Mining Advisor-Advisee Relationships in Scholarly Big Data: A Deep Learning Approach

Wei Wang, Jiaying Liu, Shuo Yu, Chenxin Zhang, Zhenzhen Xu, Feng Xia
School of Software, Dalian University of Technology
Dalian 116620, China
xzz@dlut.edu.cn

ABSTRACT

Mining advisor-advisee relationships can benefit many interesting applications such as advisor recommendation and protege performance analysis. Based on the hypothesis that, advisor-advisee relationships among researchers are hidden in scholarly big data, we propose in this work a deep learning based advisor-advisee relationship identification method which considers the personal properties and network characteristics with a stacked autoencoder model. To the best of our knowledge, this is the first time that a deep learning model is utilized to represent coauthor network features for relationships identification. Moreover, experiments demonstrate that the proposed method has better performance compared with other state-of-the-art methods.

Keywords

Deep learning; Relationship mining; Stacked autoencoders

1. INTRODUCTION

The benefit of mentorship for advisee is obvious. On the one hand, the extent to which advisee mimic their advisors' career choices and academic preferences, and learn their mentorship skills is unclear [1]. On the other hand, the lack of dataset for advisor-advisee relationships makes it not easy to handle these issues. There are several projects that aim to collect mentorships relationships, such as Mathematics Genealogy Project [1], The Academic Family Tree [2], and Academic Genealogy Wiki [3]. However, these methods heavily rely on volunteers' efforts, which results in limited records.

Fortunately, advisor-advisee relationships are usually hidden in the coauthor network [2], which enables us to automatically uncover these relationships. In this paper, we propose a deep learning based advisor-advisee relationships mining methods with a computer science bibliographic network extracted from DBLP and Academic Genealogy Wiki project. A stacked autoencoder model is used to learn both scholars' personal properties and network characteristics. Experimental results demonstrate that our proposed method has superior performance compared with four classical machine learning methods.

2. PROPOSED SCHEME

We employ the Stacked Autoencoder (SAE) as the foundation of our methods, which is a famous deep learning model. The SAE model allows users to easily inject the personal features and network characteristics as input without the manual effort of feature selection.In this section, the SAE model details and its settings are introduced.

2.1 Autoencoder

An autoencoder is an artificial network with one input layer, one hidden layer, and one output layer, which can find a lower-dimensional representation of input features. Given input vector $x \in [0,1]^N$, it aims at seeking a lower-dimensional representation $y \in [0,1]^M$ with $M < N$. The mapping function f between x and y is called encoding function and can be the logistic sigmoid function

$$f(x) = f(Wx + b) = \frac{1}{1 + exp(Wx + b)} = y \qquad (1)$$

where, W is a weighted matrix, b is an encoding bias vector. Then the autoencoder finds a second mapping function $f'(y) = f(W'x + b') = z$, such that the output z is equal to the input x.

2.2 Stacked Autoencoders

A SAE model is a series of autoencoder. Considering SAE with k layers, the first layer will be the autoencoder, with the training set as the input. After gaining the kth hidden layer, the input of the $(k+1)$th layer is the output of kth hidden layer. Thus, multiple autoencoders can be stacked together. Meanwhile, to use the SAE model to identify the advisor-advisee relationships, we put a logistic regression layer after the last output layer for relationship classification.

2.3 SAE Training Method

The SAE model is trained with a greedy layerwise unsupervised learning algorithm. The key point is to first pretrain the deep network layer by layer in a unsupervised way. Then, fine-tuning with BP is used to tune the model's parameters in a top-down direction to gain better results. To

[1] http://genealogy.math.ndsu.nodak.edu/index.php

[2] http://academictree.org/

[3] http://phdtree.org/

JCDL '16 June 19-23, 2016, Newark, NJ, USA
© 2016 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-4229-2/16/06.
DOI: http://dx.doi.org/10.1145/2910896.2925435

Table 1: Description of input features

Feature	Description
AA_i	academic age of i when first collaborating with j
AA_j	academic age of j when first collaborating with i
N_i	No. of i's publication before collaborating with j
N_j	No. of j's publication before collaborating with i
AD	academic age difference value between i and j
N_{ij}	collaborating times between i and j
CD	collaborating duration between i and j
FTA	number of times i and j being first two authors
$Cohesion$	similarity between i and j (first 8 years)

be specific, the procedure can be described as follows: 1) Train the first layer by minimizing the difference between input vector x and reconstructed vector z; 2) Train the second layer by taking the first layer's output as the input; 3) Iterate the second step for the desired hidden layers; 4) Use the output of the last layer as the input for the identification layer, and the weights of each layer as the initialized parameters for BP supervised training; 5) Optimize the parameters of all layers with BP method in a supervised way.

