Analysis and Prediction of CET4 Based on Data Mining

Haiyan Liu(✉)

Xi'an Siyuan University, Xi'an 710038, Shaanxi Province, China
tangyijie3@163.com

Abstract. CET4 analysis and prediction based on data mining is a technology that uses data mining technology to analyze system performance. It can be used as an effective tool to predict the future performance of any system or application. The main goal of this technology is to identify the weaknesses in the system in order to improve it, so as to obtain better results and efficiency. This technology helps improve the productivity of employees and improves their efficiency by identifying areas where they need more training or improvement. Based on the analysis and prediction of CET-4 scores based on data mining, this paper analyzes many factors affecting CET-4 based on multiple linear regression algorithm. According to the idea of data mining, historical data are collected and transformed appropriately, and many factors affecting CET-4 are analyzed by using statistical analysis technology. It is concluded that the linear regression relationship between CET-4 and its influencing factors has a high degree of fitting, which can be used to predict CET-4, and has a great guiding role and use value for students' learning and school teaching management.

Keywords: Data mining · CET-4 · Performance analysis

1 Introduction

In the 20th century, with the increasing updating of database technology, the of database management system has been widely popularized, and the data is growing rapidly., human beings can exchange a large number of data information and work together on the network at different places and at different times. Therefore, what we are facing now is not limited to the huge database information gathered by our departments, units and the whole industry, but a large amount of information knowledge. On the surface, most of the data are chaotic and have no rules to follow. In fact, behind these soaring data are many valuable information. At present, ordinary database systems can basically achieve the functions of effective data query, input and preliminary statistics. However, they can not find a certain relationship from the data, so they can not predict the future development direction according to the current analysis methods [1]. There is a lack of means to mine the valuable, resulting in the phenomenon of "data expansion and lack of knowledge". Behind a large amount of data, there is a large amount of information

M. A. Jan and F. Khan (Eds.): BigIoT-EDU 2022, LNICST 465, pp. 103–113, 2023.
https://doi.org/10.1007/978-3-031-23950-2_12

helpful to human development. If these data can not be well transformed into useful knowledge, assuming that there is no correct means, human beings begin to face the ocean of information, but there is nothing to do. Therefore, people began to consider how to mine effective knowledge from massive information, deal with it more deeply, and improve the utilization rate of knowledge. Facing the challenge of massive data to human beings, people first discover knowledge in databases, and then to its core technology - data mining, which naturally appears, develops rapidly and shows strong vitality [2].

Today's society is in the process of internationalization, globalization, cultural diversity and rapid and vigorous. Among them, the rapid development of educational informatization and the sharp strengthening of the attention of talents and manpower, how to use and make use of science and technology to find favorable information from a large number of cumbersome data, so as to improve the existing teaching methods and strategies to help students improve their grades, which provides a reference basis for educational reform [3]. It is very beneficial to apply data mining technology to the field of education, mine and analyze the data of students' grades, provide important decision-making basis for educational work, and use the results to serve teachers and students and make them carry out more effective teaching activities.

In the constitution of the Teaching Steering Committee, it is pointed out that we should adhere to the "people-oriented", promote the "four returns" and comprehensively improve the talent training ability of colleges and universities [4]. CET-4 is to assess students' Comprehensive English ability and make scientific prediction. These data can not be used only for query or simple statistics. Combined with my major and work, this paper uses the method of statistical analysis to find out the relevant factors affecting the passing of CET-4, and find out some factors affecting whether college students pass CET-4, so as to provide practical and effective suggestions and measures to help students learn English well and use English well, At the same time, it is of great practical significance to improve teaching quality and teaching methods for teachers [5].

In the notice of the, reasonably improve the academic challenge, increase the difficulty strictly enforce the examination discipline, strictly control the exit of graduation, and resolutely cancel the "clear examination" system. It can be concluded that if college students want to graduate smoothly, they need to learn the required courses well [6]. Otherwise, if some courses fail to pass the exam, they will face the risk of not graduating without "clear examination". Therefore, it is very necessary to identify which courses have a great impact on students' graduation, and provide early warning to students with poor grades and the risk of re studying or dropping out of school as soon as possible to help students graduate smoothly, and assist the school management department to formulate effective management methods to help students. Through the students' scores in the first four semesters during the school period to predict the students' graduation, obtain the distribution of students' scores, analyze the students' scores in various subjects through methods, classify and analyze the students' graduation, and establish relevant prediction models. It is of great practical perfect the problems existing in the existing teaching.

