What is Hot at NIPS?

Mao Hui (20297805), Pc Ng (20305846)*
Department of Electronics and Computer Engineering
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong.
{hmaoaa, pcng}@ust.hk

Abstract

NIPS is a very popular conference among the area of machine learning and computational neuroscience. Knowing what is hot at NIPS is not only beneficial to authors but also can help organizer prepare next conference more efficiently. In this mini project, we successfully mined hot topics that appeared at NIPS from 1987 to 2016 and compare the performance of difference classifiers for classifying papers into different topics. During the hot topic mining, Principle component analysis (PCA) is used for dimensions reduction and denoising, we found that PCA can boost the progress of hot topic mining and improve the performance of classification. We examines our methodology using an NIPS dataset provided by Ben Hamner[2]. An toy experiment is also conducted on 6 non-NIPS papers (2 AAAI papers, 2 ICML papers and 2 CHI) to show that our hot topic mining can work well on predicting whether one paper can be accepted by NIPS.

1 Introduction

Nowadays, prestigious conference always receive massive submissions with diverse paper quality. Some papers never even meet the theme, or the discussion of the paper is simply irrelevant to the research field of the said conference. This, undoubtedly, has increase the burden of the conference committees and reviewers, to organize the reviewing process and to review the paper, respectively. On the other sides, authors are always waiting in anxious after submitting their paper without any clue regarding the current status of their paper. However, related previous works such as [1] present a paper recommendation system based on the author's research interests, and [2] conduct some analysis regarding the metadate in IEEE Visualization Publications. And there are seldom works focus on mining hot topics of the submitted paper by far.

Motivated by these factors, this report proposes an unified methodology to do hot topics mining and predict the probability of acceptance for a submitted paper conditioned on the published papers in the history. This work aims to deliver two-fold objectives: to help the conference committees to run a quick scan on the submitted papers, and to give authors some clues based on the estimated predicted outcome. Nonetheless, the estimated prediction is used merely as a guideline. That is to say, we deem that decision of acceptance should not based on the proposed hot topics mining method alone, but through this prediction engine, at least, we can give either conference committees or authors some quick overview regarding the acceptance likelihood of a particular paper. And based on this probability of acceptance, conference committees can arrange the manual reviewing process most efficiently; whereas authors can judge for themselves whether the particular conference venue is the right venue to submit their work.

*This is a technical report for MATH 6380, mini project 2. A course conducted by Prof. Yuan Yao in The Hong Kong University of Science and Technology during Spring term 2017.

[2]https://www.kaggle.com/benhamner/nips-papers

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

The papers published in NIPS from 1987 to 2016 were used to experiment the propose methodology. Differ from the clean dataset (which did not include papers published in 2016) used by [3], this dataset tabulates the *title*, *abstract*, and *text* into each individual column with some missing values at some rows. Further effort is needed to reorganize this noisy dataset. The main contributions of this work is summarized as follows:

- **Hot topics mining**: we propose an unified hot topics mining methodology to find hot topics and academic trends at NIPS using the metadata of previous accepted NIPS papers. The mining results can be also used to predict whether one paper can be accepted.
- **Dimensionality reduction & Denoising**: the extracted feature (keyword) is in high dimensions with a lot of noisy data (meaningless keywords). These noisy data are redundant and might not give much meaningful information with their low variance. Principal component analysis (PCA) is then applied to reduce the dimensions of the data so that more efficient processing can be achieved for later training phase.
- **Automatic paper classification**: After hot topic mining, new papers can be classified into proper topics based on learned knowledge, which can improve efficiency of the processing of the conference.

The rest of the report is organized as follows. Section 2 presents the problem formulation; Section 3 describes the data pre-processing including feature extraction and dimensionality reduction; Section 4 describes the methodology of hot topics mining; and Section 5 concludes the paper.

