Predicting Publication Conference with Deep Learning

Jingyuan Liu
AndrewId: jingyual
jingyual@andrew.cmu.edu

1 Introduction

When researchers are writing a paper, they will decide which conference it should be submitted to. Submitting a paper to a proper conference is a very important. A conference with close research topics and good reputation is likely to make the paper more influential. Therefore, in general, good conferences means good quality of a paper in related research topics.

Predicting the conference of a paper is an interesting and challenging problem. Information of the paper, like words, authors, organizations and time are highly related to the paper conference information. Conference contains several latent research topics, which will overlap with the latents topics of paper submitted to it. If I could properly represent the latent research topic information and capture the relation of the latent information and conference label, then I can solve the predicting challenge. To predict the conference for a publication, the conference will be treated as the classification label, and other information of the publication will be treated as features, for example, the authors, the organizations, and paper time.

Besides, prediction method could extend to a more interesting recommendation model. If I use generative classification models for the problem, I can get all parameters after training. Then I can use the trained model to recommend conferences for a new paper to submit. Given the information of all attributes of a paper, which could be transferred to features of the model, like authors, organizations, and time, I can use the trained model to get an output of the conference. If the model can get impressive performance, this model could be quite useful to provide suggestions for those new researchers.

2 Data Description

The dataset is a mixture collection of different papers in different conferences by different authors published at different time. I treated the conferences of those published paper as the label for the paper instances. Other attributes of a paper as features. Specifically:

Authors: authors of a paper is a very important attribute of a paper. I can use an indicator vector to encode this attribute to a vector as input feature. At first, I assign every different author a index. Then for each paper, I would put 1 to the position of the index for every author in the paper, otherwise 0.

Organizations: it is similar to authors.

Time: time is also very import attribute for a paper. The research topics of a conference would change over time. Therefore, when predicting conference for a paper, I need to take the time into consideration. I could use the year as the input of the feature.

Words: words in a paper is likely the most important information for mining the latent topics of a paper. Therefore, words features are very important for prediction of a paper conference. Most NLP problems use ”bag of words” method to encode words as features.

I will use a dataset obtained online for this model. The dataset was published by Prof. Jie Tang from Tsinghua University, which contains over 20+ million papers. After preprocess, the dataset currently contains:

paper: 1.5m +

author: 60k +

conference: 60k +

year range: 1951 $\sim$ 2014

organization: 20k +

words: 600k +

The dataset is currently too large for a normal classification problem. I will further preprocess the dataset to get a more reasonable and feasible dataset with smaller size.

3 Baseline Methods

The proposal project is a classification problem in the supervise learning filed. There are many traddtional methods could be used to solve this challenge. For example, Naive Bayes, and Decision Trees, could be used to solve this problem.

Naive Bayes is a generative classification model. I can use words of a paper as feature input and conference token as label for training. Decision Tree is a discriminative classification model. I can use decision tree to generate rules and "split" data into different groups for the classification task.

4 Proposed Model

I want to use deep learning to solve this problem. Deep learning is currently becoming more and more popular and powerful in several machine learning and pattern recognition related research fileds, for example, recognizing cat faces for computer vision, and word2vec for natural language processing. With large datasets, deep learning methods usually perform obviously better than traditional classification models for supervised learning.

First, I can use word2vec to encode word features for words in a paper. Traditional classification method use the bag of words model to represent word as features, which may lose the information of word orders and word relations. Word2vec is a model proposed by Mikolov, which is a distributed representation of word and phrases. This model could encode word to a vector, which performed better than bag of words in some cases.

Next, I would use a neural network model for classification. I would at first build simple layers for the function. For example, for the first layer, I would use the sum of feature input and an activation function. For the second layer, I would use a logistic basis. And at last, I would use a softmax function to make the output a distribution. For model training, I would use Backprop and SGD to optimize model parameters iteratively.

I would use early stopping and some other method to avoid overfitting during training for NN model. After training, I would conduct cross validation with the proposed model and compared to baseline models. I will give a detailed analysis of those models mentioned above.