Admission Probability Prediction of College Entrance Examination Based on Siamese Network

Sida Gao
Shenyang Institute of Computing Technology, Chinese Academy of Sciences; University of Chinese Academy of Sciences, Beijing, China
gaosida20@mails.ucas.ac.cn

Tianming Liu
Shenyang Institute of Computing Technology, Chinese Academy of Sciences; University of Chinese Academy of Sciences, Beijing, China
liutianming21@mails.ucas.ac.cn

Lijun Fu*
Shenyang Institute of Computing Technology, Chinese Academy of Sciences; University of Chinese Academy of Sciences; Laboratory of Big Data and Artificial Intelligence Technology, Shandong University, Beijing, China
fu_lijun@ucas.ac.cn

Xueliang Chen
Shenyang Institute of Computing Technology, Chinese Academy of Sciences; University of Chinese Academy of Sciences, Beijing, China
chenxueliang20@mails.ucas.ac.cn

ABSTRACT

For students taking the College Entrance Examination, the choice of major is as important as their scores. The choice is often related to whether the candidate can be admitted with a higher probability. Therefore, how to accurately predict the admission probability has become the important problem. In this paper, by using the candidate's provincial rank in the current year and the provincial enrollment rank of each major in the past three years, Siamese network structure combined with gated recurrent unit is used to predict the admission probability of the candidate in each major. It was found that the model can accurately calculate the admission probability, and can capture the time series information of the rank change in previous years. Compared with the traditional average method, it has a great improvement.

CCS CONCEPTS

• **Computing methodologies** → Artificial intelligence; Knowledge representation and reasoning; Probabilistic reasoning.

KEYWORDS

siamese network, gated recurrent unit, similarity calculation, admission probability prediction

ACM Reference Format:
Sida Gao, Tianming Liu, Lijun Fu, and Xueliang Chen. 2022. Admission Probability Prediction of College Entrance Examination Based on Siamese Network. In *2022 6th International Conference on Electronic Information Technology and Computer Engineering (EITCE 2022), October 21–23, 2022, Xiamen, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3573428.3573488

*Corresponding author.

EITCE 2022, October 21–23, 2022, Xiamen, China

ACM ISBN 978-1-4503-9714-8/22/10…$15.00
https://doi.org/10.1145/3573428.3573488

1 INTRODUCTION

At present, there are relatively few relevant studies for high school admissions. Relevant methods mainly include line difference method, average ranking method, combined prediction model, neural network, etc.

The line difference method is a method that uses the difference between the current year's provincial control line and the previous year's admissions score and the provincial control line (line difference) as a way to predict admissions scores [1]. This method is easy to calculate, but the accuracy of the prediction is low. In particular, the further away from the minimum score line, the less accurate the prediction. The average ranking method is a modified equivalence method, which uses the average of the lowest ranking scores for many years to predict the admissions scores in the nth year. It also suffers from low prediction accuracy and is affected by changes in the total capacity of candidates and the minimum admissions line [2]. The combined prediction model is a model composed by various different single prediction models. Zhou Fan combines three different single prediction models together. The coefficients of the required optimal variable weights are derived through a least squares operation [3]. Finally, the commonly used error sum of squares is then used for comparison and validation. The experiments with Chongqing admission scores yielded a high prediction accuracy of the combined model. However, this model did not address factors such as the fact that the college entrance examination score line was affected by the difficulty of the questions [4]. Li Jingwen et al. used a combination of fuzzy mathematics, which can illustrate indistinguishable phenomena with mathematical thinking, and a grey prediction model, which requires less observational data, to construct a fuzzy grey model. This model takes into account factors such as the minimum score line, the number of students enrolled in the programme [5], and the students' preference for the school. The data from Lanzhou University and Lanzhou Jiaotong University were analysed and tested. The prediction accuracy of this model was found to be relatively good. However, the experimental data selected for this model are relatively homogeneous. It cannot better illustrate the accuracy of the model prediction. Jani used a neural network model. The three-year data was processed. The normalisation of the influencing factors was achieved by dimensionality reduction through principal component analysis. The best weight value is finally calculated by BP back propagation algorithm [6]. The prediction accuracy is improved. This method still suffers from the problem of small test sample set, which is not representative.

By conducting regression analysis on the data, Xunping et al. used the least squares method to confirm that there was some correlation between the institution's admissions score line and the local score line [7]. However, this model was only applicable to specific colleges and universities with years of historical data and was not universally adopted. Jia Zhiguo used a neural network algorithm to analyze the college admissions data. The algorithm was found to be highly accurate under non-linear requirements. The prediction of admission rank by the above model gives more ideas for solving this problem [8]. However, there are still the following problems: the prediction of admission rank is somewhat biased and the effect varies widely between samples.

