NIPS Paper Explained

Mengyue ZHA * Chutian HUANG †

Abstract

In this report, we conduct analyses on a collection of words appeared in NIPS papers ranging from 1987 to 2015. First, we find 5 topics from the words by Latent Dirichlet Allocation (LDA), namely, Computer Vision, Matrix Computation, Reinforcement Learning, Bayesian Methods and Time Series. We use t-distributed stochastic neighbor embedding (t-SNE) to visualize the results from LDA, and analyse the topic trends over the years. Next, we try to reduce the high-dimensional matrix to a low-dimension. We employ several data reduction methods, namely, Locally Linear Embeddin (LLE), Local Tangent Space Alignme (LTSA), Hessian LLE, Modified LLE (MLLE), Isometric Mapping (Isomap), Laplacian Eigenmap (Spectral Embedding), Principal Component Analysis (PCA), Multidimensional Scaling (MDS), to solve the problem. Detailed analysis is conducted.

1 Introduction

Neural Information Processing Systems (NIPS) is one of the major machine learning conferences. Over the years, the topics in NIPS has ranged from neuroscience to deep learning, computer vision, statistical linguistics, and information theory. In this report, we explore the datasets which include the title, authors, abstracts, and extracted text for all NIPS papers published on NIPS from 1987 to 2016.

We raise 3 questions:

- Does there exist certain categories of topics in NIPS and what are the topics?
- What are the trend of topics in NIPS over time?
- Can the topics be represented in a low dimensional subspace or sub-manifold?

2 Basic Data Exploration

2.1 Data Desciption and Preprocessing

There are totally 4 files in the database. First we look at the file `'NIPS_1987-2015.csv'`. It contains 11463 words and 5811 papers from 1987 to 2015. The head of the data covers the information of publishing year. However, 7 papers in them do not contain any word in the table, so we drop these 7 invalid columns, and the matrix size becomes 11463×5804. Each row represents one word, each column represents one paper, and the entry X_{ij} represents whether a word (i) is in the paper (j). It is a sparse matrix since 90% of the data are 0.

By grouping the paper-text data by publishing year, we obtain a histogram of the number of accepted papers per year, shown in figure 1.

*mzha@connect.ust.hk
†chuangat@connect.ust.hk

Department of Mathematics, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong.

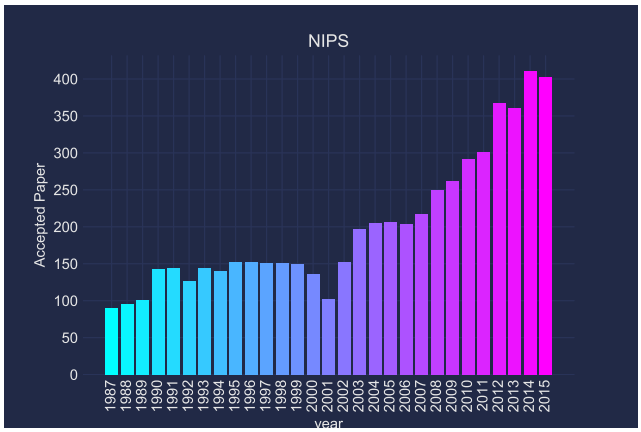

Figure 1: Accepted Papers per Year

From this bar chart, we can see that before year 2003, the number of accepted papers is relatively low, but after year 2003, the number has increased dramatically year by year. This is possibly due to the popularity of machine learning this century.

In virtue of the "worldcloud" package in Python, we draw the word cloud of the data, as shown in figure 2. The figure is drawn by deleting the common stop words, punctuation and any word with only 2 letters in advance. We can see that words like "learning, network, model,neural, algorithm" are mostly used in these papers.

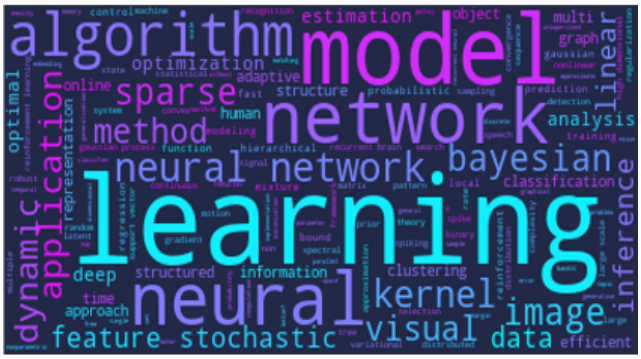

Figure 2: Word Cloud

2.2 Basic Exploration by LDA

Latent Dirichlet Allocation (LDA) is a model to generate certain numbers of topic and returns the weight of each word and each paper accounting for the topic. In detail, given the topic number k, the LDA takes the $w \times p$ word-paper matrix as input and returns two matrices, `paper_topics`, `top_topics` with size $p \times k$, $k \times w$ respectively. This model assumes each topic corresponds to a multinomial distribution of the vocabularies, and each paper also corresponds to a multinomial distribution of the topics. We apply this model here before further analysis because it leads us to a profound understanding of the intrinsic properties of the topics rather than the plain words.

