

Sample Project Proposal: Paper Categorization and Author Clustering in Computer Science Domain

1 INTRODUCTION

Given a set of research papers from top data science conferences, can machines read the papers and automatically do the following two tasks?

- Categorize a future paper into a proper conference.
- Detect data scientist communities.

These tasks can be formulated as the following problems:

- Multi-class classification: Given a training data set of papers (as data objects), conferences (as class labels) and authors/keywords/references (as features), build classification models to predict the class label of test paper.
- Clustering: Given a data set of authors (as data objects) and papers/conferences/keywords/references (as features), use clustering algorithms to find groups of authors (i.e., data scientist communities).

We plan to use a few public datasets, such as DBLP (give a url here), Microsoft Academic Graph (give a url here), and Arnetminer (give a url here). All the datasets have been downloaded...

{It will be great but not necessary for a proposal if you can write down the following:} For the classification task, we propose to try a few models including Decision Trees, Naive Bayes, Support Vector Machines, and Neural Networks. For clustering, we propose to use K-Means clustering, K-Medoids, density-based clustering (like DBSCAN) and spectral clustering...

{It will be great but not necessary for a proposal if you can write down the following:} For the classification, we will divide the data set into training and testing parts. We will use the training set to learn classification models and do cross-validation evaluation on the testing set. For clustering, we will use visualization and some quantitative measures like Beta-CV to evaluate the performance...

2 RELATED WORK

Review classification and clustering papers/code packages/tools [1].

3 PROBLEM DEFINITION

4 PROPOSED METHODOLOGY

5 DATA AND EXPERIMENTS

5.1 Data set

5.2 Experimental settings

5.3 Evaluation results

6 CONCLUSIONS

REFERENCES

- [1] Jiawei Han, Jian Pei, and Micheline Kamber. 2011. *Data mining: concepts and techniques*. Elsevier.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference'17, Washington, DC, USA

© 2016 ACM. 978-x-xxxx-xxxx-x/YY/MM...\$15.00

DOI: 10.1145/nnnnnnn.nnnnnnn