Shanghai Jiao Tong University

IE304

Mining of Academic Mentorship

Author:
Kaibin Zheng

Student Number:
515021910366

May 27,2018

Contents

1 INTRODUCTION

In this project, we predict the mentorship from the information of the papers and their authors, which is obtained from the database of "2015-2018 Acemap, Inc. Shanghai Jiao Tong University". Based on the work of Wenzheng Tao, we improve the mentorship prediction accuracy and expand its function of prediction, which enables us to predict the relationship of advisor-student, student-advisor and cooperation. Here the cooperation relationship is the relationship between ordinary partners other than advisor-student or student-advisor relationship. We improve the performance of the prediction from three aspects, i.e. feature extraction, feature processing and deep neural network model. We evaluate the performance of our prediction using the data set from the database of Acemap. Results show that the proposed prediction method provides higher prediction accuracy than existing policies.

First of all, we have to explain the reason for choosing the topic Mining of academic mentorship instead of the original chosen topic Personalized User Profiling for recommendations. Actually, we are supposed to extract academic profile in the early stage. But soon we realized that a large number of websites, for example, Acemap, are doing the same work, and it's not wise to repeat it on our own. However, there's no website, even Acemap, perform excellently on extraction of academic mentorship. Then we consider to help to improve the accuracy of mentorship extraction.

Our contributions are listed below:
(1) Extract cooperation relationship from Acemap data set and expand the data set;
(2) Extract features by considering the situation in real world that last author in the paper's author list is always the advisor;
(3) Extract features by considering relative position of the author in the paper's author list;
(4) Remove the unnecessary features;
(5) Optimize the implementation of the code;
(6) Improve the predict speed and lower the space complexity;
(7) Expand the function by three-classification model;
(8) Optimize the neural network model.

2 MODEL

In order to distinguish the relationship, we model it as a three classification problem and train a classifier based on features that reflect relationship of authors. The model is shown as Figure 1. Most of the features reflect the dominant differences,

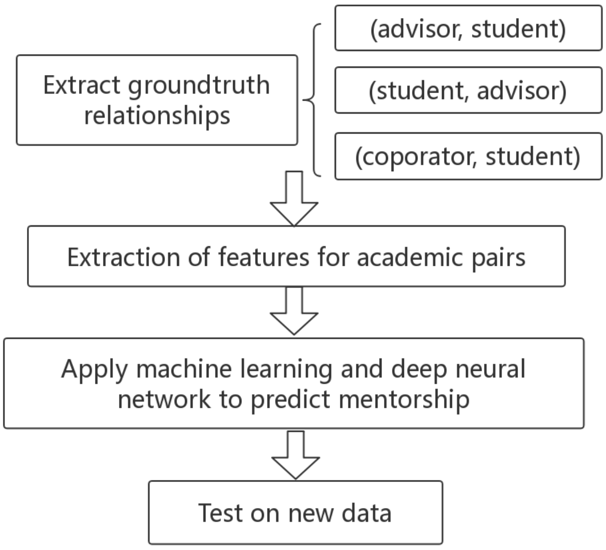
(advisor, student)
(student, advisor)
(coporator, student)
Extract groundtruth relationships
Extraction of features for academic pairs
Apply machine learning and deep neural network to predict mentorship
Test on new data

Figure 1: Model

cumulative differences, and specific differences between two authors in the cooperation relationship. The dominant difference is in anti-correlation with the author's order. In general, the more frontier it is, the more dominant he is in this paper. The cumulative difference is reflected in the number and duration of the papers published by the author before the cooperation relationship begins. In general, the advisor has published papers earlier than the student and has much more published papers than the student. The specificity of the difference can be obtained from the papers published during the cooperation period, the proportion of papers containing the other author reflects that the students' papers mostly have the names of the advisor, whereas the teachers' papers do not necessarily have the student's names. The students' specificity is strong, and the teacher's specificity is weak. Besides these features, in order to consider the condition that the advisor is always the last one in the author list of papers, we add the feature that describes whether the advisor or the student is the last author in their co-paper. If neither of them is the last one in the paper's author list, then we consider their relative position in the author list, for example, where the student is the first author and the advisor is the second author.

