A Prediction Model for High School Students’ Academic Performance in College Based on Machine Learning

Xinyu Du
Business School, Sichuan University, Chengdu 610064, P.R. China

(Received October 06 2018, Accepted June 24 2019)

Abstract. In China, the college entrance examination (CEE) score is the only criterion to gain admission for most students. However, CEE score does not necessarily reflect the students future academic performance. The development of data mining and machine learning enable us to efficiently capture information from a wide range of data to support decision making. Therefore, this paper proposes a student academic performance prediction model by applying three classification classifiers: KNN, Naive C4.5, and Bayes on student data set of University A, which contains admission information and undergraduate GPA. The main aim of this model is to help the college to select a candidate who has potential in good academic performance. The experimental results show that the KNN algorithm is better than C4.5 and Naive Bayes. In addition to the CEE score, other attributes can affect predicting academic performance too, such as gender, high school awards, and admission area.

Keywords: machine learning, classification algorithm, academic performance, prediction

1 Introduction

Every college follows a strict admission system to select suitable students. Regardless of the admission system, the standardized test score is an essential evaluation indicator for applicants of most colleges. In the United States, students apply for the college after taking the Scholastic Assessment Test (SAT) or the American College Test (ACT). Besides, the school also evaluates other indicators of the student, such as the extracurricular activity performance, individual skill, and personality, to examine whether the student is suitable for studying at the school. In Japan and South Korea, there are usually two admission tests. After taking the national exam, students will also take reviews that are held independently by each college. On the other hand, many countries do not have a nationally unified college admission system. For example, in Australia, assessment exams are different in various states; in India, colleges will independently organize exams to enroll students.[5].

Similarly, in China, millions of students take the National College Entrance Examination (CEE) to get admission every year. In 2019, the number of applicants has exceeded 10 million. However, there is a massive gap between the quota of admission and the number of applicants, especially for top colleges, which leads to high pressure on Chinese high school students. Several top colleges have flexible admission policies: some applicants (usually the winners of National High-School Olympic Competitions) can be admitted directly by the college without examination. Other candidates, who passed the independent entrance assessment of college, can get admission with a lower CEE score. Generally speaking, the CEE score is a decisive factor in college admission for most students in China.

The standardized test is a common way for colleges to select students. However, we all know that although the standard test has the advantages of low cost and relatively fairness, the test score level cannot fully represent the student’s academic performance in the future. China’s current admission policy that places too much emphasis on CEE score may lead students to spend much energy on preparing for exams and neglect the development of other aspects.
The machine learning approach allows us to predict student’s college academic performance by mining more useful information from the student’s admission data. In this way, the school can select students who may have better academic performance.

The remainder of this paper is structured as follows: Section 2 reviews related works. Section 3 shows the methodology of classification using machine learning. Section 4 presents a case study of predicting college academic performance. Section 5 shows the results and discussion. Section 6 concludes the paper.

2 Related works

Many scholars have researched the prediction of college students’ academic performance. Betts & Morrell [4] analyzed the sample of 5000 students from 2 universities and found that both SAT scores and high school GPA were significant predictors of college GPA. Bai et al. [3] found that high school achievement and admission routes are also significant predictors of college grades by using the empirical model.

With the development of data mining and machine learning, more and more scholars have begun to use them to study the student’s academic performance prediction. Some researchers [2, 12, 18] focused on using different classification algorithms such as KNN, Naive Bayes, SVM to compare and analyze the final results. Some proposed new methods to improve prediction accuracy. Francis et al. [8] proposed a hybrid algorithm combining clustering and classification approaches to predict student’s academic performance with their behavior data. Livieris et al. utilized a semi-supervised learning approach to predict secondary school students’ performance [16]. Kaur et al. [14] applied a predictive data mining model using classification algorithms to identify students who are slow learners, which further makes school a decision to take more care for them. Masci et al. [17] collected student data from 9 countries (including Australia, Canada, France, Germany, Italy, Japan, Spain, UK, and the USA) to represent different types of the educational system. They applied the tree-based method to study the relevant characteristics from the student level and school level that affect student performance. The results show that social-economic index, anxiety, motivation, gender, and parental education are most influential in the student level; in the school level, the situation varies in different countries.

