Machine Learning methods for Author Name Disambiguation

Rank 8 Solution of KDD CUP 2024 OAG-Challenge WhoIsWho-IND Task

Weiliang Ji†
Z Lab
Chengdu, Sichuan, China
626255007@qq.com

ABSTRACT

This paper describes the Rank 8 solution of KDD CUP 2024 OAG-Challenge WhoIsWho-IND Task, and the code has been released at https://github.com/zui0711/Z-Lab/tree/main/2024%20KDD-Whoiswho . The task is to develop a model to discover paper assignment errors for given authors. We take use of 3 kinds of embedding methods combining with manual feature engineering. Then we build single-models based on LightGBM and Xgboost with several subsets of features and apply an ensemble for these models aiming at a high weighted AUC.

CCS CONCEPTS

•Computing methodologies ~ Machine learning ~ Machine learning approaches

KEYWORDS

Name Disambiguation, Embedding, Gradient Boosting Decision Tree

1 Introduction

The overarching goal of academic data mining is to deepen our comprehension of the development, nature, and trends of science. It offers the potential to unlock enormous scientific, technological, and educational value. For example, deep mining from academic data can assist governments in making scientific policies, support companies in talent discovery, and help researchers acquire new knowledge more efficiently.

The landscape of academic data mining is rich with entity-centric applications, such as paper retrieval, expert finding, and venue recommendation. However, community efforts to advance academic graph mining have been severely limited by the lack of a suitable public benchmark. For KDD Cup 2024, the host presents OAG-Challenge, a collection of three realistic and challenging datasets for advancing the state-of-the-art in academic graph mining.

2 OAG-Challenge WhoIsWho-IND Task

The increasing number of online publications has made the name ambiguity problem more complex. Moreover, the inaccurate disambiguation results have led to invalid author rankings and award cheating. This competition hopes participants develop a model to discover paper assignment errors for given authors.

Given each author's profile, including author name and published papers, participants are asked to develop a model to detect incorrect paper assignments among all one's papers. The paper attributes are provided, including title, abstract, authors, keywords, venue, and publication year. The data used in this competition is collected from AMiner.cn. The dataset consists of 1,149 authors and 317,302 papers.

We adopt weighted Area Under ROC Curve (AUC) which is broadly adopted in anomaly detection as the evaluation metric. For each author, weight can be calculated as follows:

$$Weight = \frac{ErrorsOfTheAuthor}{TotalErrors}$$

For all authors (M is the number of authors), weighted AUC can be calculated as follows:

$$WeightedAUC = \frac{ErrorsOfTheAuthor}{TotalErrors}$$

3 Method

3.1 Solution overview

We take use of the common data mining method, doing feature engineering first and then modeling with these features. We now take a quick look at our solution, it's shown as Figure 1.

*Article Title Footnote needs to be captured as Title Note
†Author Footnote to be captured as Author Note

WOODSTOCK'18. Iune. 2018. El Paso. Texas USA

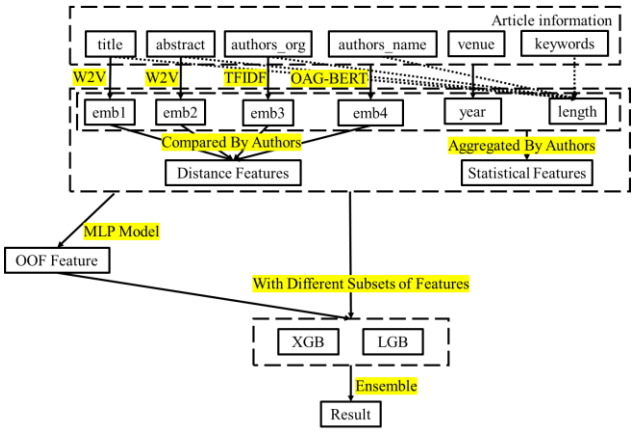

Figure 1: Solution Overview

We mainly use 3 different kinds of embedding methods to convert text fields to dense vectors, including W2V [1], TFIDF [2] with SVD [3] and OAG-BERT [4]. And we simultaneously perform manual feature engineering, features such as length and distance are constructed.

3.2 Features Engineering

Our features can be split into 3 parts: the base features of papers, derived features generated by aggregating authors and a feature created by a neural network model based on the above two kinds of features.

3.2.1 Basic Paper Features. Paper features mainly conclude embedding features and length features. We take use of W2V to represent the *title* field and the *abstract* field. Both of two field are convert to 32-dimensional vectors. When extracting organization information of authors from the *author* filed, we can use TFIDF with SVD to represent author organizations as 32-dimensional vectors. What's more, taking all fields of paper as input at the same time, we are able to get the embedding of the paper from OAG-BERT, which is a 768-dimensional vector.