Choosing am optimal number of topics, k, could be rather tricky. There are many methods to address this problem, for instance, using the perplexity value as a criteria. In our report, we choose $k = 5$ by empirical analysis, since the topics in NIPS mainly focuses on several specific topics in machine learning, like deep learning, computer vision, reinforcement learning, information theory, etc. Thus, we get the matrices `paper_topics`, `top_topics` with size 5804×5, 5×11463 repectively. We also obtain a table consisting of the mostly used vocabulary for each topic, shown in table 1.

Table 1: Top 6 words in 5 topics

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
training	algorithm	learning	model	model
learning	data	state	data	time
image	function	algorithm	models	neural
set	matrix	time	distribution	figure
network	learning	policy	inference	network
features	problem	function	gaussian	neurons

From the table above, we can see that indeed every topic has a rough theme. For instance, topic 1 includes words "training, learning, image, network, features", so it might concerns with deep learning training, feature extraction and computer vision. Topic 2 is possibly about matrix computational problem. Topic 3 could propose reinforcement learning algorithms, since it contains "state, policy, time". Topic 4 obviously employs Bayesian opinions in statistics due to the frequent usage of "distribtion, inference, gaussian". Topic 5 considers deep learning models, possibly about recurrent neural network with time series.

We draw the line chart of topic trends over the years using the `paper_topics` matrix, shown in figure 3. We can see that the beginning of this century is a turning point of the topic amount. It is consistent with figure 1 because more paper the conference accepts, more topics it will have. We can see that topic 1, matrix computation and topic 3, statistical models increases dramatically, and topic 2, reinforcement learning has also become more and more popular.Topic 0 and topic 4 do not increase as much as other topics, this is possibly because these two topics are dependent and are both about neural networks, so the total amount of these two topics is still large.

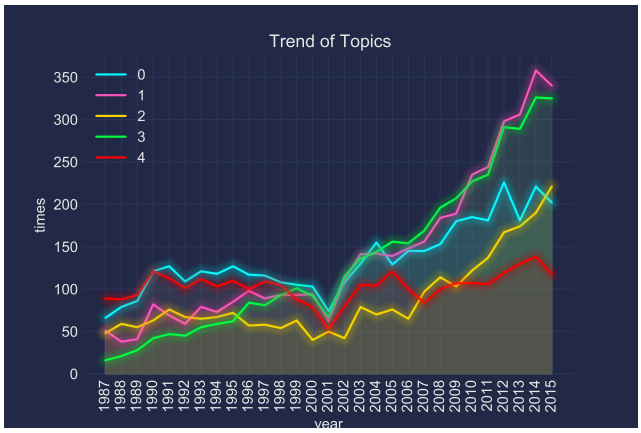

Figure 3: Trend of topics by Year

3 Data Reduction and Visualization

As we know, high-dimensional datasets can be very difficult to visualize. While data in two or three dimensions can be plotted to show the inherent structure of the data, equivalent high-dimensional plots are much less intuitive. To aid visualization of the structure of a dataset, the dimension must be reduced in some way. Linear dimensionality reduction ways include PCA, MDS. But linear frameworks assumes the data lies in a low dimensional linear subspace, it often misses the non-linear structure. Manifold learning addresses this problem by generalizing the linear frameworks to non-linear ones. In this section, we use PCA, MDS, LLE, LTSA, Hessian LLE, Modified LLE, Isomap, and Spectral Embedding method to visualize our `paper_topics`, `top_topics` matrix respectively.

First, let's briefly introduce these methods.

- PCA

Principal Component Analysis (PCA) is defined by a transformation which transforms the data to a new coordinate system such that the variance in descending order by some scalar projection of the data comes to lie on the coordinate in descending order. Thus the first few coordinates could store most information of the data.

- MDS

  (Classical) Multidimensional Scaling (MDS) projects the high-dimensional points to a low subspace by preserving the pairwise Euclidean distances of these points.

- LLE

  Locally Linear Embedding (LLE) writes each point as a linear combination of its neighbors. It seeks a lower-dimensional projection of the data which preserves the weights within local neighborhoods. A vital issue in this method is how to choose the number of neighboring points k. Here we choose $k = 6$.