2 Related Work

2.1 Overview of Data Mining

Data mining is a process of extracting hidden, unknown but potentially useful information and knowledge from a large number of incomplete, noisy, fuzzy and random data by using various analysis methods and technologies. In other words, data is extracted from a database called knowledge retrieval (KDD). Data mining is an interdisciplinary subject integrating artificial intelligence, database technology, model recognition, machine learning, statistics and data visualization. At present, it is widely used in finance, retail, insurance, medicine, communication, electronics, aviation and other fields. One of the most common standards in the industry is the cross industry data extraction process. According to this standard, the process of data mining is divided into six stages: 1) understanding the operation, 2) understanding the data, 3) preparing the data, 4) modeling, 5) evaluation and 6) deployment. Modeling must be based on different business scenarios, namely [7].

2.2 Data Mining Steps

The ultimate purpose of data mining is application-oriented. Only through application can we reflect its value and promote the development of data mining. The successful application of data mining in various fields has attracted more and more attention. Its main applications are divided into two categories: analyzing the hidden laws of data and practical support in decision-making. For example, cheating inspection, background analysis, competitive sports skills, risk analysis and management, customer relationship management (CRM), web analysis based on data mining and target market for consumption are being applied. In addition, some new fields have emerged, such as stream data mining and DNA and biological data analysis [9].

Among various applications, the application of knowledge mining (KDD) in database has become the core of data mining. Because data mining, the data of data mining basically comes from the database, development direction. The basic steps of database knowledge mining are shown in Fig. 1 below.

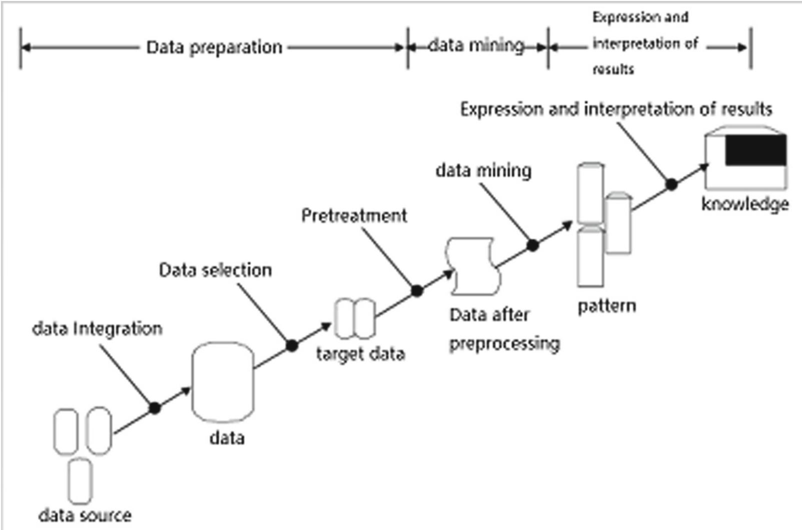

Fig. 1. Basic steps of knowledge mining

To carry out data mining, we must know the basic steps of data mining. The first step is to understand the application fields of data mining, including relevant knowledge and application objectives, so as to know which knowledge or patterns you are interested in, and only after understanding relevant knowledge can you reasonably process; The second step is to select data - data transformation; The third step is to select the function of data mining and the algorithm of data mining. This step is the key step of data mining, which determines the performance and effect of data mining; The fourth step is to find the model of interest, eliminate the misleading model, evaluate the model and express knowledge; Finally, use the excavated knowledge.

is shown in Fig. 2. Its bottom layer, that is, the object of data mining, that is, the source of data. There are many types of databases, but the most perfect development in technology is relational database, and the related tools are also the most, so relational database is the most used in data mining. The upper layer of the database, which is responsible for data storage, query, update, transaction processing and other related functions. It is responsible for transforming the data mining engine on the server into the data mining engine on the server. After data mining obtains the pattern of interest, it is to evaluate the pattern. It is responsible for comparing with the knowledge base to see whether the pattern meets the conditions [8]. This process is determined by the upper layer of the data mining engine, that is, the pattern evaluation layer. Finally, the interested patterns should be presented to users in a way that users can understand.