In the neural network training processing, we first build the network using Tensorflow model and Keras API. We use multi-layer network in Figure 2 with densely-connected layer with the activation function "tanh" and the fully-connected network as the output layer. In order to avoid over-fitting, we apply Dropout layer to the input. Dropout layer consists in randomly setting a fraction rate of input units to 0 at each update during training time, which helps prevent over-fitting.

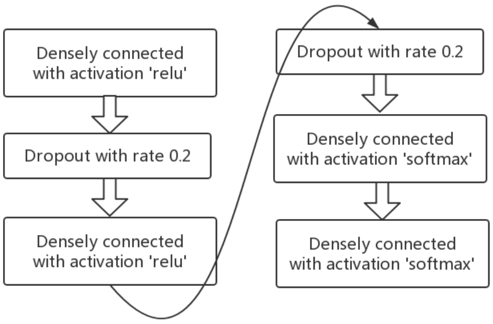
Densely connected with activation 'relu'
Dropout with rate 0.2
Densely connected with activation 'relu'
Dropout with rate 0.2
Densely connected with activation 'softmax'
Densely connected with activation 'softmax'

Figure 2: Neural Network

3 EXPERIMENT

In the project, we have two python program. The first one is the feature extraction and feature processing program, and the other one is the neutral network training and prediction program.

In order to make three-classification prediction possible, we have to expand the training data set first. Therefore, we extract cooperation relationship from the Acemap data set by removing mentorship in the paper's author list.

After we obtain the author id from the training set, we can extract features from the Acemap database online. The relative difficult part is to extract features of the last author. First, we need to search the paper list of the advisor and student in the training set, and then find the coincident part. After that, for each paper we have found, we search the author's position in the author list and find the largest one, which can be later used to compare with the advisor and the student's position in the author list and be extracted as features. We remove the feature linear calculation in the reference code because the linear calculation between the raw feature data is unnecessary, which could be learned by the neural network easily. Doing these could not improve the mutual information among the features and may lead to over-fitting in the neural network training.

After feature extraction, we use the multi-layer neural network and deep learning method to build our network. The output layer of the network is set to be the fully connected layer with output dimension being equal to the number of classification, i.e.two for classification of mentorship and three for classification of mentorship and cooperation. We divide the data set with labels and features into training set and testing set. After debugging the parameters of the neural network, we have found the best parameters and then train the neural network. Here, we set the number of neurons in the fully connected layer to be 30. From the experiment, we find that the number of neurons in the input layer should be large enough, which is significant for the neural network to transfer information to the later layer. Besides, the more deeper network with dense layer has better performance, which would be show later. After that, we calculate the accuracy and loss in the testing set and obtain the result. Later, we predict the relationship in new pair of authors.

In previous work of predicting mentorship, they set a large author-paper dictionary containing the key as the author and value as all of the papers published by this author. Similarly, they set a paper-author dictionary. These dictionaries would be convenient for them to search for the information when predicting the relationship after training the neural network. However, this method increases the space complexity and need a lot of memory space. In order to expand the usability of the code and lower the complexity, we search information from the database online and obtain what we need to extract the features, and then use the trained neural network to do prediction. We believe that after this improvement, the prediction model will work well in predicting mentorship and cooperation with higher speed and lower memory space.

4 RESULT

Table 1: Prediction Result

classification	**training set**	**cooperation**	**batch size**	**epoch**	**accuracy**	**loss**
2	110000	no	5	30	98.32%	0.052
2	490000	yes	500	100	98.79%	0.04697
3	160000	yes	100	50	96.01%	0.1240
3	490000	yes	500	100	96.67%	0.1160

The result of the experiment is shown as Table 1. The classification of 2 means we classify between advisor-student while 3 means we classify among advisor-student, student-advisor and cooperation. The cooperation item means whether we include the features of cooperation as the training set.

