Extracting Latent Social Dimensions

Felix Gutmann
felix.gutmann@barcelonagse.eu

Nicholas Halliwell
nicholas.halliwell@barcelonagse.eu

June 24, 2016

Introduction

The focus of this project is around data this is not independent and identically distributed. Most, if not all social media datasets are like this, as there exists dependent relationships between individuals in the network. These networks are said to be multidimensional, given the many ways people can connect. The authors in [1] propose a method to extract unobserved variables the network, using relational learning to address the dependency issue. These "social dimension" describe how users in the network connect with each other. Essentially the authors are labeling nodes in a network graph, however, nodes can obtain multiple labels. This makes predictions difficult, the method of extracting "social dimensions" proposed in [1] was shown to outperform relational learning methods. In this project, we implement this method of "social dimension" extraction, and using the same dataset, attempt to improve their results.

Summary

Here we summarize the results of [1] where they propose and implement an algorithm to extract these "social dimensions" and compare classification results on several datasets. First and foremost, the authors are interested in labeling nodes of a social media network where nodes can take multiple labels. The authors give the following example; Actor one connects with Actor two on social media as they work for the same company. Actor two connects with Actor three as they attend the same gym. The data given to us is the interests of Actor one, however, given that Actor two and Actor three connect to Actor one, can we learn the interests of the other actors?

The proposed algorithm starts with a modularity matrix of the graph. After computing this from the graph, the algorithm then takes the k largest eigenvectors. The authors recommend k to be between

$400 - 600$, we chose 500 for convenience. These eigenvectors are the social dimensions, used as features for supervised multi-label learning. The authors use a Support Vector Machine using a one-vs-all method and compare classification performance using the Micro-$F1$ and Macro-$F1$ metric. They compare their method to six others, finding their method beats all others in most cases.

The dataset we use comes from Flickr, a website where users upload photos, connect with friends, and tag photos. These tags serve as node labels. This dataset consists of 195 different labels, $80,513$ actors, who are represented as nodes in the graph, and $5,899,882$ links connections between these actors.

Implementation and Results

We attempt to replicate the results from [1] using the Flickr dataset. We subset this dataset for computational convenience, taking 10% of the original dataset, and constructing the test set to be 25% of the training set. We decide not to use a Support Vector Machine, as running this algorithm on such a large dataset was computational infeasible. After over 8 hours of running, we had no results. After subsetting the data, we run the SVM used in [1], however, we were not able to match their results. The authors do not explain the parameters they used to tune the algorithm. In addition, they also leave out the size of the test set, perhaps explaining why we their classification scores were so much higher than ours. We also implemented a random forest using a one-vs-all method, a classifier known for performing well on large datasets. Even with this classifier, we were not able to match the authors results. We attribute this due to the missing information regarding the size of the test set.

Conclusion

References

[1] Tang, Lei, and Huan Liu. "Relational Learning via Latent Social Dimensions." Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '09 (2009)