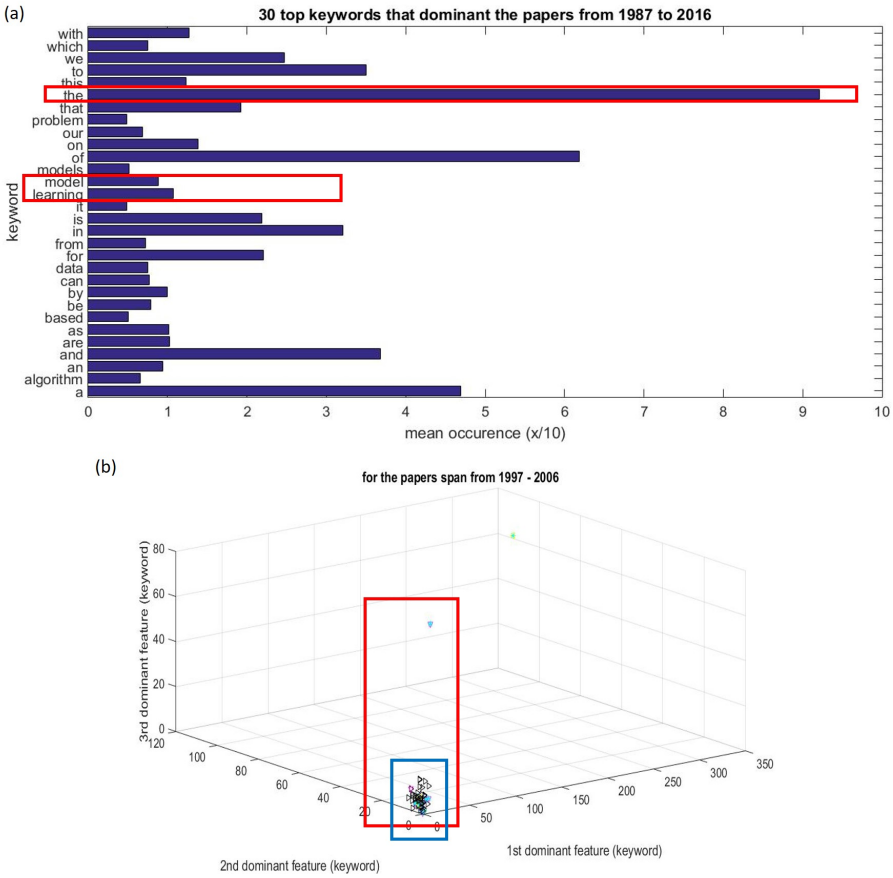

Figure 1: The mean occurrence of keywords does not convey meaningful information regarding the papers: (a) a plot of top N keywords in connection to its mean occurrences in each paper, and (b) visualization of each paper based on the top 3 dominant keywords.

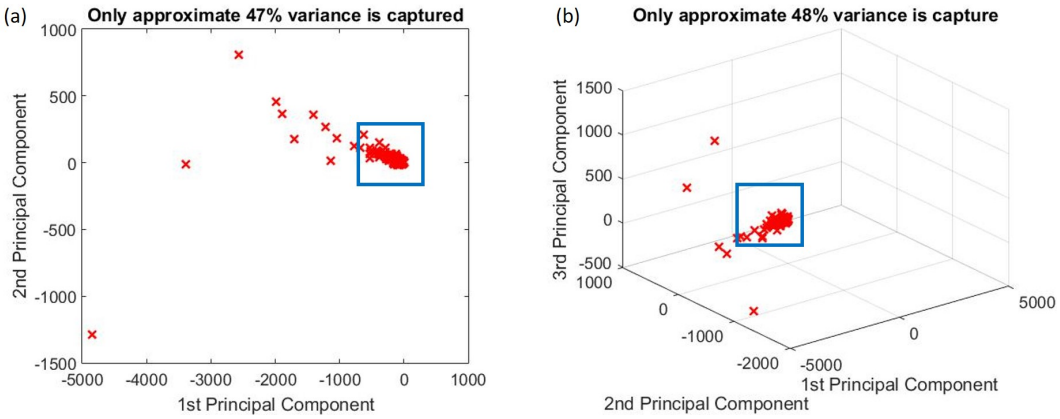

Figure 2: From the total PC(s): (a) approximate 47% variance is captured when top 2 PC(s) are used, and (b) approximate 48% variance is captured when top 3 PC(s) are used.