2.3 Factor Analysis of CET-4

Although colleges and universities have abolished the linkage between CET4 scores and degree certificates, this does not mean that CET4 scores are no longer important. CET-4 score is an important index to measure a student's English level. It can reflect

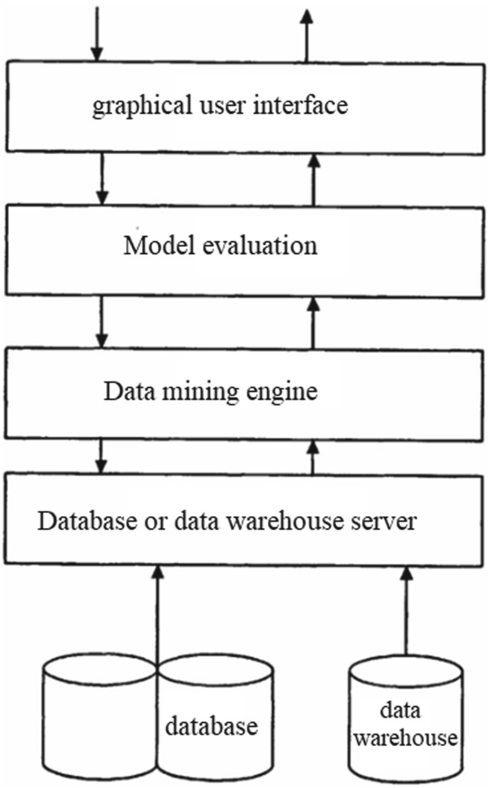

Fig. 2. Structure of data mining system

a student's comprehensive English level. In addition, the passing rate of CET-4 can also reflect the level. The content and form of English test also have certain guidance for College English teaching. Colleges to CET-4 achievement test. It can be seen that factors affecting the achievement of CET4. Therefore, in this chapter, this paper will take the relevant information of college students, the scores of four college English tests and CET-4 as the research data, and study and analyze the relationship and closeness between students' CET-4 passing and various factors [10].

2.4 Data Preparation

In order to analyze the factors related to CET-4 scores, this paper collects the relevant information of students from three colleges of a university. The information includes the following contents: gender, major, place of origin, entrance English scores, college entrance examination scores, four positive College English test scores during the school period, and CET-4 scores (including listening, writing and reading scores) during the four semesters during the school period. There are a total of 2821 pieces of data, each of which has the following characteristics: gender, major, college, place of origin, admission

English, college entrance examination results, course No. 1, English score 1, course No. 2, English score 2, semester, listening, writing, reading and grades.

In terms of data adoption, data sorting is carried out based on the particularity of data. On the one hand, some students pass CET-4 for the first time. Although these students have not completed all college English courses, if they still use the future English scores to analyze and predict the passed CET-4 scores when analyzing the data, it will be unreliable and unreasonable; On the other hand, some students may need to take CET-4 for many times to pass. Therefore, the data of students who repeatedly take CET-4 are regarded as independent data in this paper. As shown in Fig. 3.

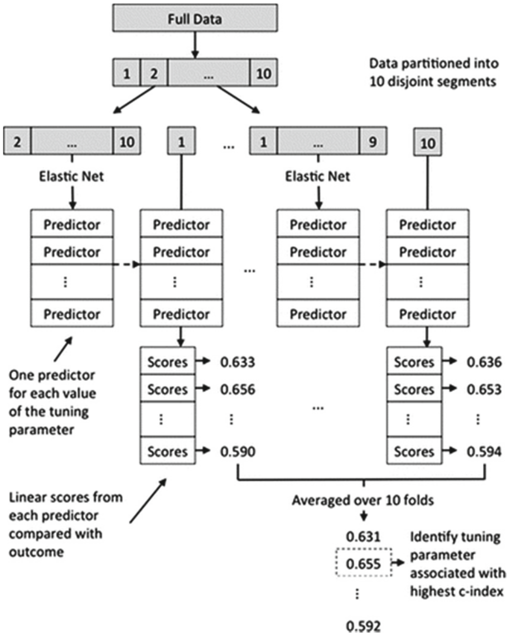

Fig. 3. Prediction model of CET-4 scores

2.5 Data Preprocessing

The original data set collected often contains some incomplete and noisy data. Such as abnormal value and missing value. In addition, the data set may contain some data that is not suitable, data sets need to be preprocessed such as data cleaning and data transformation. The specific data preprocessing is as follows:

First, use python programming to make a simple statistical analysis of the data to see which features contain missing values and abnormal values. The results are as follows:

As shown in Fig. 4 above, students' "entrance English scores" and "college entrance examination scores" have missing values. After calculation, the proportion of missing values in the total data is 2.37% and 2.06% respectively. Through further inspection, 118 of them are of grade 2013 and 5 of them are of grade 2014. Considering that "entrance English score" may have a strong correlation with CET4 score, and in order not to cause data imbalance, this paper temporarily retains the two characteristics of "entrance English score" and "college entrance examination score". It can also be found in the above table that the lowest score of "college entrance examination score" is 329, which is significantly lower than 400. Since the colleges and universities studied in this paper are provincial key colleges and universities, their enrollment scores should not be lower than the provincial key score line. It can be seen that the college entrance examination scores below 400 should be abnormal data. After checking the abnormal data, it is found that the students corresponding to the data are from Jiangsu Province and Shanghai. Through verification, the total scores of the college entrance examination scores of these two provinces and cities in that year are 480 and 600 respectively, which is different from the full score of 750 in other cities. In order to keep the data consistent, this paper needs to convert the student scores of Jiangsu Province and Shanghai.

count	2763	2754	2821	2821	2821	2821	2821	2821	2821
mean	67.06207021	533.2377563	72.69266218	69.09961007	3.064870613	143.153492	132.2860688	149.471464	424.911024
std	20.14536335	28.04436905	13.98852795	12.11898765	0.907245984	23.5255598	17.60689886	27.77197522	54.2560866
min	0	329	14	0	2	0	0	0	238
25%	59	520	63	62	2	127	121	131	389
50%	70	536.09212	77	68	3	142	132	151	424
75%	80.5	549.108115	83	77	4	157	144	170	460
max	109	601	99	99	4	220	200	223	616

Fig. 4. Statistical information table.

The specific calculation formula is as follows:

Jiangsu students' scores after processing = college entrance examination scores $\times$ 750 $\div$ 480.

Shanghai students' scores after processing = college entrance examination scores $\times$ 750 $\div$ 600.

In order to better understand the correlation between the adopted data and CET4, as well as the importance of each feature. From the perspective of correlation, this paper calculates the correlation coefficient of each data feature, as shown in Fig. 5:

Because the Pearson correlation coefficient between "college entrance examination scores" and CET-4 scores is only 0 081, so it is considered that there is basically no inevitable connection between CET-4 and college entrance examination scores. This -- conclusion is inconsistent with Zheng Lidan's conclusion that college entrance examination scores determine students' CET-4 scores. The Pearson correlation coefficient between entrance English score and CET4 score is 0.39, and there is a weak correlation between them. This feature is retained in this paper for better analysis and prediction. The correlation coefficient between reading and CET-4 is 0.84, which shows that there is a great relationship between CET-4 and reading score, that is, reading comprehension has a great impact on whether students can pass CET-4. The correlation coefficient between "English score 2" and CET4 is 0.62, which is higher than that between "English score 1" and CET4 58. This is consistent with the inference of this paper. College English