3 Methodology

This part mainly introduces the necessary steps and methods for establishing a predictive model based on machine learning.

3.1 Identify the business

Before conducting a machine learning project, we need to understand the actual problem of the study, identify the available data, and determine whether the mission is classification, regression, or clustering. In this study, our goal is to use students’ admission data to predict their academic performance at the university.

3.2 Data pre-processing

After collecting data from the source, we need to pre-process the data. Data pre-processing is considered to be a vital step in the knowledge discovery process so that we need to spend much time on it. The purpose of data pre-processing is to improve the prediction accuracy of the model by enhancing data quality.

3.2.1 Data cleaning

Some times there may exist some missing values or abnormal values (out of the normal range, called outliers) in the data we collect for several reasons, such as logging errors and personal privacy problem. These conditions may affect the final performance of the model, so we must deal with them. Methods such as deletion, replacement, and interpolation can be used to dispose of missing values. For abnormal values, we can adopt other techniques such as deletion and correction.

WJMS email for contribution: submit@wjms.org.uk
3.2.2 Discretization and normalization

Data discretization refers to dividing the value of continuous attributes into discrete intervals and representing each sub-interval with a specific number. Data discretization can simplify the data structure, improve learning efficiency, and make it easier to understand and explain the impact of attributes on the results. At present, data discretization methods are divided into two types: unsupervised one and supervised one [9]. Unsupervised discretization (e.g., binning) is challenging to obtain idealized results so that some researchers will use the supervised method such as entropy-based discretization.

Different attributes are sometimes measured in different units, which may affect the results of data analysis. Normalization is often required to eliminate this effect and resolve comparability issues between attributes. Commonly used methods are min-max normalization and z-score normalization.

3.2.3 Attribute selection and dimension reduction

The more attributes mean the more information, but too many attributes will increase the difficulty of model calculation. Not only that, redundant attributes will cause errors to reduce the model accuracy. What’s more, overly complicated models can lead to the over-fitting problem: model parameters are too suitable for training set data, but not work well on test sets. Therefore, we commonly use attribute selection and data dimension reduction to process the data, which can simplify the model while ensuring accuracy and reducing running time.

(1) Attribute selection: Select a subset of existing attributes by retaining only the most relevant to the mission. The attribute selection method using algorithms such as FOCUS [1] and Relief [15] is referred to filter approach by John et al. [13], which completely ignores the effects of selected attribute subset on the machine learning performance. For this reason, John et al. proposed a wrapper approach: train each candidate attribute subset on the training set, and select one which has a minimum error on the test set. However, the computational performance of a wrapper approach is worse than that of a filter approach. Here is also an embedded approach that integrates attribute selection algorithms into machine learning. The decision tree, such as ID3 [20] and C4.5 [19], is the typical one.

(2) Dimension reduction: The attribute vector for each data point is multidimensional, and data reduction can project the original data into a lower dimensional space [10]. We usually use PCA to realize dimension reduction, which forms several new indicators by constructing some linear combinations of original variables. At the same time, it can remove the correlation between attributes and make low-dimensional data retain the original information as much as possible. LDA is a method derived from Fisher’s work [7], which aims to make the data points easier to distinguish after dimension reduction. In addition to PCA and LDA, there are many non-linear methods suitable for linearly inseparable data, such as Isomap and LLE.

3.2.4 Model training

Our task belongs to supervised learning, in which the class tag has been artificially labeled to train a model. After dividing the data into a training set and a test set, the prediction model can be obtained based on the data and existing tags of the training set.

In this paper, three classification algorithms will be used to build separate models, including KNN, Naive Bayes, and C4.5.