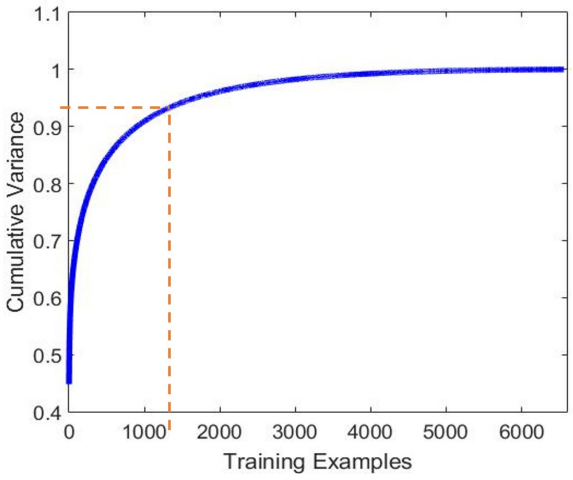

Figure 3: Approximate 1675 PC(s) are needed to ensure the features still convey more than 95% variance (information).

2 Problem Formulation

This section sets forth the problem to be examined in this report; that is, we would like to predict if a paper will be accepted for NIPS publication. Consider n number of papers submitted to NIPS, there are only k number of papers will be accepted for publication. According to NIPS history, the number of accepted papers k is generally smaller than the number of rejected papers, i.e., $k < n - k$. Such simple argument implies that the number of accepted papers is twice less than the number of submitted paper. In other words, we can say that more than half of the papers will be rejected. Mathematically,

$$2k < n \tag{1}$$

where both k and n are non-negative and non-zero integers.

Given such a low acceptance rate, we would like to investigate if the submitted paper is most likely to be accepted or rejected. This give rise to a simple Bernoulli process, with random variable X_i denotes the paper submitted to NIPS, where $i \in \{1, 2, ..., n\}$. It is obvious that this sequence of random variables are independent. Intuitively, the chance of acceptance for submitted paper X_1 will not affect the acceptance possibility for the submitted paper X_2. Let's denote the probability of acceptance as

p, then based on the argument in Eq. 1, we can say the the acceptance of a paper is bounded by

$$P(X_i = 1) = p < \binom{n}{k} p^k (1-p)^{n-k} \tag{2}$$

Since our final objective is to predict, for a given paper (random variable X_i), the chances to get accepted by NIPS, the final prediction outcome conditioned on Eq. 2 can be described as follows:

$$X_i = \begin{cases} 1, & \text{if } \hat{p} \geq p \\ 0, & \text{if } \hat{p} < p \end{cases}, \tag{3}$$

where $\hat{p}$ is the predicted probability based on the similarity measured obtained from SVM. Intuitively, we can say that a paper is most likely to be accepted if $\hat{p}$ is greater or equal than the lower bound of acceptance p.

3 Data Processing

This section describes the methodologies we had applied to process our data including feature extraction, cleaning and dimensionality reduction using PCA.

3.1 Feature Extraction and Cleaning

As discussed, the dataset consists of noisy elements such as *link to the pdf*, *event*, etc. Furthermore, some of the elements are with missing value, particularly the *abstract* column. To fix the *missing abstract* problem, we first search through the whole text file for the identified papers with *missing abstract*. If the *abstract* is unavailable in the text file, we will take the first 200 words from the *text* and add to the *abstract*. Once the above problem is fixed, we extracted the keywords available in both *title* and *abstract*. The intuition is that the main idea of a paper is usually captured in these two elements. In total, 22253 features (keywords) are extracted from the total of 6560 papers.

