An Optimal Investment Strategy for University Endowment Fund

Yu-xuan XIONG\(^1\), Xiao-hong DENG\(^2\), Xin LIU\(^1\) and Ya-nan JIANG\(^1\)

\(^1\)Software College, Jilin University, China
\(^2\)Department of Psychology, Faculty of Education, Hubei University, China
*Corresponding author

**Keywords:** University endowment fund, Data mining, Portfolio optimization, Support vector machine.

**Abstract.** Automatic educational endowment funding systems suffer from two problems, i.e., lack of historical data and labels. In this paper, we propose a simple but effective investment portfolio model, which helps us to determine which campuses should be invested in, how much money should be distributed, and how long funds should last. Our algorithm first uses the PCA framework to assign each campus a label, and then uses SVMs to learn ranking list of campuses. We further develop a greedy strategy to optimize portfolio. Experiments on the ROC metric indicate the effectiveness of our model.

**Introduction**

With the explosive growth of modern universities, the high educational expenditure becomes more and more limited. More recently, the university endowment fund is major additional support to universities, and many institutions will benefit from the funds. How to raise and manage these funds is very important [1]. However, university endowment funds are mainly determined by manual rules and human experience.

Traditional methods usually analyze return of investment (ROI) directly. Means-variance model of Markowitz is the most typical model in the field of portfolio [2] and has many limitations: 1) Means and variance can describe how the income deviates from mathematical expectation, but cannot give out the direction of such a deviation. However, people always care about the loss of an investment action. 2) Since these methods use means and variance of historical data of ROI, more information about campuses cannot be taken into account. But some behaviors including academic reputation and students’ performance should be considered [3]. Merton used campuses’ performance such as education, training, research, and storage of knowledge to establish a model deciding which campuses deserved to be invested in [4]. Unfortunately, this model only took the quantity of activities into account and could not evaluate a campus in the round.

Data mining approaches including support vector machine (SVM) have been applied to predict the closing price of stock [5] and gained good result in return. But university endowment problem is different from stock investment because people can collect the data of closing price of stocks everyday while the data of campuses’ endowment are usually obtained once a financial year. Because data mining can be used to effectively explore the inner relationship of variables, so we wanted to build an investment portfolio model of university endowment funds using data mining.

When we optimized the investment strategy, it was a challenge to search for most effective variables. Except for money describing benefits from the investment, we selected academic performance, repayment, student, earning, and cost to evaluate campuses. Because the database provided by Integrated Postsecondary Education Data System (IPEDS) was enormous, principal component analysis (PCA) was used to reduce dimension of original data. We created a new column labeling the campuses because of no apparent labels classifying each campus. Since the condition of campuses will remain stable within a short time, it was acceptable that we evaluated the campuses with a constant way in the next 5 years. As a university consists of several campuses and each campus has relatively separated management system, these campuses could be evaluated separately.
Method

Data Resource

The most recent cohorts’ data was derived from IPEDS where there are a lot of information about colleges, universities, and vocational technical schools supported by federal government and foundations. The information includes utility costs, registration, academic behavior, etc.

The data-set we used is a table including information about 7804 campuses. There are 79 features settled down in the table.

Description of the Problem

Because foundations intended to donate 100 million US dollars per year in candidate campuses in the next 5 years, we need provide a ranking list of the top candidate campuses and make a strategy about how to donate to the selected campuses and how long the funds should be provided to have the highest likelihood of producing a strong positive effect on campus performance.

Data Preprocessing

As we evaluated students’ performances based on repayment, earning, academic performance, etc., a 7804*77 matrix was extracted from the original data-set manually. The 77 columns were related to students’ performances. Before training SVM, we selected some appropriate columns and filled the null blanks in these columns using interpolation. Normalization was used to compress data into a common section between -1 and 1 because there were more than 120 columns in original data-set and the range of variation was quite different.

PCA was used to reduce dimensionality for some reasons: 1) It might take a lot of time to train the SVM as the columns in 7804*77 matrix were enormous. 2) Some columns selected manually needed to be abandoned as they were not significant. 3) PCA could produce some meaningful intermediate products describing the weight of each new linear combination. 4) The intermediate products could be used to calculate the label of each row (see Table 1).

Table 1. The cumulative score of principal components.

