text classification with Naive Bayes method on Apache Spark," in *Proc. 25th Signal Process. Commun. Appl. Conf. (SIU)*, May 2017, pp. 1–4.
[27] E. Sarbazi, M. Uysal, M. Abdallah, and K. Qaraqe, "Ray tracing based channel modeling for visible light communications," in *Proc. 22nd Signal Process. Commun. Appl. Conf. (SIU)*, vol. 1, Apr. 2017, pp. 1–4.
[28] S. Zhang, "Learning k for k-NN classification," *ACM Trans. Intell. Syst. Technol.*, vol. 8, pp. 1–19. Oct. 2017.
[29] W. S. Noble, "What is a support vector machine?" *Nature Biotechnol.*, vol. 24, pp. 1565–1567, Oct. 2006.
[30] M. Melucci, "Vector-space model," *Encyclopedia Database Syst.*, vol. 10, pp. 3224–3440, Aug. 2017.

ZENGZHEN SHAO was born in Weifang, China, in 1976. He received the Ph.D. degree from Shandong Normal University, China. He is currently a Professor with the School of Information and Computer Science, Shandong Women's University. His main research interests include data analysis and prediction, machine learning, and emergency evacuation.

HONGXU SUN was born in 1994. She is currently pursuing the master's degree with the College of Computer Science and Engineering, Shandong Normal University, China. Her research interests include data mining and processing.

XIAO WANG was born in 1994. She is currently pursuing the master's degree with the College of Computer Science and Engineering, Shandong Normal University, China. Her research interest includes big data processing.

ZHONGZHI SUN was born in China, in 1996. He is currently pursuing the master's degree with the School of Information Science and Engineering, Shandong Normal University. His main research interests include education data mining and artificial intelligence.