In two-classification, by comparing between without and with cooperation data set, the accuracy is improved from 98.32% to 98.57%. We can suppose this situation: the advisor-student and student-advisor relationship have opponent features, while student-advisor and cooperation don't have similar features. But here in two-classification, we set student-advisor and cooperation as the same label and treat them as the same classification. Therefore, the network could only separate between advisor-student and non advisor-student relationship, which does not improve the performance of the two-classification model very much. Thus the more reasonable way is to model the prediction problem as three-classification problem when using the cooperation data set.

In three-classification problem, I can find that when the training set is large enough, the accuracy can reach to 96.56%, which shows that our prediction model can separate the three classification quite well. However, if I predict new author pair using the trained model, I can find that if the true relationship is mentorship, the model predicts very accurately, but if the true relationship is cooperation, the model will predict some of them as mentorship with high probabilities. As we know, the data set of co-author is obtained by excluding advisor and student relationship in the paper's author list. So I propose that the training data set of mentorship may not fully labeled by users. So the true mentorship is seemed to be cooperation in the data set, which may be the reason that leads to the wrong prediction in the cooperation relationship.

Table 2: Results of mentorship prediction on different dataset with different number of features and different networks.

Data set	Features	Network	Train loss	Train accuracy	Test loss	Test accuracy
Ex	Ex 22	2*Dense	*	98%	*	95%
Optimized	Ex 22	2*Dense	0.1209	96.01%	0.11471	96.247%
Optimized	Newly 63	2*Dense	0.1018	96.49%	0.10262	96.48%
Optimized	Ex 22	Dense-Dropout-Dense-Dense	0.0953	0.9669%	0.0872	97.077%
Optimized	Newly 63	Dense-Dropout-Dense-Dense	0.0809	97.23%	0.07723	97.38%
Optimized	Ex22	Complex densely connected	0.1007	96.63%	0.0876	97.073%
Optimized	Newly 63	Complex densely connected	0.0261	99.20%	0.05497	98.797%

To compare our work with previous work, in two-classification model, we compare our method with them in terms of dataset, features and network and the result is shown in Table 2. Here, the optimized data set means the data set with cooperation we obtain from Acemap. In the item feature, the previous work used 22 features but we expand and optimize them to 63 features. From the results, we can see that the optimized data set is better to show the difference of advisor-student and non advisor-student relationship. Moreover, our features are better to show the inner information of the relationship. As for the network construction, the deeper and more dense the network is , the better the performance is. And from 2, we could see our network is much better than before. Compared to the previous work of the mentorship prediction, We have improved the prediction accuracy and expand its function by classification among three kinds of relationship.

5 CONCLUSION

In this project, we predict the mentorship and cooperation relationship by the classification model. In brief, our contributions are listed below:
(1) Optimize extracting features;
(2) Optimize the implementation of the code;
(3) Expand the function by three-classification model;
(4) Optimize the neural network model.

After the doing the project, I have some ideas to optimize the model further. First, from the truth that an advisor always have many students, otherwise not. So I can model the prediction network with more limitation. Secondly, as we know, the mentorship among the advisor's advisor, the advisor and the student is not circular,i.e., the student is not possible to be the advisor's advisor. However, the cooperation relationship among co-authors may be looped. By adding this limitation, we can avoid predicting the advisor-student or student-advisor relationship to be cooperation. I think that if we have more time, we can improve the performance of our model.

At last, I sincerely appreciate the Professor Fu, Doctor Jia for their guide. And I sincerely appreciate the senior Wenzheng Tao and Ziyu Wang for their help. Thanks for my partner Yi Fang. We enjoy a lot during the project.