Analyze the key issues and Visualize the research trend through Principal Component Analysis

MAO Hui, Pc Ng
Department of Electronic and Computer Engineering, HKUST.

Introduction

NIPS is a prestigious conference that showcase the research development in neural information and machine learning since 1987. To obtain further insight regarding the research trend in NIPS and also other communities in related domain, we had applied principal component analysis to analyze the key issues that received highly attention and also visualize the research trend through lasso. The key features were extracted from the title and abstract for all the papers span from 1987 to 2016. It was observed that the extracted features contain a number of noise, and feature selection and denoising are applied during the PCA process in order to obtain a more representable data model. Later, 10 hot topics is extracted from NIPS papers between 1987 and 2016.

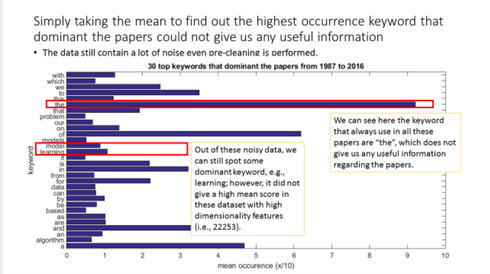
Contributions

The contributions of this work is summarized as follows:

- Prior to feature extraction, data recovery is conducted to retrieve those missing data especially the missing abstract.
- Data preprocessing is performed to clean the data to ensure highly relevant features can be extracted.
- Denoising is processed during PCA to further eliminate the noise components that affect the data representation.
- Visualize and analyze the topic trends of NIPS papers.

Methods

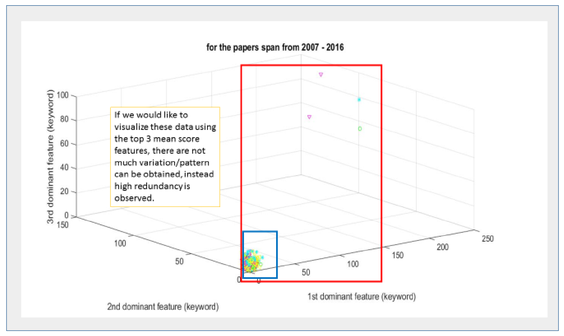
- The datasets from NIPS are used. It consists of all the papers published in NIPS from 1987 to 2016. The information of author, paper title, abstract, paper text are tabulated in 3 different csv files accordingly.
- Since some data did not come with complete information, that is about 3000++ data are missing the abstract out of the total 6560 papers.
- We extracted every words in each paper's title and abstract, the special characters and whitespace are cleaned during the preprocessing step before feature extraction.
- Average mean score is used to basics statistics of the data. The data dimension is 22253, and 30 top features are shown in below figure.



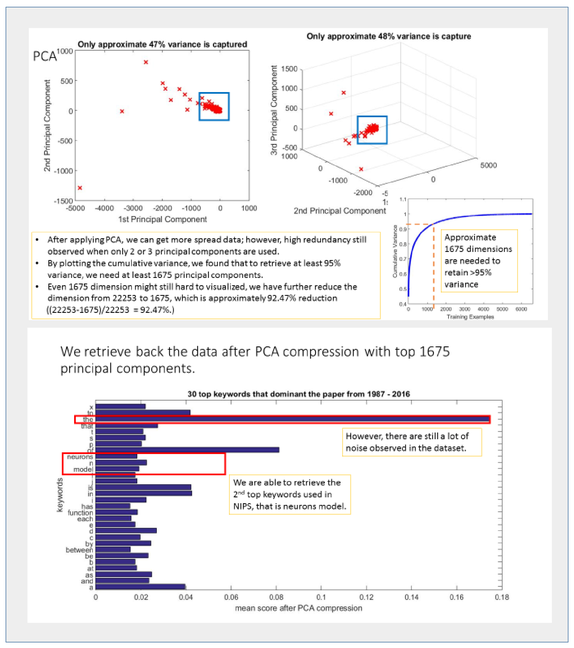
- PCA is applied to find out the best projection direction of the dataset such that the dimension of the dataset can be further reduced to retain only the features that give the highest variation.
- TF-IDF is applied to extract feature from content of the NIPS papers.
- SVD is applied to find the hot topics at NIPS papers.

Preliminarily Observation

- During the initial state (prior to PCA), we picked the top 3 features that give the highest mean score to try to retrieve some insights regarding this dataset. Unfortunately instead of showing some meaningful pattern, the data exhibits high redundancy and not much information can be obtained.

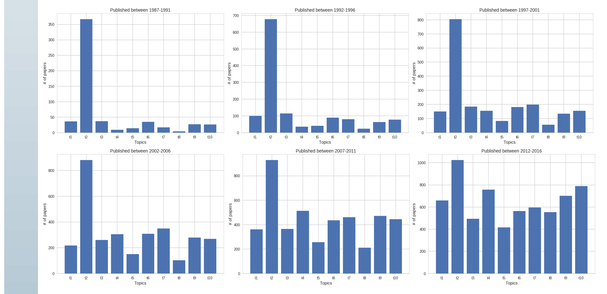


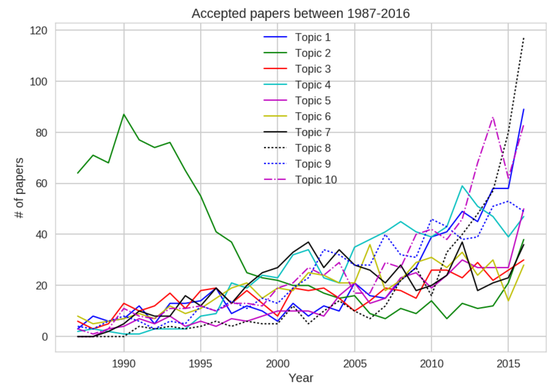
- To combat the issue with high dimensional data, PCA is employed. With PCA , we can further reduced the dimension by project the data to the direction that gives the highest variation.



Research Trend Analysis

- 10 Hot Topics: Algorithms, Neural Network, Reinforcement Learning, Statistical Models, Image Recognition, Decision Theory, Graph, Clustering.
- Topic 2 (Neural Network) is popular.





Conclusion

By leveraging PCA, we can have a better representation to model our data. For this project, we applied PCA for denoising at the same time secure the key features that give the highest variation. This give us a clear picture about the dataset, and in our case, PCA helps to identify the key issues that researchers were keen to discuss over the last few decades.

hmaoaa@connect.ust.hk (MAO, Hui)
pcng@connect.ust.hk (Pc Ng)

This is a poster for Math 6380 Mini Project 1, March 2017.