Since the extracted features still contain a lot of noisy data even though pre-cleaning is applied before feature extraction. To solve this problem, we applied another round of cleaning to filter out the 26 common alphabets and unrecognized characters. The occurrences of each keyword in each paper is counted and tabulated accordingly. In order to have a quick overview of the extracted keyword, we calculated the mean occurrence of each keyword and plotted a bar chart as shown in Fig. 1(a). It is obvious that by simply taking the mean of the keywords, it did not convey much information regarding the paper published in NIPS. From Fig. 1(a), we can see that there are too much noisy keywords (e.g., *with, the, we, which, to, it, from*, etc.) affect the correct interpretation of the paper's key idea. This meaningless keywords also impose high dimensionality curse to the desired features. Fig. 1(b) further show that there is no much variance can be observed from the top 3 dominant keywords obtained from Fig. 1(a). In other words, it simply means that these keywords are common in almost every paper. Such scenario drives us to employ PCA tool to further reduce the dimensionality and get rid of unwanted noisy keyword. The PCA process is discussed in the next subsection.

After applying PCA, we can get more spread data with reduced dimension; however, high redundancy still observed when only 2 or 3 PC(s) are used, as shown in Fig. 2. In general, it is acceptable to retrieve at least 95% variance, so to see how many PC(s) are needed in order to retain this 95% variance, we plotted the cumulative variance as shown in Fig. 3. From the graph, we can see that we only need to use approximately 1675 PC(s) to guarantee the features returns at least 95% variance. With original features in 22253 dimensions to 1675, PCA successfully reduce the dimensionality with 92.47% reduction.

3.2 Dimensionality Reduction

PCA [4] is a widely used statistical tool to deal with data with high dimensionality. Before applying PCA to our data, we first performed the normalization on each element obtained during the feature extraction phase. Given n dimensions features, and m number of papers, the features of each paper

Table 1: Eight topics with ten key words

Topics	Key Words
Topic 1	algorithm matrix convex theorem bound log loss problem optimization function
Topic 2	network networks units layer training neural input hidden output learning
Topic 3	policy state action reward agent actions reinforcement learning policies states
Topic 4	model data models distribution posterior latent bayesian inference gaussian likelihood
Topic 5	image images object visual features objects model feature recognition pixel
Topic 6	neurons spike neuron synaptic stimulus firing cells activity cell time
Topic 7	graph tree node nodes clustering graphs cluster algorithm clusters edges
Topic 8	kernel data training classification svm learning kernels xi classifier feature

Table 2: One paper's title under each topic

Topics	Titles
Topic 1	Skeletonization: A Technique for Trimming the Fat from a Network via Relevance Assessment
Topic 2	network networks units layer training neural input hidden output learning
Topic 3	policy state action reward agent actions reinforcement learning policies states
Topic 4	model data models distribution posterior latent bayesian inference gaussian likelihood
Topic 5	image images object visual features objects model feature recognition pixel
Topic 6	neurons spike neuron synaptic stimulus firing cells activity cell time
Topic 7	graph tree node nodes clustering graphs cluster algorithm clusters edges
Topic 8	kernel data training classification svm learning kernels xi classifier feature

can be represented as follows:

$$\Phi = \begin{bmatrix} \phi_{1,1} & \phi_{1,2} & \cdots & \phi_{1,n} \\ \phi_{2,1} & \phi_{2,2} & \cdots & \phi_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{m,1} & \phi_{m,2} & \cdots & \phi_{m,n} \end{bmatrix} \tag{4}$$

where $\phi_{i,j}$ denotes the total occurrences of keyword j for ith paper.

Normalization can be done by subtracting each element in Eq. 4 with its sample mean and dividing the result with the sample standard deviation. Mathematically, this can be achieved as follows:

$$\mathbf{A} = \frac{\Phi - \bar{\Phi}}{\sigma_{\Phi}} \tag{5}$$

where $\bar{\Phi}$ and σ_{Φ} are $m \times 1$ sample means and sample standard deviation, respectively. Note that the division here is element-wise division.

To find the PC coefficients, we just merely calculate the eigenfunctions of the sample covariance matrix. The sample covariance matrix is calculated based on the normalized matrix in Eq. 5, and we can get its variance by taking its diagonal elements. With this, the PC coefficients $\mathcal{P}$ can be obtained as follows:

$$\mathcal{P} = \mathbf{A} \times eig(cov(\mathbf{A})) \tag{6}$$

where $eig(cov(A))$ calculates the eigenfunctions of sample covariance.

