ITR Lab Week 3: Dimensionality reduction and clustering

Current Usage Investigation

As we are in an age where data grows exponentially, datasets available for sociological study tends to have more and more complex structure: demographic and socio-economic databases are containing ever-expanding lists of variables, while other types of data, such as text data, are presenting even greater structural complexity intrinsically. On one hand, increasing amount of available variables has enhanced datasets' potential of revealing multi-dimensional social phenomenon. On the other hand, this tendency also encourages researchers to apply dimensionality reduction in their research, either to exclude features with minor relevance to manage data size and optimize data structure, or to directly uncover meaningful 'dimensions' of culture or society within complicated datasets. Thus, the primary applications of dimensionality reduction in sociological analysis can be grouped into two major categories: as a part of the preprocessing pipeline, or as an unsupervised learning method.

As a part of the preprocessing pipeline, dimensionality reduction can improve performance of model training processes, especially for time-intensive models such as word embedding models, which can become increasingly resource-demanding as data size grows. For example, Grupta et al. applied kernel PCA to word similarity matrices to derive morphological information of words, which was then provided to word embedding models including word2vec and fastText to assist model training [2]. The result showed that this extra process not only increased the accuracies of word embedding models, but also significantly reduced the time spent on model training.

As an unsupervised learning methods, dimensionality reduction can be used to identify inherent structures within datasets without annotations. In the study of Ge et al. about Chinese traditional culture and risky assets holdings, researchers applied PCA to 2016 China Family Panel Studies and discovered that households that are more influenced by Confucian culture are less likely to participate in the risky financial market [1]. In this research, the 'influence by Confucian culture' is an underlying dimension discovered through dimensionality reduction that leads to another social phenomenon: diversity in household's participation in the risky financial market. Dimensionality reduction can also provide statistical insights for traditional structural sociological theory. To provide empirical evidence for the components of social capital, Saukani et al. applied Categorical PCA to explore the inner structure of variables related to social capital [3]. Apart from the traditional 3-dimensional definition of social capital, networks, norms and trust, researchers discovered that in Malaysia society, spirituality is also an important dimension that is rather independent from the other three. Dimensionality reduction is also widely used in meta-analysis of studies about certain academic topic. For example, considering the low consensus nature of Sociology, Schwemmer and Wieczorek applied a rather simple dimensionality reduction model 'wordfish' to compress high dimensional data into a continuum [4], and discovered the division of authors, topics and journals considering different schools of methods, as well as the marginalization of mixed methods approaches.

Data Application

This week I applied the dimensionality reduction and clustering methods on my dataset that contains all articles from People's Daily since its publication. Before, applying the algorithms, the raw dataset is first reprocessed to adjust its form accordingly to the requirement of these algorithms. Firstly, I extract articles from the database for each March, as this is the month in which the International Women's Day is celebrated every year. Therefore, articles published during this month put more emphasis on social gender issues and provide a clearer view of the Chinese Communist Party's stance and priorities on social gender-related topics. Secondly, every single article is tokenized and transferred into a list of words, among which predefined stop words and single characters are removed. Different from last week's ITR lab, the stop words dictionary has been adjusted to better fit the use case. For example, some words related to newspaper publication are added to the stop words list, and Chinese expression of numbers are removed using regular expressions. Finally, tf-idf value is calculated for different words and the tf-idf matrix is stored as a sparse matrix.

After that, I first directly applied the K means algorithm directly on the tf-idf matrix, however, as is expected, the silhouette scores for clustering results with k ranging from 2 to 20 are all not very satisfying. Therefore, as the final part of the preprocessing pipeline, I tried to first apply PCA and SVD to the tf-idf matrix before give it to the K means algorithm. The silhouette score for the K means model based on PCA-processed data is much more satisfying (although is still lower than 0.5). By checking the representative articles that are classified into the two categories. I also conducted temporal analysis on the count of articles classified into different categories across years.

Survey data about values on family and role of women can also provide insights into the status quo and development of public gender stereotypes. Therefore, I applied PCA to the Chinese General Social Survey 2017 (CGSS2017), a Chinese version of GSS. The original dataset has 12582 samples from different provinces of China and 783 variables, among which 25 variables are selected which are attitude variables about social gender. PCA is applied to the dataset to reduce its dimension from 25 to 2. We can regard observations from a same province as a cluster, and by averaging over every observation's score on different dimensions, variation among areas considering gender stereotypes can be revealed.

