MATH 6380Q Project 1 Peer Review

09. CHAN, Lok Chun. Realization of Recent Trends in Machine Learning Community in Recent Years by Pattern Mining of NIPS Words.

[ Reviewer 2 ]

In poster nine, “Realization of Recent Trends in Machine Learning Community in Recent Years by Pattern Mining of NIPS Words”, author used PCA and linear regression on the frequency distribution of the words being used to identify the recent trends in machine learning-related research fields. The NIPS word dataset contains the word count of 11463 words in 5812 journal articles collected from 1987-2015. Author used PCA and linear regression to get top 50 words from the dataset and then analyze them. The conclusion is that these top 50 rising words are not strongly related to “deep learning”. It is conjectured that the dataset may have certain complicated underlying structure which cannot be clustered effectively by linear projection method like sparse PCA, such that the signals of certain important features are masked in some low PCs.

The strength of this poster is that it uses PCA and linear regression method and finally get the result.

The weakness of this poster is the dataset is may not suitable for using PCA to analyze it.

Evaluation on Clarity and quality of writing: 4. There is a good use of examples and figures and the report is clearly written. Suggestion is using more figures to get the paper clearer.

Evaluation on Technical Quality: 3.

Overall rating: 3.

Confidence on my assessment: 3.

[ Reviewer 3 ]

This project adopted sparse PCA and linear projection on dataset of paper published in NIPS from 1987 to 2015 to identify the recent trends in machine learning field.

The initial idea is novel which is intended to discover the ascending key words in NIPS. However, the sparse PCA’s first principle component cannot distinguish those trends with highest frequencies.

In the introduction part, I suppose that ‘linear regression’ should be replaced by ‘linear projection’ based on the words on other parts. Then, there are limited examples and figures to help understand the results.

In the technical extent, sparse PCA may not give a good result on the NIPS dataset. I think the author can try other methods. Thus, he cannot elucidate the strength of this method. He talked mainly about the weakness of the method. The results are easy to replicate.

The overall rating is 3 and my confidence is 2.

[ Reviewer 4 ]

This report is clearly written and well organized.

Evaluation on Clarity and quality of writing: 5

Evaluation on Technical Quality:3

Overall rating:3

Confidence on your assessment:2

[ Reviewer 5 ]

Summary: In this study, analysis has been done on a collection of words appear in journal articles published on NIPS from 1987 to 2015. Sparse PCA and linear regression on the frequency distribution of the words being used, are employed to identify the recent trends in machine learning-related research fields.

Strengths: Linear regression is combined with sparse PCA.

Weakness: More results can be included in a more compact poster.

Evaluation on Clarity and quality of writing: 5.

Evaluation on Technical Quality: 4.

Overall rating: 4

Confidence on your assessment: 2

[ Reviewer 6 ]

8.1 Summary

For Chan Lok Chun’s work, he constructed a principle component analysis for identifying the recent trend in machine learning community.

8.2 Strength and Weakness

Chan has done a good job on explaining his idea, how does he use the PCA technology and how the problem based on this dataset are defined. The weakness of this report is that the content of analysis can be more developed.

8.3 Score

8.3.1 Clarity and Quality of Writing

The structure of this report is good. didn’t find any typo in this report. One flaw is that the analysis is a bit monotonous. I will give him 3/5 on this aspect.

8.3.2 Technical Quality

Good application of PCA, more content will be better. I will give him 3/5 on this aspect.

8.3.3 Overall

The overall score for this poster is 3/5.

[ Reviewer 7 ]

* **Summary of this report:** In this report, the SPCA is used to identify the recent trend in machine learning related research fields. It is concluded that the dataset cannot be clustered effectively by linear projection method like SPCA
* **Describe the strengths of the report:** The top 30 principle components and the corresponding projected data are shown.
* **Describe the weaknesses of the report**: The methods using in this report are not effective to analyze the complicated dataset.
* **Evaluation on Clarity and quality of writing (1-5):** 4

**Details of typos:** The list of words were (was) then analyzed manually to check which subfield of machine learning community do (does) they belong to.

* **Evaluation on Technical Quality (1-5): 3**
* **Overal ratings: 3.5**
* **Confidence on your assessment: 2**

[ Reviewer 8 ]

•• Summary of the report.

Use sparse PCA to analyze NIPS words data.