4 Hot Topics Mining

This section presents the methodology of hot topics mining subject to the problem formulation discussed in Section 2. 8 hot topics are summarized using our hot topics mining method, the trends of those topics are analyzed. The classification performance of different classifiers are also compared in this section.

4.1 Eight Hot Topics

Text analysis is a major application field for machine learning algorithms. However the raw data, a sequence of symbols cannot be fed directly to the algorithms themselves as most of them expect

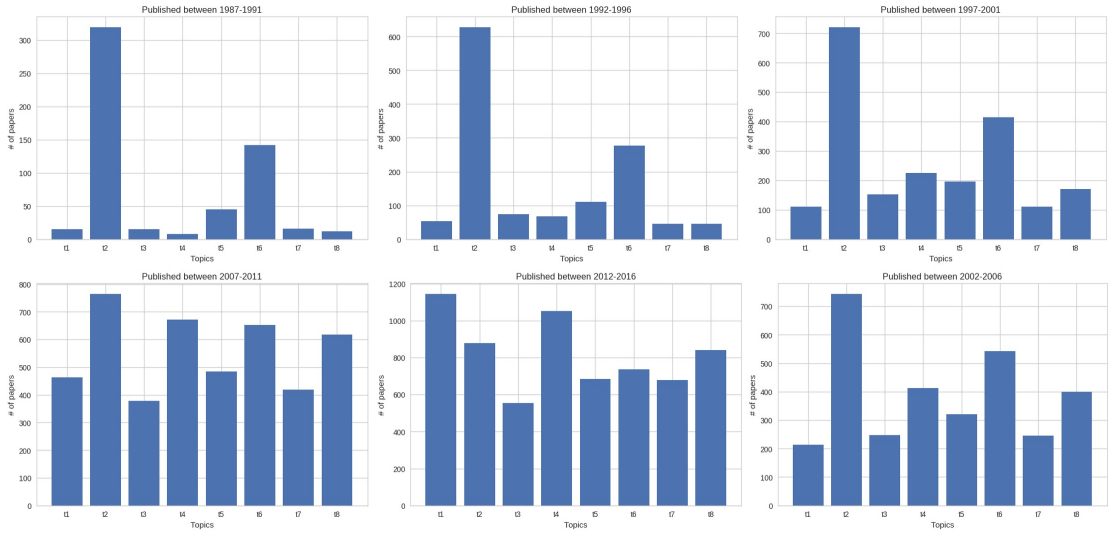

Figure 4: This figure illustrated the trend of eight hot topics at NIPS. We calculate the number of papers that being published every five years, this shows clearly that topic two is always popular in previous years, but in recent years, topic one becomes the most popular one.

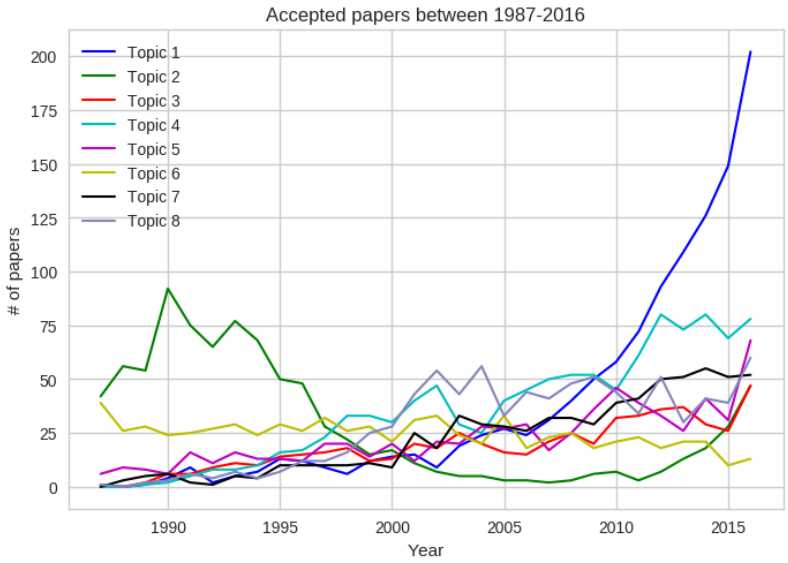

Figure 5: This figure illustrated the trend of the whole number of papers related to eight hot topics at NIPS. We added the number of papers that being published every years, this shows clearly that the number of papers related to topic one grows fast in recent five years.