Reflection

The clustering analysis made on the People's Daily corpus showed that the discussion about gender issue can be generally divided into two categories. In the selected texts from Cluster 0, articles are mainly consisted of social critique and depiction of societal issues. Women's participation in the political fighting for social reforms and political reforms such as anti-corruption battles are vividly described in these reports, aiming to promote women's political engagement, as was suggested by last week's ITR lab result.

In Cluster 1, the texts generally center around the topics of development and achievements, or to say, the positive aspects of societal progress. These texts include introduction to women's participation in scientific progress, international friendships and sports achievements, focusing on themes related to national pride. They reflect a narrative style that emphasizes the collective efforts of groups of individuals including women working towards advancement in fields like science, sports, and culture. However, the result of the temporal analysis is not very satisfying: there seems to be no apparent change of numbers of articles categorized into different cluster across the years, except for the difference caused by the changes in the overall number of pages in the newspaper.

The PCA analysis on CGSS 2017 also provided meaningful insights. As is shown in the following picture, an observable pattern is that provinces that are close geographically or economically are plotted closer. For example, more advanced provinces such as Zhejiang, Guangdong, Liaoning and major municipal cities are closer to each other on the left part of the plot. In future study, I will consider taking other socio-economic or demographic information into the study and explore their potention relationship with score of different geographic areas on the two dimensions.

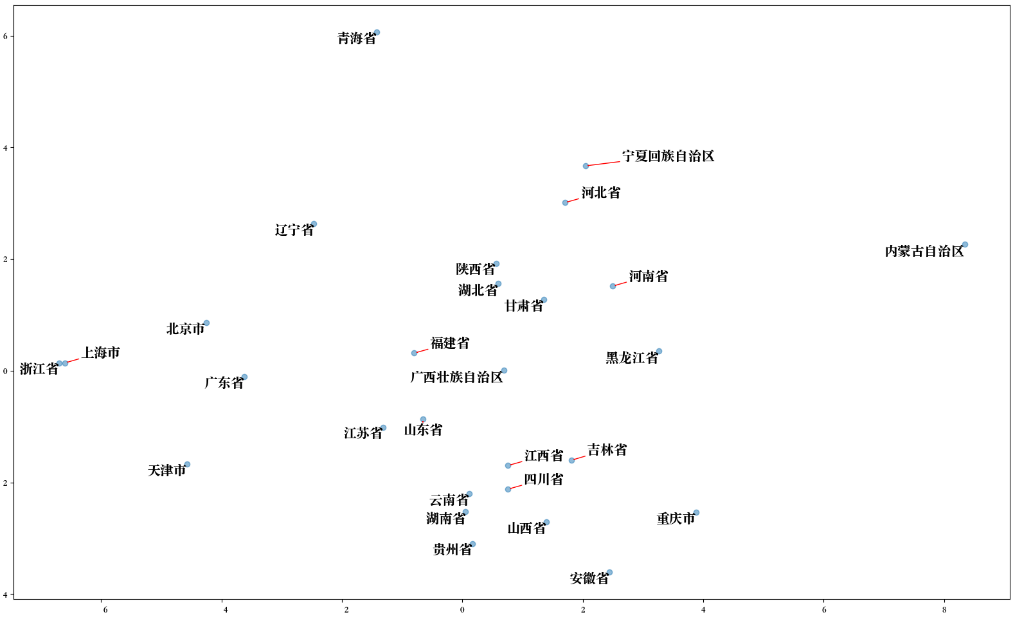

Figure 1: visualization of PCA result

References

[1] Yongbo Ge, Xiaoran Kong, Geilegeilao Dadilabang, and Kung-Cheng Ho. The effect of Confucian culture on household risky asset holdings: Using categorical principal component analysis. *International Journal of Finance*

& Economics, 28(1):839–857, January 2023.

[2] Vishwani Gupta, Sven Giesselbach, Stefan Rüping, and Christian Bauckhage. Improving Word Embeddings Using Kernel PCA. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 200–208, Florence, Italy, 2019. Association for Computational Linguistics.

[3] Nasir Saukani and Noor Azina Ismail. Identifying the Components of Social Capital by Categorical Principal Component Analysis (CATPCA). *Social Indicators Research*, 141(2):631–655, January 2019.

[4] Carsten Schwemmer and Oliver Wieczorek. The Methodological Divide of Sociology: Evidence from Two Decades of Journal Publications. *Sociology*, 54(1):3–21, February 2020.