Machine Learning Project Proposal

Bing He, Huili Huang, Xingchi Li, Zuoxin Tang, Jingfeng Yang

September 2019

1 Introduction

There are numerous machine learning methods to process and extract the information from texts: clustering, supervised classification, skip-gram, etc. The difficulty of text classification is sparity, polysemy and synonymy between words-words, words-documents, and the situations they applied to. We want to investigate a method for text classification using various datasets. We choose predictive text embedding (PTE), a semi-supervised representation learning method for text data, to deal with this due to its efficiency in incorporating arbitrary unsupervised and supervised information when constructing embeddings, compared with supervised models, like CNN and unsupervised models, like Skip-gram.

2 Methods

In semi-supervised sentence classification, to leverage both annotated data and unannotated text data, we first build a heterogeneous graph whose edge consists of word-sentence, word-word, and word-label. Then we learn the representations of each vertex in the graph. Finally, we could use them as input of the text classification task. This framework is proposed by Tang et al.[2] There are several methods we would like to try in the second step. The first is to use PTE [2] training process. The second is to use Graph Convolutional Networks (GCN) [1]. Yao et al. [3] has tried GCN in the text classification task, but they did not include existing labels as vertices in the graph. We argue that label vertices are crucial, because it can introduce supervised information when training word embeddings. Because of the heterogeneity, we would like to try techniques in Heterogeneous Graph Neural Network. We plan to use three datasets to test our model: one movie review dataset[1] and two sentiment classification datasets[2].

3 Results

We will be able to determine the approach with the best performance using the three datasets and the contribution of labeled and unlabeled data in semi-supervised learning to the final result, as well as the efficiency and the parameter used. Finally, we also present document visualizations to vividly illustrate our result.

4 Discussion

PTE relates the unlabeled training data directly to the specific task, outperforming methods that train word/semantic embedding separately. Thus, we expect that utilizing recent advancements in graph-based neural network structure and traning method like proposed in [1] can further enhance PTE method. We also expect that techniques in Heterogeneous Graph Neural Network like attention mechanism will improve our performance in this project because of our heterogeneous construction of graph and the nature of text analysis task.

References

[1] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[2] Jian Tang, Meng Qu, and Qiaozhu Mei. Pte: Predictive text embedding through large-scale heterogeneous text networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1165–1174, New York, NY, USA, 2015. ACM.

[3] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7370–7377, 2019.

[1] Movie Review Data: `http://www.cs.cornell.edu/people/pabo/movie-review-data/`

[2] Multi-Domain Sentiment Dataset (version 2.0): `http://www.cs.jhu.edu/~mdredze/datasets/sentiment/`
Twitter Sentiment Analysis Training Corpus (Dataset):
`http://thinknook.com/twitter-sentiment-analysis-training-corpus-dataset-2012-09-22/`