Application of Data Mining in Higher Vocational Recruitment

Xinyi Zhou[a], Chuanmin Su[b]

[a]178342361@qq.com, [b]214847241@qq.com

Dalian Vocational and Technical College, Dalian, China

Abstract. With the rapid advancement of information technology, data mining has emerged as a crucial approach to extracting valuable information and knowledge. Given the continuous growth of higher vocational enrollment and the intricacy of enrollment processes, the utilization of data mining techniques has garnered increasing attention. This paper aims to investigate the application of data mining technology in the enrollment process of vocational colleges, with the goal of offering more scientifically-informed strategies for vocational college admissions and enhancing the accuracy and efficiency of enrollment practices.

Keywords: Higher Vocational Education, Admissions, Data Mining, Decision Tree

1 Background introduction

In January 2019, the Chinese State Council issued a significant document called the "National Vocational Education Reform Implementation Plan" (referred to as the "Plan" hereafter). This document marked a groundbreaking step towards the development of vocational education. While the expansion of enrollment 20 years ago in 1999 aimed to relax the "threshold" of higher education, the expansion of higher vocational education over the past two decades has provided an opportunity for deepening reforms in this field[1]. The "Plan" emphasizes the shift from pursuing scale expansion to improving quality, moving away from a general education model towards greater enterprise and social participation, and transforming vocational education to have distinctive professional characteristics. It aims to significantly enhance the modernization level of vocational education in the new era. The goal of this reform is to provide high-quality vocational education to society, supply technical and skilled personnel for key national industries, and achieve a parallel expansion and improvement in the quality of vocational education.

EIMT 2024, March 29-31, Wuhan, People's Republic of China

DOI 10.4108/eai.29-3-2024.2347626

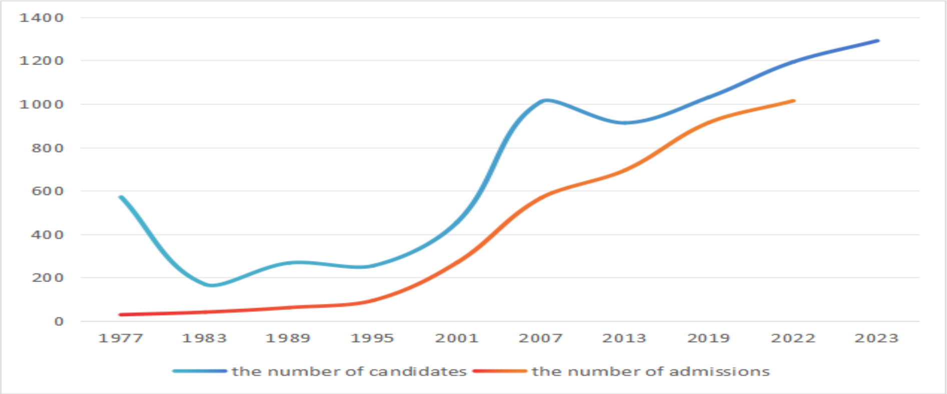

Fig.1 Number of College Enrolment and College Enrolment in China from 1977 to 2023

The enrollment of higher vocational colleges is not only vital for the development of these institutions but also an integral part of the national talent development strategy[2]. With the expansion of higher vocational education[Fig.1], the enrollment of higher vocational colleges faces various challenges. These challenges include the diversity of candidates' backgrounds, regional economic development imbalances, and differing social perceptions of vocational education. These factors can have an impact on the enrollment outcomes. Consequently, higher vocational schools are faced with the task of using information technology effectively to improve the enrollment rate and enhance enrollment decision-making. Data mining techniques, particularly decision tree algorithms, have been extensively employed across various fields due to their effectiveness in processing large-scale data sets, identifying potential patterns, and predicting outcomes. In the context of vocational college enrollment, data mining can not only aid in understanding the key factors that influence enrollment, but also predict future enrollment trends. It can provide a scientific basis for formulating enrollment policies and optimizing resource allocation[3].