Apart from embedding features, we construct some basic statistical features from text fields, details as follows:

1. Length features, such as length of words in title, abstract and keywords.
2. Number of the authors of a paper

3.2.2 Aggerated Paper Features. Completing conducting basic paper features, we aggregated authors to caculate the mean values of the basic paper features to represent an overall view of the papers assigned to one author (considering the fact that for most authors, the papers assigned to them are right). Then we can calculate the ratio of one feature of an paper with its mean value to measure the deviation of this paper in all papers assigned to this author. Similarily, for embedding features, we are able to calculate the distance between it and the mean embedding of this author representing the center of embedding vector space, which can get similar result to calculating the pairwise distances between papers assigned to this author and computing the average. However, our method has ability to significantly reduce the computational load.

We caculate the ratio of basic features and mean features as additional features. The ratio of venues of is also caculated.

3.2.3 NN Feature. Taking advantage of basic paper features and aggerated paper features as inputs with the label (normal or outlier) as output, we can generate another feature by training a neural network model to integration information from dataset better. In detail, our neural network model simply consists of linear layers, which is shown in Figure 2.

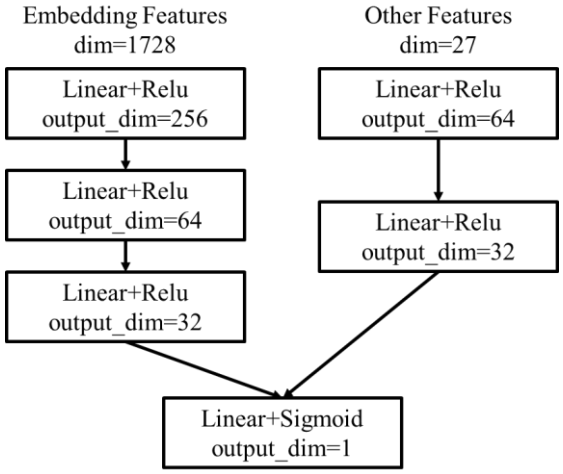

Figure 2: Architecture of NN Model

3.3 Modeling

We perform cross validation on the dataset grouped by author ID to avoid papers of one author simultaneously appear in the train dataset and the validation dataset which may lead to information leak. Taking use of features mentioned above, we build our models based on XGBoost [5] and LightGBM [6]. Specifically, we select multiple feature subsets by crossing the aforementioned features to train multiple models. Actually for every feature subset, we use 3 different random seeds to train the model to get more single-models. Get all single-models trained, we ensemble them according to the scores on leaderbord.

4 Experiments

Our experimental hardware consists of an AMD 5600X CPU, 64GB of memory, and an NVIDIA 3080Ti GPU with 12GB of VRAM. The conclusion of experiments are summarized as Table 1. We can find out the fact tha the key point of out solution is taking use of various kinds of embeddings to

represent papers and distance features to measure the deviation of one paper in all papers assigned to this author.

Methods	Test WeightedAUC
Baseline (with Basic Static Features)	0.624
Add Aggerated Paper Features (without Embedding)	0.678
Add Embedding Features (without OAG-BERT)	0.778
Add OAG-BERT Features	0.790
Add NN Featurs	0.795
Ensemble	0.799

Table 1: Record of experiments

5 Conclusion

Combining 3 kinds of embedding methods with manual feature engineering and the nerual network modeling feature, we are able to describe the authors and papers from many different aspects. Futhor more, we develop our model based on LightGBM and Xgboost, which can accurately discover paper assignment errors for given authors, reaching a high weight auc at 0.79941.

In the future work, we are going to make use of Large Language Models (LLMs) and Graph Nerual Networks (GNNs) to collect more information from dataset aming at achieving higher detection accuracy.

REFERENCES

[1] Mikolov, Tomas, et al. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.
[2] Jones K S. A Statistical Interpretation of Term Specificity and Its Application in Retrieval. Journal of Documentation, 28(1): 11-21, 1972.
[3] Liu, Xiao, et al. Oag-bert: Towards a unified backbone language model for academic knowledge services. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022.
[4] Demmel, J. and Kahan, W. Computing Small Singular Values of Bidiagonal Matrices With Guaranteed High Relative Accuracy. SIAM J. Sci. Statist. Comput. 11 (5), 873-912, 1990.
[5] Chen, Tianqi, and Carlos Guestrin. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.
[6] Ke, Guolin, et al. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 30, 2017.