- Hessian LLE

  When the number of neighbors is greater than the number of input dimensions, the matrix defining each local neighborhood is rank-deficient. The conditioning problem will lead to distortions in LLE maps.

  Hessian Locally Linear Embedding (Hessian LLE), also known as Hessian Eigenmapping, remedies the issue by revolving around a hessian-based quadratic form at each neighborhood to recover the locally linear structure. Unfortunately, the computational complexity of this method is quite high.

- MLLE

  Modified Locally Linear Embedding (MLLE) is another method to address the conditioning issue by using multiple weight vectors projected from orthogonal complement of local PCA.

- IsoMap

  Isometric Mapping (IsoMap) algorithm is one extension of the classical MDS method by changing the Euclidean distance into geodesic distance. It seeks a lower-dimensional embedding which maintains geodesic distances between all points.

- LTSA

  Local Tangent Space Alignment (LTSA) is motivated by the idea that when a manifold is correctly unfolded, all of the tangent hyperplanes to the manifold will become aligned. Similar to LLE, it starts by seeking geometric structure by computing the k-nearest neighbors of every point. Afterwards, LTSA characterizes the local geometry at each neighborhood via its tangent space, and performs a global optimization to align these local tangent spaces to learn the embedding.

- Spectral Embedding

  Laplacian Eigenmaps finds a low dimensional representation of the data using a spectral decomposition of the graph Laplacian. The graph is constructed from the neighboring points and can be considered as as a discrete approximation of the low-dimensional manifold in the high-dimensional space, each node representing a point and each connectivity between nodes representing the proximity of neighboring points.

We implement these methods to the `paper_topics`, `top_topics` matrix extracted by LDA, and obtain figure 5, 6. The legend of the following figures is shown in figure 4.