(1) KNN: It is a simple but efficient algorithm. In a feature space, if the majority of the k nearest neighbors of a sample belongs to a specific category, then the sample also belongs to this category. For the KNN algorithm, different k values may bring different classification results, so it is essential to select the appropriate K value [11]. There are many ways to choose k value, and the easiest way is to run the algorithm with different k values and choose the best one.

(2) C4.5: It is an algorithm used to construct a decision tree. The decision tree is a tree data structure composed of decision nodes and leaves, where each internal node represents a test on an attribute, each branch represents a test output, and each leaf node represents a category [21]. A very complicated tree can cause poor
generalization ability. Therefore, we need to remove some nodes to simplify the model, which is called pruning. One benefit of C4.5 is that it can discretize continuous attributes in the tree construction process.

(3) Naive Bayes: It is a classification method based on Bayes’ theorem and conditional independence assumption. This method is relatively simple, which requires few parameters and is less sensitive to missing data. For discrete attributes, it takes the prior probability and the class conditional probability as parameters. For continuous attributes, there are generally two ways to deal with them: the first is to make discretization, but the size of interval division will affect the model quality; the second is to use the mean and variance of continuous attributes as parameters [6].

3.2.5 Model performance evaluation

After running the trained model on the test set, we need to evaluate its quality. In this research, three indicators will be used for the evaluation, including f-score, AUC, and accuracy.

F-score uses a weighted harmonic mean to deal with the possible contradiction between precision rate and recall rate. Recall rate represents the proportion of correctly classified cases in the total number of cases belong to a certain class. The precision rate represents the proportion of correctly classified cases in the total number of cases predicted into a particular category [22]. Accuracy is the ratio of the total number of properly predicted cases [8]. AUC is the area under the ROC curve. As a numerical value (generally between 0.1 and 1), it can intuitively evaluate the classifier quality.

4 Case study

The student data we can get is from the School of Economics and Management of University A, which is one of the top universities in China. The data contains student’s college grades, CEE score (including total score and subject score), and basic information such as gender, age, admission type, and admission area.

Most students in China are admitted to a university with the regular CEE score, so we focus our research on them. Those who take admission through other routes will not be considered by us, such as students who admitted without an exam. According to college GPA, we divide student’s academic performance into two categories, namely “good” and “poor”.

Thus, by sorting out the data, we get a data set. Table 1 shows all the attributes it contains. We will use this data set to predict the student’s academic performance and compare the classification quality of different classifiers. At the same time, we will examine the prediction results which using only CEE score.

In this article, most of the operations will be done in WEKA (an open source tool).

5 Results

5.1 Selected attributes

The categories in this data set are unbalanced, that is, instances marked as “poor” only account for about 14.3% of all instances. In this case, the performance of classifier will be affected. Even if the model classifies all instances into a majority category, it can achieve excellent classification accuracy.

We can solve this problem by re-sampling. In this experiment, we use the SMOTE approach to increase the number of instances of a minority category. After re-sampling, we use ReliefF approach to select attributes. Fig. 1 displays the attribute importance.

The results show that gender, CEE total score, CEE subject score, and admission area are more important compared to high school awards and social work, while student type, minority and enroll age are least important. We select top 9 attributes which are shown in Table 2.

To make a comparison, we also make predictions with only CEE score. In this situation, attribute selection is no longer performed because the number of attributes is already small.
Table 1: Attributes and description