The prediction of admission rank by the above model provides more solutions to this problem, but there are still the following problems: there is a certain deviation in the prediction of admission rank, and the effect of each sample is quite different. Only the prediction of admission ranking can not intuitively reflect the problem that users are more concerned about and how large the admission probability is. To improve the above problems, this paper proposes an admission probability prediction algorithm [9], which uses the Siamese network model and the long short-term memory algorithm to make full use of the time series information of recent admission data, and has achieved satisfactory results in experiments and applications.

2 MATERIAL AND METHODS

To intuitively and accurately calculate the admission probability of each institution, the main contributions of this paper are as follows.

We organized the construction of a batch of admission probability sample data sets. Determined the correlation between the rank information and the admission probability over the years under the guidance of experts in the field of education, and organized the staff in the field of education to label.

Aiming at the prediction task of admission probability, this paper designed a prediction method of admission probability based on the Siamese network. By modeling the association of admission rank. Current rank and admission probability in recent years. The model was used to fit, and the admission probability concerned by the user was calculated. The experiments on the admission probability sample data set prove the effectiveness and accuracy of the model.

2.1 Construction of Admission Probability Sample Dataset

Dataset annotation refers to the standard formulated by experts in the field, and the labeled data set is defined as a binary classification task. The samples of each major including the rank of the past year and the rank of the student are labeled, and the label is whether they can be admitted. The data set is stored in JSON, and the specific form and content are shown in Figure 1


```
{
    "school_name": "Pku",
    "school_id": "A001",
    "major_name": "Law",
    "rank1": 86,
    "rank2": 77,
    "rank3": 90,
    "update_score": 691,
    "update_rank": 139,
    "student_score": 690,
    "student_rank": 166,
    "resign_label": 0
},
{
    "school_name": "Pku",
    "school_id": "A001",
    "major_name": "Law",
    "rank1": 86,
    "rank2": 77,
    "rank3": 90,
    "update_score": 691,
    "update_rank": 139,
    "student_score": 691,
    "student_rank": 139,
    "resign_label": 1
},
```


Figure 1: Example of annotated data

The meanings of each field are: school_name is the sample's university name, school_id is the sample's university code, major_name is the sample's major name, Rank1 admission position in the previous year, Rank2 admission position in the first two years, Rank3 admission position in the first three years, update_score is admission score, Update_rank is admission position, student_score is sample's candidate score, student_rank is sample's candidate position, resign_label whether the candidate can be admitted.

The data set is annotated by people in the field of education, and the size of the data is 150,000, and each data is annotated by more than three annotators. It has high annotation accuracy and strong data reliability, which can provide great support for model training.

2.2 Construction of Admission Probability Prediction Model

Data processing layer. When analyzing the data, we found that two factors in the data, major ranking, and undergraduate admission ranking, had a great impact on the results, so we should focus on these two groups of data. In the data processing layer, mainly to professional ranking and admission to the undergraduate rankings for processing, the subtraction, and division, the interaction characteristics between two groups of data, through this step, the two groups of data contain information fusion, after processing the data of representative admitted ranking cross relations with undergraduate admissions ranking, said similar to normalized calculation [10]. The network structure designed in this paper is shown in Figure 2

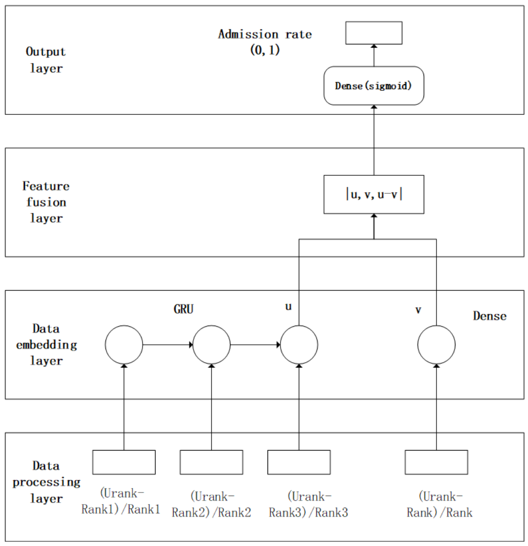

Figure 2: Plot of admission probability prediction model based on Siamese network