2 Overview of data mining technology

Data mining technology is an interdisciplinary discipline integrating computer science, statistics, and artificial intelligence, aimed at extracting valuable information and knowledge from large-scale data. Its development is closely linked to the advancement of computer technology and data storage methods, providing substantial support for various applications.

Data mining involves the systematic processing of vast amounts of data to uncover reliable, innovative, valid, and comprehensible patterns. It draws upon research findings from machine learning, pattern recognition, databases, statistics, and artificial intelligence. When applied to data mining, researchers face a plethora of credible and valid data sources. The objective is to interact with users or knowledge databases, and discover the knowledge and patterns that they desire. The extracted knowledge must be articulated in natural language, facilitating comprehension and practical application.

It is important to stress that data mining results are obtained based on specific premises and constraints, yielding valuable knowledge applicable to a particular field. These results are often relative as different premises and constraints may lead to diverse mining outcomes.

The data mining process[Fig.2] consists of three main stages: data preparation, data mining, and result presentation and interpretation. During the data preparation phase, raw data is initially collected from various sources. This data is then thoroughly cleaned and integrated to ensure its quality and consistency. This step is crucial as it ensures that the subsequent analysis is based on reliable and accurate data. The data is also carefully selected and transformed to make it suitable for mining purposes. This involves formatting and structuring the data in such a way that it can be effectively analyzed and processed by the mining algorithms. Additionally, the data is divided into training and test sets to facilitate model construction and evaluation.

The data mining phase is the central stage of the process where the actual model building and mining take place. Various machine learning and data mining techniques are employed to analyze the data and identify patterns and relationships within it. Commonly used techniques include classification, clustering, and association rule mining. These techniques enable the discovery of valuable insights and knowledge hidden within the data. By applying these technologies, significant and meaningful information can be extracted, contributing to informed decision-making and problem-solving.

The final phase, known as result presentation and interpretation, involves visualizing and interpreting the obtained results. This is important as complex models and findings can be easily communicated through intuitive and understandable visual representations. Visualization techniques help present the results in a format that allows for easy comprehension and analysis by stakeholders. Moreover, the results can be used as a basis for further analysis and decision-making. Adjustments and improvements can be made based on the insights gained from the results, effectively enhancing the overall data mining process.

In conclusion, the three primary stages of the data mining process serve distinct functions in data preparation, data mining, and result interpretation and expression. Through the systematic execution of these stages, valuable knowledge and information can be effectively discovered from vast datasets.

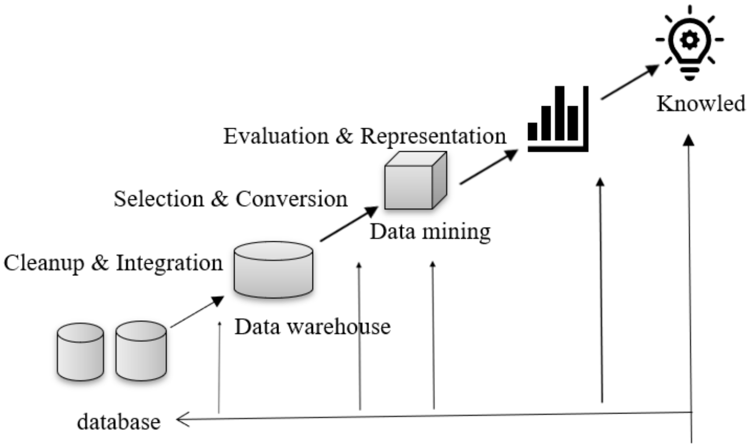

Fig.2 The Process of Data Mining

3 Application of decision tree algorithm in data mining

The decision tree algorithm is a widely used machine learning algorithm in the field of data mining that is based on a tree-like structure. It constructs a tree structure by dividing a dataset to classify and predict data. In the analysis of enrollment data for vocational colleges, the decision tree algorithm can be employed to identify factors that have a significant impact on admission outcomes for candidates. Additionally, it can generate results that are interpretable and easy to comprehend. Specifically, in the context of higher vocational enrollment, the decision tree algorithm can create a decision tree model based on the candidates' personal information (such as gender, age, household registration, etc.), their college entrance examination scores (e.g., Chinese, mathematics, foreign language, professional course scores, etc.), and their voluntary selections (e.g., first choice, second choice colleges and majors, etc.). This model can then reveal the influence of various factors on enrollment outcomes, offering a scientific foundation for the formulation of enrollment policies for higher vocational colleges.