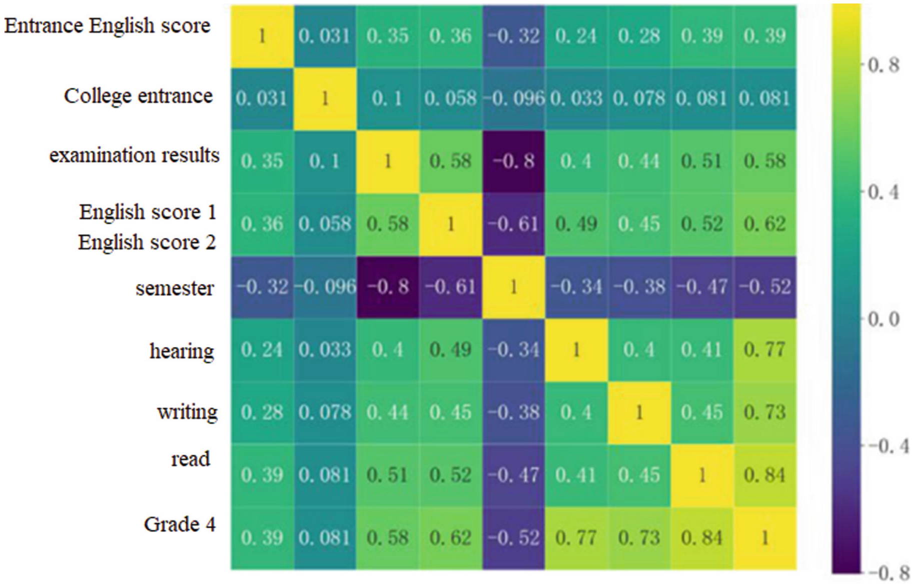

Fig. 5. Pearson correlation coefficient

scores that are longer away from CET-4 have less impact on CET-4 scores. In order to accurately understand the relationship between each factor and CET-4, the relationship between each factor and CET-4 will be analyzed one by one. The correlation between "semester" and CET-4 is -0.52, indicating that there is a strong negative correlation between "semester" and CET-4. It can be seen from this that the higher the "semester", the worse the grade of CET-4. That is to say, the later the students take CET-4, the less likely they are to pass CET-4.

3 Analysis and Prediction of CET4 Based on Data Mining

When using k-nearest neighbor algorithm for prediction classification, it is necessary to determine the appropriate number of nearest neighbors K and select important features. In order to make the classification accuracy higher and the classification performance more stable, this paper uses an iterative method to select k and important features. Firstly, take all features as input variables to determine the nearest neighbor number k of several, then reduce the features successively under the condition that the K value remains unchanged, select important features according to the standard of prediction effect, and finally take the important features as input variables to determine the appropriate K value.

Using python programming, the curves of accuracy, accuracy, recall and F - value in the above 8 cases are obtained. The accuracy reflects the ability of the classification model (algorithm) to correctly classify and predict, including predicting positive cases as positive cases and negative cases as negative cases. The accuracy rate reflects the ability of how many of the predicted positive cases are real positive cases, while the recall rate reflects the ability of how many of the positive cases can be correctly classified and predicted.

As shown in Fig. 6, the accuracy in eight modes increases with K-shaped amplification. $K \geq 11$, the accuracy is basically stable under all working conditions With the increase of potassium content, some changes have taken place, but the change is very small, usually no more than 2%. Case 1 has the largest input property, but the prediction accuracy is not the highest in terms of display accuracy. Although the input property is very small, the average accuracy of Sect. 11 prediction can be $\leq K \leq 29$, and the variation range is relatively small. In the accuracy chart and inverse modulus chart, the variation range of each curve is larger than that of the accuracy chart and f chart, indicating that the number and nature of the recent K have a great impact on the prediction accuracy and callback rate. In these eight cases, the precision, recall coefficient and F value of condition 8 are not the highest, but the precision, recall coefficient and F value of condition 8 are relatively high. As for the purpose of this prediction, I want to predict students' grades in College English. Finally, considering the importance of indicators to the prediction of College English test results, only three characteristics of English entry, English 1 and English 2, are selected.

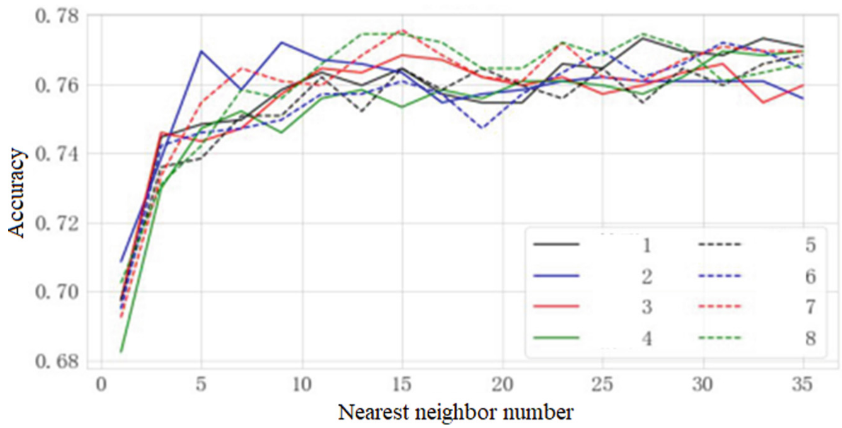

Fig. 6. Accuracy chart

Considering that the relevant scores have been standardized, and the distance between samples is less than 3, this paper selects Gaussian function as the function of weight. The specific steps of realizing weighted k-nearest neighbor classification are as follows:

(1) Initialize the value of K;
(2) Loading data;
(3) Calculate the Euclidean distance between the samples to be classified and the training data;
(4) Sort the distance values in ascending order;
(5) Extract the first k distances D1, D2,…, DK from the sorted array;
(6) Using Gaussian function, the voting weights W1, W2,…, WN of the first k distances are calculated;

(7) Win the sample type corresponding to the initial distance K, and calculate the total weight of each category
(8) Go back to the results of the plan.

In order to predict the results of CET-4 more accurately, the data are processed in advance, the classified variables are converted into factors, and the quantitative variables are standardized. Then select the input prediction function and nearest neighbor K according to various indicators. The nearest neighbor classification model of K is established Finally, a K-Neighborhood subclass weighted model is established to accelerate the classification of CET4 results.

English scores have the greatest influence on CET-4 scores and play a decisive role. The teaching effect is not ideal simply according to the quality of English performance, because there are many reasons for the quality of English performance. For example, some students have a good foundation, but do not study after enrollment, resulting in poor English performance and CET-4 can not pass; Some students can't keep up with the teaching progress because of their poor foundation. Although they study hard, their English grades are still not very good, and they can't pass CET-4. This paper lists English achievement as a reference attribute and studies the influence of other factors on CET-4 achievement. Among the factors that affect CET-4 academic performance, students' entrance performance has the most significant impact on CET-4 performance in addition to English performance factors, because students with good entrance performance have a relatively solid foundation, can actively study, and a relatively high pass rate of CET-4 examination. The impact of public course scores on CET-4 scores is weaker than that of admission scores, but stronger than that of professional course scores on CET-4 scores. The reason is that students pay attention to the study of professional courses and invest less in public courses, while English is a public course.

4 Conclusion

Firstly, this paper points out the importance of CET-4 to college students, then expounds the concepts of algorithm, and focuses on the principle. Finally, some theories of data mining are applied influencing factors and prediction of College Students' CET-4. It has clear conditioning, in-depth analysis and close to reality, which has certain reference value and guiding significance. The main features are: (2) the data comes from the accumulation of many years of work, which is true and effective; (2) It can skillfully transform and process individual field data, making the analysis more illustrative and universal; (3) The analysis and prediction of CET-4 has certain practical value for students, teachers and teaching managers. Of course, not all the influencing factors are taken into account, such as students' family situation and English teachers' teaching situation, which leads to a certain gap between the predicted and actual results of very few students. In the future actual prediction, the influencing factors should be taken into account as much as possible to reduce the error. However, the model will also become relatively complex at that time.

References

1. Wang, S.H., Ji-Rong, L.V., Feng, B.: Analysis and prediction of performance of CET4 based on data mining. Comput. Knowl. Technol. (2014)
2. Wang, H.: Analysis and prediction of cet4 scores based on data mining algorithm. Complexity **2021**(12), 1–11 (2021)
3. Liang, S.W.: Research and analysis of the accuracy of listening question of cet4 based on the multi-media environment
4. Qiao, Q.S., Wang, H.T., Wang, Z.Y., et al.: CET4 passing rate analysis based on fuzzy decision tree induction and active learning. In: International Conference on Machine Learning & Cybernetics. IEEE (2011)
5. Reflection of the Reform in CET4 writing based on language testing theory. 2015(8), 2 (2015)
6. Xin, Z., Shuang, G., Yu, S.: A study on Post-CET4 college English teaching model based on intercultural communication
7. Zhang, S.Z., Shang, L.L., Lu, X., et al.: The analysis and discussion of CET4 scores based on data statistics. J. Jiamusi Vocational Institute (2017)
8. Paterva Evolution Data Mining for the Google Generation. https://www.facebook.com/Symantec
9. The Effects of MI Theoy-based Learning Strategies on Vocational College Students'CET4 Vocabulary Learning
10. Yan, X.U., Zheng, D.F.: Based on CET4 and CET6 new topic of the listening comprehension ability training. Sci. Technol. Vis. (2017)