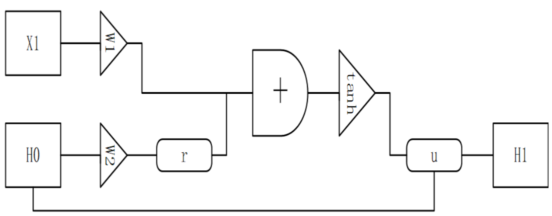

Figure 3: GRU network structure

After the data has been processed. A new data normalised to suit the operation is obtained. Different forms of coding are adopted in the twin network due to the different data formats on both sides. The GRU network was used to encode the data for the combined three-year professional ranking and undergraduate admissions ranking data[11].Will step on the three years of data obtained using one-way GRU helped network coding, encoded is expressed as $N_3=[n_1, n_2, n_3]$, $H_3=[h_1, h_2, h_3]$ represents the ith of digital after GRU helped network coding vector of the hidden layer, to obtain the final unit vector of the hidden layer. Because it involves the whole sequence information, as general data of the three years as u. The GRU network structure is shown in Figure 3

The candidate data s in the current year is encoded through the Dense layer, so that its dimension is consistent with the hidden layer u, and v is used to represent the encoded vector. The vector representation formula is shown in equation 1) and equation 2):

$$\mathrm{u} = H_3 = \mathrm{GRU}\,(N_3) = [h_1, h_2, h_3] \tag{1}$$

$$\mathrm{v} = \mathrm{Dense}\,(\mathrm{s}) \tag{2}$$

Feature fusion layer. For the encoded vectors u, and v, to calculate and represent the similarity between them, the model is completed by concatenation and subtraction. In the feature fusion layer, the three vectors u, v, and u-v are concatenated to represent the relationship between the two vectors. Experimental results show that compared with the method of multiplication, the fusion method can provide more information about the relationship between two vectors for the model.

Output Layer. Feature fusion, namely after fusion of vector | u, v, u - v |, in the output layer, the Dense layer on vector calculation, using a sigmoid function as the activation function, get the final results, calculating said such as equation 3):

$$\text{Admission rate} = \mathrm{sigmoid}\,(\mathrm{Dense}\,(|\mathrm{u}, \mathrm{v}, \mathrm{u} - \mathrm{v}|)) \tag{3}$$

The GRU has two gates, namely a reset gate and an update gate. Intuitively, the reset gate determines how to combine the new input information with the previous memory. And the update gate defines the amount of the previous memory saved until the current time step. The update gate helps the model decide how much information from the past should be transferred to the future or how much information from the previous time step and the current time step should be continued [12]. This is very powerful because the model can decide to copy all the information from the past to reduce the risk of the gradient vanishing. The reset gate mainly determines how much past information needs to be forgotten [13]. This is commonly used when dealing with sequence information.

Table 1: information of the data set

Data Set Information	Amount
Total Amount	150000
Scores Range	200
Amount of Colleges	300
Amount of Specialties	80

3 RESULTS AND DISCUSSION

3.1 Dataset

This article for nearly three years in Shandong province enrollment plan and recruit students number information, mainly covers the admission, admission scores, undergraduate admission information such as rankings, data cleaning, and organize a group of staff education field to indicate. For different ranking examinees. Different ranks of colleges and universities. Mark whether students can be admitted to universities. The annotation data quantity is about 150000. It involves a score interval of 200 points, 300 reference colleges, and 60 majors, and the data as a whole has a strong representation. The information of the data set is shown in Table 1.

3.2 Experimental and results

The training data and validation data used in this experiment are divided according to the ratio of 9:1. The GRU and Dense coding dimensions are 100, the parameter optimization method dimension is Adam, and the Loss function is mean-squared Loss.

Taking a student with a ranking of 50,000 as an example, the model is verified to predict the probability of the student being admitted by each institution, and the overall distribution is mapped into a graph. The probability of the student being admitted by each

Table 2: The probability admitted to each institution

College Name	Major Name	Three-year Average Rank	RankOne Year	RankTwo Year	RankThree Year	Admission-p
TUST	MSR	49974	48481	57867	59670	78.12%
KUST	FIN	49961	49767	50321	49767	78.53%
SUES	MTA	50015	43440	56992	55789	74.62%
......	......	......	......	......	......	......
HUT	FST	50005	51602	48229	48770	81.33%
NEPU	EDU	50045	52338	46575	50106	82.36%