numerical feature vectors with a fixed size rather than the raw text documents with variable length. In order to address this, TF-IDF technique is used to convert a collection of paper text of NIPS to a matrix of TF-IDF features. The we apply Non-negative matrix factorization (NMF) to the TF-IDF matrix. Unlike PCA, the representation of a vector is obtained in an additive fashion, by superimposing the components, without subtracting. Such additive models are efficient for representing text. We got eight hot topics (each topic contains ten key words) shown in Table 1 from previous papers of NIPS. In order to use a phrase to conclude each topic, we evaluate the top three papers' title under each topic, example is shown in Table 2. At the end, we summarized eight hot topics: *optimization algorithms*, *neural network application*, *reinforcement learning*, *bayesian methods*, *image recognition*, *artificial neuron design*, *graph theory* , *kernel methods*.

Table 3: Results of Naive Bayes

Topics	precision	recall	f1-score	support
Topic 1	0.80	0.83	0.82	351
Topic 2	0.71	0.40	0.51	75
Topic 3	1.00	0.03	0.05	73
Topic 4	0.30	1.00	0.46	147
Topic 5	1.00	0.19	0.32	99
Topic 6	1.00	0.43	0.61	23
Topic 7	1.00	0.01	0.02	103
Topic 8	0.95	0.37	0.54	99
Avg / Total	0.79	0.56	0.51	970

Table 4: Results of SVM

Topics	precision	recall	f1-score	support
Topic 1	0.91	0.98	0.94	351
Topic 2	0.91	0.92	0.91	75
Topic 3	0.92	0.90	0.91	73
Topic 4	0.92	0.90	0.91	147
Topic 5	0.93	0.96	0.95	99
Topic 6	0.84	0.91	0.87	23
Topic 7	0.97	0.86	0.91	103
Topic 8	0.95	0.78	0.86	99
Avg / Total	0.92	0.92	0.92	970

4.2 The Trend of Hot Topics

After got all the hot topics, we analyzed the trend of those hot topics by each five years (Fig. 4). The trend of whole number of papers related to each topic are also analyzed (Fig. 5). We can easily concluded that in previous years topic 2 (*neural network application*) is popular, but topic 1 (*optimization algorithms*) is more and more popular in recent five years. Researchers preferred to propose more optimization algorithms in recent five years. From this trend, in the future, papers related to topic 1, topic 2 and topic 4 will have a larger chance to be accepted.

4.3 Automatic Classification

In this section, we compare the performance between different classifiers (Naive Bayes, SVM and Random Forest). We use papers from 1987 to 2014 to train the classifiers and use papers from 2015 to 2016 to test the performance of the trained classifiers. The performance of Naive Bayes, SVM and Random Forest are shown in Table 3, Table 4 and Table 5. We can see that the performance of SVM is the best. The confusion matrix of three methods are shown in Fig. 6, Fig. 7 and Fig. 8.

5 Conclusion

This work proposed an unified methodology to do hot topic mining, automatic classification and based on the learned knowledge, it can predict whether one paper can be accepted. The dataset from NIPS is used for training and prediction purpose. Overall, total 22253 features are extracted from all the papers published from 1987 to 2016. Such high dimension features pose difficulty for further feature training and analysis. Furthermore, no all the features convey meaningful information regarding the main idea of the published paper. To combat the issue of high dimensionality, PCA is applied and about 92.47% dimension reduction is achieved to retain at least 95% of the information variance. We also found that PCA can boost the progress of hot topic mining and improve the performance of prediction.