• Describe the strengths of the report.

Clearly written. Interesting analysis.

• Describe the weaknesses of the report.

The methods (SPCA and linear regression) might be a little bit simple.

• Evaluation on Clarity and quality of writing (1-5): Is the report clearly written? Is there a good use of examples and ﬁgures? Is it well organized? Are there problems with style and grammar? Are there issues with typos, formatting, references, etc.? Please make suggestions to improve the clarity of the paper, and provide details of typos.

4

• Evaluation on Technical Quality (1-5): Are the results technically sound? Are there obvious ﬂaws in the reasoning? Are claims well-supported by theoretical analysis or experimental results? Are the experiments well thought out and convincing? Will it be possible for other researchers to replicate these results? Is the evaluation appropriate? Did the authors clearly assess both the strengths and weaknesses of their approach? Are relevant papers cited, discussed, and compared to the presented work?

3

• Overall rating: (5- My vote as the best-report. 4- A good report. 3- An average one. 2below average. 1- a poorly written one).

3~4

• Conﬁdence on your assessment (1-3) (3- I have carefully read the paper and checked the results, 2- I just browse the paper without checking the details, 1- My assessment can be wrong)

2

[ Reviewer 9 ]

1. **Summary**

The author applied Sparse PCA and linear regression on the frequency distribution of the words being used, are employed to identify the recent trends in machine learning-related research fields. From their results, no key words strongly related to “deep learning”. The author assumes that some information may be masked in the low PCs.

1. **Strength of the report**

The objective of the task is clearly stated and conducted. The conclusion is reached using the method. The limitation of the results is stated.

1. **Weakness of the report**

The results of linear regression are not clearly discussed.

1. **Evaluation of clarity and quality of writing**

The report is clearly written and well organized. The figures are clearly explained. No obvious problems of the style, grammar, typo, and formatting are examined.

1. **Evaluation on technical quality**

The results are technically sound and no obvious flaws in the reasoning are found. Both the strength and limitations of PCA are discussed.

1. **Overall rating**

An overall rating of 3 is given.

1. **Confidence on your assessment**

2- I just browse the paper without checking the details

[ Reviewer 10 ]

*Summary:*

Used PCA and linear regression to check the trend ‘deep learning’ as a keyword in NIPS publication.

*Strength:*

Provided a creative way to identity the trend of keywords in NIPS.

*Weakness:*

The reasoning of the method does not explain well. Cannot get how the slopes of linear regression result is the top rising words in NIPS.

*Evaluation on Clarity and quality of writing (1-5): 2*

The course code may be better replaced by this course, that is, CSIC5011. (Do not know if the course has another course code of MATH6380J)

Since there is only one group member, maybe change ‘our’ to ‘my’.

*Evaluation on Technical Quality (1-5): 3*

It would be better if visual aids for “top 50 rising words” was provided.

*Overall rating (1-5): 3*

*Conﬁdence on your assessment (1-3): 1*

[ Reviewer 11 ]

Summary: This report analyzes the NIPS word dataset using PCA.

Strengths: The report is well-organized and is easy to understood.

Weaknesses: Some figure may need more explanations.

Writing: (5) The report is well-written and well-organized.

Technical Quality: (4) The claims are well-supported by the experiments. But I think Fig.1 need more explanations: the y-axis is the cumulative variance, but we do not know the total variance, why not plot it in percentage?

Overall rating: (4)

Confidence on my assessment: (2)

[ Reviewer 12 ]

**Summary:** This project uses Sparse PCA to analyze the NIPS word dataset. It draws the conclusion that most of the rising words are not related to machine learning and SPCA is not suitable for this dataset to extract important features.

**Strengths:** The topic uses SPCA to deal with this dataset and preprocess the raw data by averaging the counts to get relative frequencies. And it tries to explain the underlying reasons for its results and give some reasonable explanations.

**Weaknesses:** Figure 2 is not clearly labeled and it does not explain how to get the trending words from the extracted sparse PCs. The average method only remains the relative frequency of each word while losing the absolute value information. This makes it not possible to compare the absolute counts among different words and masks the mainstream words which always have a high number of counting.

**Evaluation on Clarity and quality of writing:** 3

**Evaluation on Technical Quality:** 3

**Overall rating:** 3

**Confidence on your assessment:** 2