<table>
<thead>
<tr>
<th>component</th>
<th>contribution</th>
<th>cumulative contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5901</td>
<td>0.5901</td>
</tr>
<tr>
<td>2</td>
<td>0.1757</td>
<td>0.7658</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>15</td>
<td>0.0016</td>
<td>0.9988</td>
</tr>
<tr>
<td>16</td>
<td>0.0012</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Since the float in 15 row was 0.9988 and the float in next row was 1.0000, first 15 rows in score matrix were used to represent the original matrix. The weighted average of each row was stored in the 16th column.

Classification by SVM

To classify the campuses, labels describing whether the campuses deserve to be invested in were given manually. The average of the 16th column of new data matrix was calculated, and the number in the 16th column was compared with the average. If the number in 16th column was greater than the average, label “class 1” was assigned to this row and was stored in column 17, or the column was labeled “class 2”. Class1 and class 2 represented the campuses deserving and not deserving to be invested in respectively.

Libsvm in Matlab was used when we classified the campuses. It was our innovative point that we used grid search method to optimize the process of searching for an appropriate parameter set for SVM. 7/10 of the data-set was used as train set and other 3/10 as test set.
In order to select the best kernel function for SVM, four types of kernel functions (linear, polynomial, radial basis function, and sigmoid) in libsvm have been tested on train set and test set. The results of SVM classifier using linear, polynomial, radial basis function, and sigmoid as kernel functions are 98%, 87%, 99%, and 96% respectively. So we selected radial basis function as kernel function.

The result for parameter selection of SVM was showed in Figure 1. The trained SVM model was tested by 2341 samples in test set and the accuracy of test set was 98.334%.

![Figure 1. The result of parameter selection.](image)

The SVM was reliable because the area between the receiver operating characteristic curve (ROC) of SVM and the diagonal was quite close to 0.5 (see Figure 2).

![Figure 2. The ROC curve.](image)

**Rank of Candidate Campuses**

Candidate campuses were ranked based on the intermediate products named dec_values. The score of classification was between -1 and 1 and had linear relationship with dec_values. The larger the dec_values, the more reliable the result of classification. The larger dec_values of a campus was, the more worthy the campus was to be invested in. The partial results of ranking were showed and ID was the number of campuses in original data-set (see Table 2).

<table>
<thead>
<tr>
<th>ranking</th>
<th>Dec_value</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00E-07</td>
<td>1938</td>
</tr>
<tr>
<td>2</td>
<td>1.00E-07</td>
<td>3747</td>
</tr>
<tr>
<td>3</td>
<td>1.00E-07</td>
<td>4438</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>6</td>
<td>1.08E-07</td>
<td>2032</td>
</tr>
</tbody>
</table>

**Portfolio Optimization**

After generating the ranking list of candidate campuses, a greedy strategy was used to establish a
portfolio model. Foundations were going to distribute the total 500 million dollars based on the list order. One assumption was that foundations could obtain more benefits after investing more money in better campuses. So the top rank campuses will be invested in first.

The formula (1) was used to calculate the proportion of the selected campuses:

\[
pro_i = \frac{n + i - 1}{\sum_{i=1}^{n} i},
\]

In the above formula, \(i\) represented the campus ranking. If a campus was the best in the list, \(i\) equaled to 1. \(n\) represented the number of selected campuses. \(Pro\) represented the proportion a campus could be distributed in total 500 million dollars.

The duration of donation for a campus was calculated with the formula (2):

\[
\text{range} = \left[5 \times \sum_{i=1}^{n} pro_i, 5 \times \sum_{i=1}^{n} pro_i\right]
\]

Here we took \(n=2\) as example. Two top rank campuses were 1938 and 3747. They were given 1000/3 and 500/3 million dollars respectively. The duration of donation for 1938 was the first to 10/3 th year and for 3747 was the 11/3th to 5th year.

Foundations can stop investment whenever the campus invested in is not worth to be distributed any more.

**Discussion**

To improve the accuracy of the model, the further research and refinement are needed: 1) The portfolio optimization strategy is sometimes risky as it can’t share risk effectively. A more reliable strategy should be provided. 2) The ranking list would be better if we used a training score of each campus rather than dec_values. 3) Because different campuses at a university probably influence each other, a method analyzing the inner relationships between these campuses should be explored and a mathematical model should be built to modify the score of each campus. 4) The accuracy of the model would be better if we got more historical data.

**Conclusion**

A simple but effective investment portfolio model is established, which helps us to determine which campuses should be invested in, how much money should be distributed, and how long funds should last. Our algorithm first uses the PCA framework to assign each campus a label, and then uses SVMs to learn ranking list of campuses. We further develop a greedy strategy to optimize portfolio. Experiments on the ROC metric indicate the effectiveness of our model.

**References**