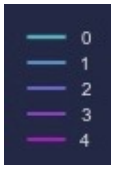

Figure 4: Color Legend

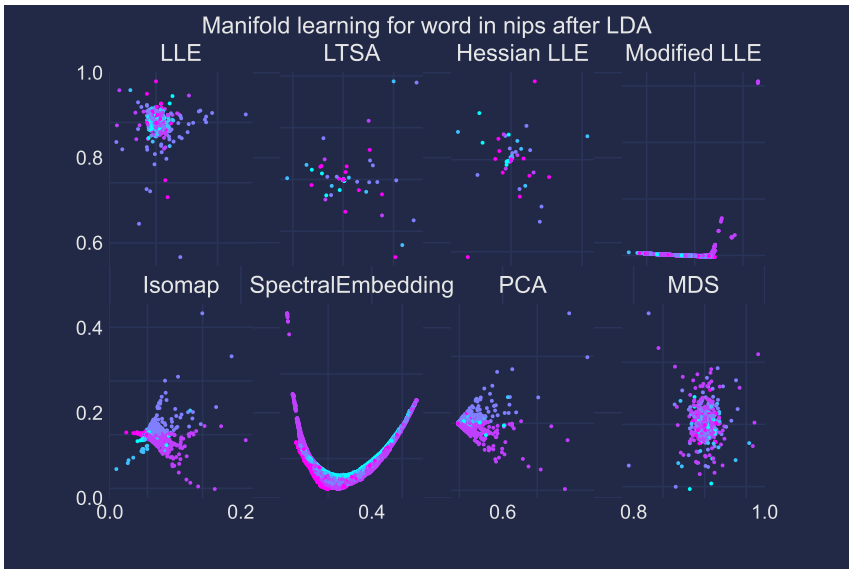

Figure 5: Manifold learning for *top_topics*

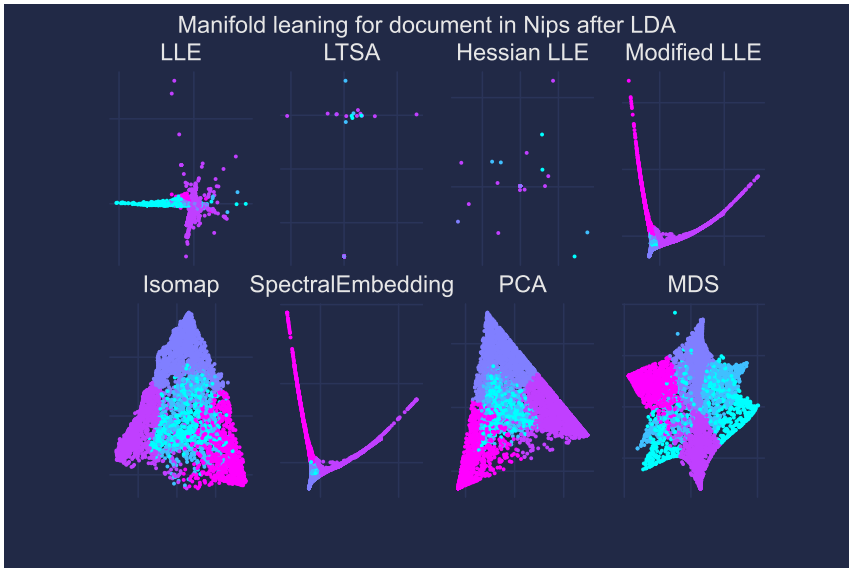

Figure 6: Manifold learning for *paper_topics*

These figures are colored in the following method: the bright green represents topic 0, the grey- blue represents topic 1, the grey-purple represents topic 2, the pink represents topic 3, and purple-purple represents topic 4.

Figure 5 shows the manifold learning results for `top_topics` matrix. Figure 6 shows the manifold learning results for `paper_topics` matrix. We can see that manifold learning for the matrix corresponding to the weights of topics in words do not show manifest results, but manifold learning for the matrix corresponding to the weights of topics in papers, some methods show satisfactory results. This is possibly because every paper certainly has one explicit topic; but according to linguists' research results, for most words used in text, the connection between the word and topic is quite weak, so the structure for words is hard to capture.

Now we focus on figure 6. From the eight methods, we can see that the methods in LLE family do not perform as well as PCA and the MDS family. This is possibly because the data does not lie

well in the manifold defined by LLE family. Spectral Embedding and MLLE have similar results by exhibiting topic 3 and topic 4 well. PCA, MDS and Isomap are able to exhibit the 5 classes of topics more distinctly than other methods, and the topics except topic 0 and 1 ramify out as branches in these methods.

We also drew the t-SNE results for `paper_topics`, `top_topics` matrix. The results are shown in figure . This is consistent with the the above manifold learning results and our analysis.

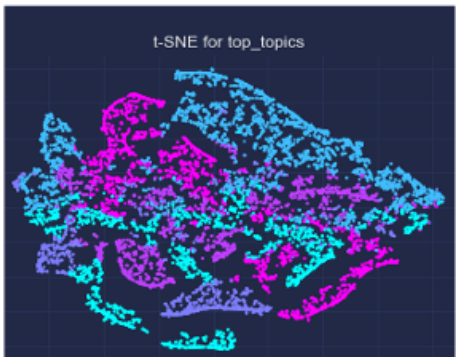

Figure 7: t-SNE for *top_topics*

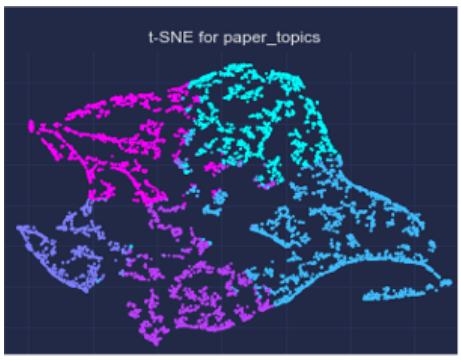

Figure 8: t-SNE for *paper_topics*

4 Conclusion

Now we can answer the questions raised at the beginning.

Given the large-scale matrix consisting of paper and text words from 1987 to 2015 in NIPS, we generate five topics from these texts, such as deep learning, matrix computation, reinforcement learning, statistics.

The beginning of this century is a turning point of the paper amount and topic amount, the number of them increase dramatically afterwards. The most popular topics around 2015 involves words like "matrix", "distribution" and "network".

The paper topics could be represented in a low-dimensional subspace through PCA, MDS, Isomap, but it degenerates using the LLE family methods. The words cannot have an explicit structure with the topics.

5 Individual Contribution

ZHA Mengyue mainly contributes to the codes and poster.

HUANG Chutian mainly contributes to the theory support and the report.

References

[1] Sklearn Document: Manifold learning. `https://scikit-learn.org/stable/modules/manifold.html`.

[2] Arpit Dwivedi. The Hottest Topics in Machine Learning. `https://www.kaggle.com/arpitdw/the-hottest-topics-in-machine-learning`.

[3] Gu Hanlin, Huang Yifei, and Sun Jiaze. A Dive Into NIPS Words. *CSIC 5011*, 2017.

[4] Wikipedia. Nonlinear dimensionality reduction. `https://en.wikipedia.org/wiki/Nonlinear_dimensionality_reduction`.

[5] Yohan. LDA visualized using t-SNE and Bokeh. `https://www.kaggle.com/yohanb/lda-visualized-using-t-sne-and-bokeh`.