In the process of constructing a decision tree[Fig.3], the algorithm selects the best feature from the dataset as the dividing criterion for the current node. This selection is made by calculating the information gain, which measures the reduction in uncertainty after splitting the data based on a particular feature. The dataset is then divided into subnodes based on the value of the selected feature. This process is repeated for each child node until a termination condition is met, such as the number of samples in the node falling below a predetermined threshold or all available features being utilized. The end result is a fully grown decision tree that can be used to classify or predict new samples based on their characteristics.

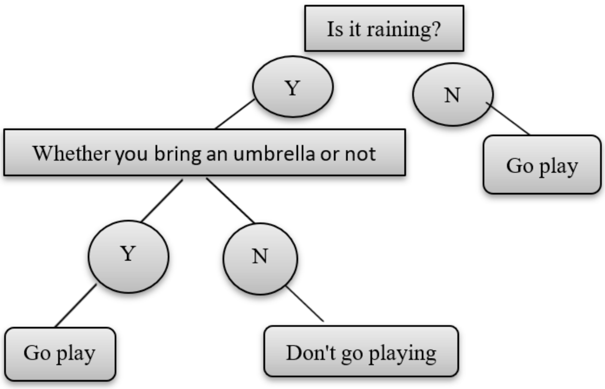

Fig.3 A typical decision tree

The C4.5 algorithm is a fundamental component of the decision tree algorithm, employing the concept of information gain for feature selection[4]. Initially, the C4.5 algorithm computes the information gain for each feature, subsequently choosing the feature with the highest information gain as the partition feature for the current node. This process is then recursively applied to construct the subtree until the termination condition is met. The C4.5 algorithm offers notable advantages including high efficiency, accuracy, and interpretability, rendering it extensively utilized in practical data mining applications. Below are several key concepts and formulas associated with the C4.5 algorithm:

Information entropy: Represents the uncertainty of information, and the formula is defined as follows:

$$Ent(D) = -\sum_{i=1}^{n} p_i \, log_2 p_i \tag{1}$$

Information gain is a measure that quantifies how much the information entropy decreases before and after splitting the dataset. It is utilized to assess the relevance of a feature in the classification task. The formula for information gain is as follows:

$$Gain(D, a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v) \tag{2}$$

Information gain ratio: To address the potential bias of information gain towards features with larger numbers of distinct values, the C4.5 algorithm incorporates the concept of gain ratio. The formula for gain ratio is defined as follows:

$$Gain - ratio(D, a) = \frac{Gain(D,a)}{IV(a)} \tag{3}$$

where IV(a)represents the intrinsic value of feature a, defined as:

$$IV(a) = -\sum_{v=1}^{V} \frac{|D^v|}{|D|} log_2 \frac{|D^v|}{|D|} \tag{4}$$

The higher the gain ratio, the more significant the contribution of the feature to the classification task, and it helps compensate for the issue of having a larger number of feature values[5].

4 Application of data mining in higher vocational enrollment

Using the enrollment and admission data of a higher vocational college in Dalian over the past five years as the research subject, the enrollment and admission source data is exported from the recruitment system. It consists of multiple attributes, including a wide range of fields such as the candidate's name, gender, ID number, admission ticket number, date of birth, ethnical code, political appearance code, candidate's subject code, graduation category code, home address, postal code, professional volunteer, file score, admission score, and more.