Table 5: Results of Random Forest

Topics	precision	recall	f1-score	support
Topic 1	0.75	0.87	0.80	351
Topic 2	0.66	0.55	0.60	75
Topic 3	0.89	0.67	0.77	73
Topic 4	0.59	0.79	0.67	147
Topic 5	0.84	0.75	0.79	99
Topic 6	0.77	0.74	0.76	23
Topic 7	0.72	0.38	0.50	103
Topic 8	0.58	0.49	0.53	99
Avg / Total	0.72	0.71	0.70	970

Acknowledgments

We would like to thanks Prof. Yuan Yao for his earnestness in teaching.
Our contributions are as follows:
Mao Hui (20297805): Hot topic mining, automatic classification and prediction
Pc Ng (20305846): Feature extraction, cleaning and dimensionality reduction (PCA)

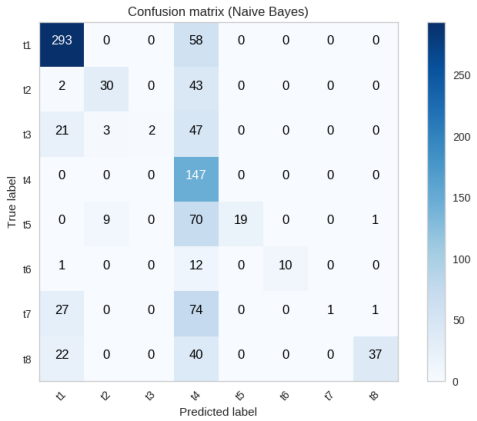

Figure 6: The confusion matrix of Naive Bayes

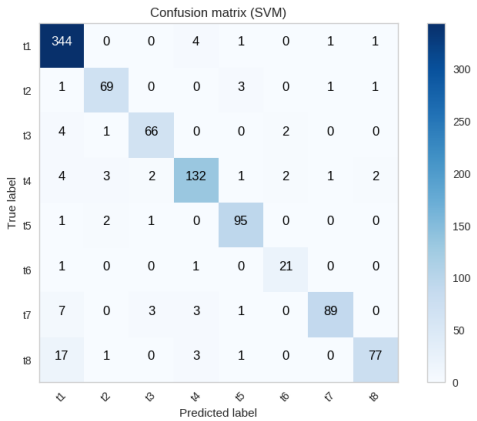

Figure 7: The confusion matrix of SVM

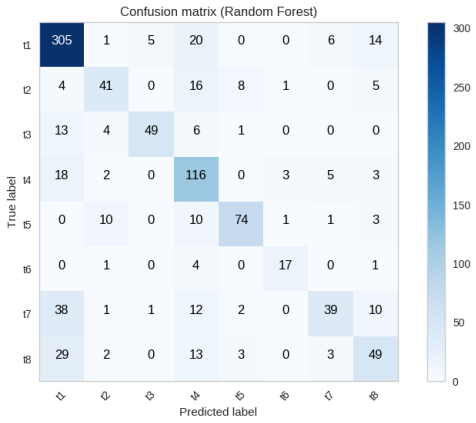

Figure 8: The confusion matrix of Random Forest

References

[1] Kazunari Sugiyama and Min-Yen Kan, "Towards higher relevance and serendipity in scholarly paper recommendation by kazunari sugiyama and min-yen kan with martin vesely as coordinator," *ACM SIGWEB Newsletter*, , no. Winter, pp. 4, 2015.

[2] Petra Isenberg, Florian Heimerl, Steffen Koch, Tobias Isenberg, Panpan Xu, Charles Stolper, Michael Sedlmair, Jian Chen, Torsten Moller, and John T Stasko, "vispubdata. org: A metadata collection about ieee visualization (vis) publications," *IEEE Transactions on Visualization and Computer Graphics*, 2016.

[3] Valerio Perrone, Paul A Jenkins, Dario Spano, and Yee Whye Teh, "Poisson random fields for dynamic feature models," *arXiv preprint arXiv:1611.07460*, 2016.

[4] Ian Jolliffe, *Principal component analysis*, Wiley Online Library, 2002.