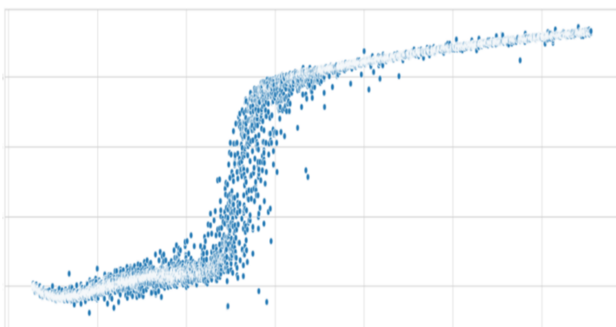

Figure 4: Probability distribution diagram of overall student admission

institution is shown in Table 2. The overall probability distribution of admission is shown in Figure 3

Figure 4 shows the probability of the candidate ranked 50,000 being admitted to the major of each institution. The horizontal axis is the three-year average ranking of the institution. The vertical axis is the probability of the candidate is admitted.

As you can see, this distribution is more suitable for real-world scenarios and can handle different years. We assume that the three years have a weighted rank of 54,000, one of the years may have a higher rank. Such as 45,000. Success for a weighted rank of around 54000 will fall in the range of 0.4 to 0.8, so the model will give you a combined result after considering three years of rank fluctuations. If it is three years in the rank of about 54000, the final weighted 54000. Then this major can be stable, the success rate will be high. If there are fluctuations in the three years, such as more than 40,000 a year, more than 50,000 a year, more than 60,000 a year, and finally weighted to get 54,000. This kind of professional should be carefully registered. The success rate will be low.

4 CONCLUSIONS

From the experimental results. We can find that the model proposed in this paper can accurately fit the admission probability of each major in the prediction effect. And can cope with the special changes of major rankings and make predictions. During the prediction, the data of the past three years and the ranking information of the students themselves were fused. The Siamese network was coded respectively to map the data to a higher dimensional space for comparison, and a large number of sample data were calculated to draw the probability distribution map. Through the distribution representation of the overall probability, it can describe the difference in the admission probability of each major. Express the trend of the admission probability of each major with the change of the ranking, which is reasonable and interpretable. And can better solve the problem of admission probability prediction.

ACKNOWLEDGMENTS

The authors wish to thank University of Chinese Academy of Sciences for providing equipment and technical support for this experiment. Thanks to the help of the students, the experiment was completed.And thank my supervisor who gave valuable suggestions that have helped to improve the quality of the manuscript.

REFERENCES

[1] Dey, R., & Salem, F. M. 2017, August. Gate-variants of gated recurrent unit (GRU) neural networks. In 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS) (pp. 1597-1600). IEEE.
[2] Graves, A. 2012. Long short-term memory. Supervised sequence labelling with recurrent neural networks, 37-45.
[3] Guo, Q., Feng, W., Zhou, C., Huang, R., Wan, L., & Wang, S. 2017. Learning dynamic siamese network for visual object tracking. In Proceedings of the IEEE international conference on computer vision (pp. 1763-1771).
[4] Shao, H., Zhong, D., Du, X., Du, S., & Veldhuis, R. N. 2021. Few-shot learning for palmprint recognition via meta-Siamese network. IEEE transactions on instrumentation and measurement, 70, 1-12.
[5] Krotov, D., & Hopfield, J. J. 2019. Unsupervised learning by competing hidden units. Proceedings of the National Academy of Sciences, 116(16), 7723-7731.
[6] Song, L., Gong, D., Li, Z., Liu, C., & Liu, W. 2019. Occlusion robust face recognition based on mask learning with pairwise differential siamese network. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 773-782).
[7] Kim, M., Alletto, S., & Rigazio, L. 2016. Similarity mapping with enhanced siamese network for multi-object tracking. arXiv preprint arXiv:1609.09156.
[8] Becker, S. 1991. Unsupervised learning procedures for neural networks. International Journal of Neural Systems, 2(01n02), 17-33.
[9] Fu, R., Zhang, Z., & Li, L. 2016, November. Using LSTM and GRU neural network methods for traffic flow prediction. In 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC) (pp. 324-328). IEEE.
[10] Batet, M., & Sánchez, D. 2015. A review on semantic similarity. Encyclopedia of Information Science and Technology, Third Edition, 7575-7583.
[11] Gers, F. A., Schmidhuber, J., & Cummins, F. 2000. Learning to forget: Continual prediction with LSTM. Neural computation, 12(10), 2451-2471.
[12] Sunilkumar, P., & Shaji, A. P. 2019, December. A survey on semantic similarity. In 2019 International Conference on Advances in Computing, Communication and Control (ICAC3) (pp. 1-8). IEEE.
[13] Huang, Z., Xu, W., & Yu, K. 2015. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.