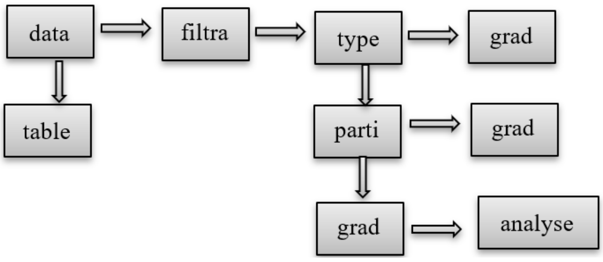

Fig.4 The enrollment data mining flow chart

Data preprocessing[Fig.4] is a crucial step that needs to be undertaken prior to data mining[6]. The selection of suitable methods and techniques for data preprocessing depends on a thorough understanding of the enrollment business characteristics, the objectives of enrollment data mining, and an understanding of the source data itself.

To begin with, data integration is necessary to combine data from various sources into one comprehensive dataset. During this process, issues such as data format, data naming conventions, and data units need to be resolved to ensure data consistency and availability. We will merge the enrollment data from the past two years to conduct a comprehensive analysis. During the merging process, we will add a field called "Admission Method" to differentiate between "Single Admission" and "Unified Admission" in order to reflect the impact of different admission systems.

Following data integration, data cleansing is required to address missing values, outliers, and duplicate values in the dataset. Missing values can be dealt with through techniques such as filling, deleting, or interpolating. Outliers can be handled statistically or based on domain knowledge, while duplicate values can be deleted or merged. With regard to the data mining task, we will eliminate certain fields that are irrelevant to the task, such as admission ticket number, date of birth, and contact number. Additionally, during the data cleansing process, we need to handle specific fields, like the "check-in or not" field. Since this field is manually filled during new student check-in, there may be a few instances of missing data. For these missing values, we will adopt either a deletion or population strategy. Meanwhile, the data regarding the retention of student status will be processed as "yes," while the data regarding withdrawal will be processed as "no," in order to ensure data integrity and accuracy.

The next step involves transforming non-numeric data into numeric data. For instance, gender attributes such as "male" and "female" can be converted into 0 and 1 respectively, while ethnic codes can be transformed into their corresponding numeric codes. This conversion facilitates subsequent data analysis and model building. Furthermore, when processing score data, since scores are continuous and there is a significant disparity between individual entrance examination scores and general entrance examination scores, we will initially standardize the scores. Subsequently, we will discretize the scores and categorize them into three grades: A (excellent), B (moderate), and C (poor). This will aid in conducting further data analysis.[Table1]

Data effect after preprocessing enrollment data:

Table 1. Preprocessed datasets.

serial number	gender	ethnic group	Political outlook	Category	Candidate Category	grades	Professional volunteering	Admission Method:	Whether it is reported or not
1	0	Han nationality	The masses	literature and history	town	C	politics of the firm	Alternative Route to Vocational Higher Education	yes
2	1	Han nationality	The masses	literature and history	town	A	Big Data and Accounting	Unified recruitment	yes
3	0	Han nationality	League member	literature and history	town	A	Financial Services Management	Unified recruitment	no
4	0	minority nationality	The masses	science and engineering	rural area	C	Architectural engineering design	Alternative Route to Vocational Higher Education	yes
5	1	Han nationality	The masses	Vocational education	rural area	B	Hotel Management	Alternative Route to Vocational Higher Education	yes
6	1	minority nationality	The masses	Vocational education	rural area	A	Big Data and Accounting	Alternative Route to Vocational	yes

								Higher Education	
7	1	Han nationality	League member	Art category	rural area	A	preschool education	Unified recruitment	yes