2.4 Settings of SAE Model

To apply the SAE model to mining advisor-advisee relationships, we need to determine the number of input features, the number of hidden layers, and the number of hidden units in each hidden layer. For the input features, we both consider the advisor and advisee personal properties, and the ego network properties. Specifically, given a advisee i and his/her collaborator j, the input features can be seen from Table 1. These features are collected from the first 8 years of i's academic career. One's first academic age is the time point of publishing first paper. The Cohesion between i and j after collaborating $t (t \leq 8)$ years can be calculated as:

$$Cohesion_{ij}^{t} = \frac{T_{ij}}{2}(\frac{1}{T_i} + \frac{1}{T_j}) \tag{2}$$

where, T_{ij} is the number of co-publications between i and j in t years, T_i is the number of i' publications, and T_j is the number of j' publications. We calculate the cohesion values every year between two collaborators. Thus, we have 16 input features in total. Meanwhile, we normalize all the input features into [0, 1].

In this work, we choose the hidden layer size from 1 to 6, and the number of hidden layer units from 1 to 15. After performing grid finding task, we acquired the best setting for our methods. The best settings consist of two hidden layers, and the number of hidden layer units in each layers is 8.

3. EXPERIMENTS

3.1 Data Description

The proposed deep learning model was applied to the data collected from the Academic Genealogy Wiki project. The mentorship dataset is collected from 16 famous universities such as Carnegie Mellon and Stanford in the field of computer science. The dataset contains 3423 advisees and corresponding 343 advisors. We then gain collaborators of each advisee and their properties from DBLP. Thus we can get the ego network of each advisee. Obviously, one of the collaborators in each scholar's ego network is his/her advisor. We randomly select 80% nodes as the training set and the rest as the testing set.

Table 2: Performance comparison for SAE, LR, KNN, and SVM

Method	Accuracy	Precision	Recall	F1-Score
LR	0.89	0.90	0.86	0.88
KNN	0.87	0.91	0.83	0.87
SVM	0.91	0.85	**0.94**	0.81
DT	0.83	0.84	0.82	0.83
SAE	**0.91**	**0.92**	0.91	**0.91**

3.2 Results

To evaluate the effectiveness of our model, we use four performance indexes, which are the Accuracy, Precision, Recall, and F1-Score. We compared our method with four supervised learning methods. They are Logistic Regression (L-R), K Nearest Neighbor (KNN), Support Vector Machine (SVM), and Decision Tree (DT).

Table 2 summarizes the performance comparison of above methods on DBLP data set. From the experimental results, we can see that SAE model outperforms other machine learning methods. The stacked autoencoder as our deep learning architecture result in a accuracy of 0.91. Meanwhile, other machine learning methods can also reach the accuracy more than 0.83. Similarly, our proposed methods has the highest precision and F1-Score. The higher accuracy, precision, and F1-Score underlines the idea that deep learning algorithms can outperform classical machine learning algorithms. Meanwhile, the results also indicate that the deep learning approach can be successfully applied into social network analysis.

4. CONCLUSION

In this work, we propose a deep learning approach with a SAE model for mining advisor-advisee relationships. We consider both the scholar's personal properties and network characteristics as input features. Experimental results show that our method can achieve better performance compared with several classical machine learning methods. In the future work, we will apply our proposed model to the whole DBLP digital library to obtain a large-scale mentorship data set, which will enable us to study the interesting application such as mentor recommendation.

5. REFERENCES

[1] R. D. Malmgren, J. M. Ottino, and L. A. N. Amaral. The Role of Mentorship in Protégé Performance. *Nature*, 465(7298):622–626, 2010.

[2] C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, and J. Guo. Mining Advisor-Advisee Relationships From Research Publication networks. In *Proceedings of the 16th KDD*, pages 203–212. ACM, 2010.