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admission area</td>
<td>The area where student comes from</td>
</tr>
<tr>
<td>Gender</td>
<td>Student’s gender</td>
</tr>
<tr>
<td>Enroll age</td>
<td>Age when entering college</td>
</tr>
<tr>
<td>Minority</td>
<td>Whether the student is an ethnic minority or not</td>
</tr>
<tr>
<td>Student type_urban</td>
<td>Whether the student comes from urban area or not</td>
</tr>
<tr>
<td>Student type_first</td>
<td>Whether the student takes the college entrance examination for the first time or not</td>
</tr>
<tr>
<td>CEE_type_liberal</td>
<td>Whether the student is a liberal arts student or not</td>
</tr>
<tr>
<td>Award_outstanding</td>
<td>Whether the student ever won the outstanding student title in high school or not</td>
</tr>
<tr>
<td>Award_outstanding_top</td>
<td>Whether the outstanding student title is province-level or not</td>
</tr>
<tr>
<td>Award_competition</td>
<td>Whether the student ever won the competition in high school or not</td>
</tr>
<tr>
<td>Award_competition_top</td>
<td>Whether the competition award is national-level or not</td>
</tr>
<tr>
<td>Social work</td>
<td>Whether the student ever had social work experience or not</td>
</tr>
<tr>
<td>CEE total score</td>
<td>Total score of the college entrance examination</td>
</tr>
<tr>
<td>CEE score_eng</td>
<td>English subject score of the college entrance examination</td>
</tr>
<tr>
<td>CEE score_math</td>
<td>Mathematics subject score of the college entrance examination</td>
</tr>
<tr>
<td>CEE score_chin</td>
<td>Chinese subject score of the college entrance examination</td>
</tr>
<tr>
<td>Performance</td>
<td>The student’s college academic performance</td>
</tr>
</tbody>
</table>

![Attribute Weight](image)

Fig. 1: Attribute weight

5.2 Classifier performance

We use the KNN, C4.5, and Naive Bayes classifiers for training. 80% of the data set is used as the training set, and the rest is used as the test set. Table3 and Fig. 2 display the experiment results.

By comparison, we can find that although the accuracy of the KNN classifier is slightly lower than C4.5, it has a higher F-score and AUC value. Overall, the KNN classifier works better than others. Also, we can find that the model performance using only CEE score is inferior to that using CEE score + extra attributes.
Table 2: Selected attributes

<table>
<thead>
<tr>
<th>Rank</th>
<th>Selected Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gender</td>
</tr>
<tr>
<td>2</td>
<td>CEE total score</td>
</tr>
<tr>
<td>3</td>
<td>CEE score_eng</td>
</tr>
<tr>
<td>4</td>
<td>CEE score_chin</td>
</tr>
<tr>
<td>5</td>
<td>Admission area</td>
</tr>
<tr>
<td>6</td>
<td>CEE score_math</td>
</tr>
<tr>
<td>7</td>
<td>Award.outstanding</td>
</tr>
<tr>
<td>8</td>
<td>Award.competition</td>
</tr>
<tr>
<td>9</td>
<td>Social work</td>
</tr>
</tbody>
</table>

Table 3: Classifier performance

<table>
<thead>
<tr>
<th></th>
<th>CEE Score + Extra Attributes</th>
<th>Only CEE Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>F-score</td>
</tr>
<tr>
<td>KNN</td>
<td>0.798</td>
<td>0.796</td>
</tr>
<tr>
<td>C4.5</td>
<td>0.803</td>
<td>0.789</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.779</td>
<td>0.771</td>
</tr>
</tbody>
</table>

Fig. 2: Classifier performance

6 Conclusion

In this article, we built a model for predicting student’s academic performance in college based on the classification algorithm using the student’s admission information. In particular, we compared the performance of three classifiers (KNN, C4.5, and Naive Bayes) using different attribute subsets. According to the results, KNN is a better classifier. Besides, in addition to the CEE score (including CEE total score and CEE subject score), gender, high school awards, and social work are also important factors predicting academic performance. The prediction performance using CEE score and extra attributes is superior to that using only CEE score.

This model can alleviate the limitation caused by the admission system that only concerns about the CEE score. In this way, schools can choose applicants who potentially to have better academic performance. The main weakness of this study is that the attributes used to construct the predictive model are not necessarily optimal due to the limitation of data collection. There are still some critical features that need to be captured and used in the model.

Therefore, in future research, more attributes that may affect a student’s academic performance will be applied for better prediction accuracy, such as high school grades, classroom performance, and psychological factors.

WJMS email for contribution: submit@wjms.org.uk
References