A decision tree was established, and according to the C4.5 algorithm, 5646 freshman data were obtained after processing the average of the total number of two years, and the information amount, information gain, and information gain rate of each attribute in the enrollment data were calculated. Taking the check-in situation as the classification attribute as T and F, T=5284, F=362, the candidate category is selected as the test attribute to establish the decision tree, and the amount of information obtained from the decision-making attribute "check-in status" is:

$$Info[362,5284] = -\frac{362}{5646}log_2\left(\frac{362}{5646}\right) - \frac{5284}{5646}log_2\left(\frac{5284}{5646}\right) = 0.34357 \tag{5}$$

"Candidate Category(kslb)" attribute, urban and rural candidates' reports: the number of unregistered F and registered T are [145, 2201] and [217, 3083], respectively, and the amount of information is calculated as follows:

$$Info[145,2201] = -\frac{145}{2346}log_2\left(\frac{145}{2346}\right) - \frac{2201}{2346}log_2\left(\frac{2201}{2346}\right) = 0.33458 \tag{6}$$

$$Info[217,3083] = -\frac{217}{3300}log_2\left(\frac{217}{3300}\right) - \frac{3083}{3300}log_2\left(\frac{3083}{3300}\right) = 0.34989 \tag{7}$$

The information entropy of the candidate category (KSLB) is calculated:

$$E(kslb) = \frac{2346}{5646} \times Info[145,2201] + \frac{3330}{5646} \times Info[217,3083] = 0.34538 \tag{8}$$

To calculate the information gain and information gain ratio for each attribute：

$$Gain(kslb) = Info[362,5284] - E(kslb) = 0.34357 - 0.34513 = -0.00181 \tag{9}$$

$$GainRatio(kslb) = \frac{Gain(kslb)}{E(kslb)} = \frac{-0.00181}{0.34538} = -0.00524 \tag{10}$$

The decision tree is constructed using test attributes such as subject category, grades, professional preferences, and admission methods. The information gain ratio of several attributes is calculated to determine their ranking from highest to lowest. The attribute with the highest information gain ratio is selected as the root node. Subsequently, the information entropy and information gain ratio of other attribute values are calculated in both the single move and unified move cases, and the attribute with the relatively highest information gain ratio is chosen as the next branch node. This process is repeated continuously to divide each branch, thereby obtaining a complete decision tree representing the decision of whether or not to report. Further elaboration will not be provided here.

5 Conclusion

According to the decision tree algorithm, the enrollment data was analyzed, resulting in a prediction accuracy of 88.76%. Among the factors influencing the registration situation, the order of importance is as follows: admission method, professional choice, place of origin, candidate category, candidate category, and gender. The decision tree algorithm is simple and easily understandable, allowing admissions staff who are not familiar with data mining to easily comprehend it due to its tree-like structure. The mining results provide certain assistance in enrollment management, and the prediction accuracy meets the requirements. In the future, additional data can be incorporated to enhance the decision tree algorithm and improve its application in the field of enrollment data mining.

Acknowledgments. 2020 Dalian Vocational and Technical College (Dalian Radio and Television University) scientific research project "Research on the Parallel Construction Path of Enrollment Expansion and Quality Improvement of Higher Vocational Colleges" (Project No.: DZDD2020B-06).

References

[1] Li Keqiang.2019 Government Work Report[EB/OL]. [2019- 03-05]. http://www.gov.cn/zhuanti/ 2019qglh/ 2019lhzfgz-bg/index.htm.

[2] Jiang Dayuan.On the great changes and new occupations brought by the expansion of higher vocational education to vocational education[J].China Vocational and Technical Education,2019(10):5-11.

[3] Pan Yan . Application of Enrollment Data Mining in Higher Vocational Colleges Based on CHAID Algorithm [J]. Yangtze River Information and Communication,2021,34(7):111-113.

[4] Bai X . Design and implementation of college enrollment data mining and visualization system based on decision tree algorithm [D]. Lanzhou:Lanzhou University, 2019

[5] Christopher W.Brown,Michael Jenkins.Analyzing proposals for improving authentication on the TLS-/SSL-protected Wed[J.] International Journal of Information Security,2016,15(6).

[6] Mysql A.MySQL:The world's most popular open source database[J] . 2010.