Improving the Performance of sLDA with SVI, Feature Engineering and Active Learning

Duanfeng Gao
duga8843

You Lu
yolu1055

Shuo Zhang
shzh3550

Jing Guo
jigu7566

ABSTRACT

The goal of our project is to improve the performance of sLDA. In order for better running performance, we introduced stochastic variational inference(SVI) to optimize the sLDA algorithm. To improve the prediction accuracy, we applied feature engineering into our model. Then in order to deal with the difficulty of labelling the documents, we built an active learning framework of sLDA.

Keywords

sLDA, SVI, Active Learning

1. INTRODUCTION

Nowadays, topic modeling is one of the hotspots in the area of Natural Language Processing. LDA[2] is acknowledged as the foundational topic model and sLDA[1, 7] is a supervised extension of LDA.

In sLDA, we add to LDA a response variable associated with each document, so that the model can uncover the latent structure of a dataset as well as retains the predictive power for supervised tasks. For regression, the response variable is drawn from a Gaussian distribution. For classification, it is drawn from a softmax distribution.

In this project, we investigated three methods that can improve the performance of classification sLDA. Firstly, we applied a novel training method, that is, stochastic variational inference (SVI)[3], to train sLDA. Secondly, we employed feature engineering methods to improve the model prediction accuracy. Thirdly, we introduced an active learning framework to reduce the need for the costly labelling of documents.

This report is organized as following. Section 2 is a brief introduction on sLDA and SVI for online sLDA. Section 3 is about the active learning framework. Experiments and results are introduced in Section 4 and the last section is the conclusion.

2. SLDA

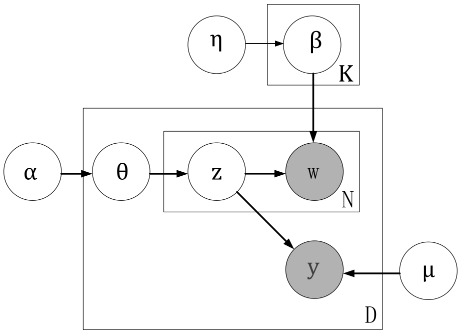


Figure 1: sLDA

With the dramatic development of the Internet, there is a growing need to analyze electronic texts. Most topic models, including LDA, are unsupervised models. Supervised latent Dirichlet allocation is the supervised version of LDA, and is becoming a popular method to apply supervised classification in the analysis of documents.

The generative process of sLDA[1, 7] is as follows:

1. Draw topic $\beta_k \sim Dirichlet(\eta, ..., \eta)$ for $k \in \{1, ..., K\}$
2. For each document $d \in \{1, ..., D\}$:
   (a) Draw topic properties $\theta \sim Dirichlet(\alpha, ..., \alpha)$
   (b) For each word $w \in \{1, ..., N\}$:
      i. Draw topic assignment $z_{dn} \sim Multinomial(\theta_d)$
      ii. Draw word $w_{dn} \sim Multinomial(\beta_{dn})$
   (c) $c \mid z_d, \mu \sim softmax(\bar{z}, \mu)$ where $\bar{z} = \frac{1}{N}\sum_{n=1}^{N} z_n$ and softmax provides the following distribution:

$$P(c \mid z, \mu) = \frac{e^{\mu_c^T \bar{z}}}{\sum_{l=1}^{C} e^{\mu_l^T \bar{z}}}$$

α and β are Dirichlet hyper-parameters. K is the number of topics, D is the total number of documents. C is the total number of classes. z_n is the topic assignment of word w_n. Figure 1 illustrates sLDA as a graphical model.

Typically, Gibbs sampling or batch variational inference method are applied to train topic models. These methods are not very efficient because they need to analyze all the documents in the corpus in order to update the global parameters at each iteration. To train classification sLDA by using these two methods is especially harder and slower,

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

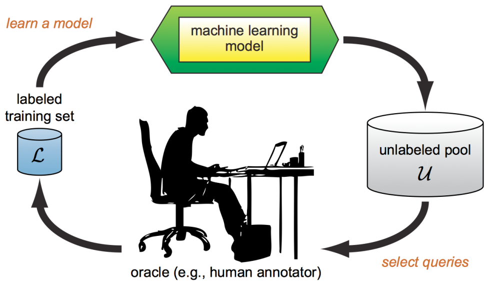


Figure 2: Pool-based Active Learning Cycle

because the response variable is nonlinear over the topic assignment, and the softmax distributionâĂŹs parameters, and the normalization factor strongly couples the topic assignment of each document [9]. To address this problem, we apply a novel training method, that is, stochastic variational inference (SVI) [3] to train sLDA. SVI for sLDA (online sLDA) is summarized in Algorithm 1.

As shown in Algorithm 1, SVI only needs to analyze a subset (mini-batch) of documents at each iteration. Therefore, SVI becomes much faster than the traditional training methods.

Algorithm 1 SVI for sLDA

```
Initialize $\lambda^{(0)}$
Set the learning rate for $\lambda$, i.e., $\rho_\lambda$ and the learning rate for , i.e., appropriately.
while unconvergent do
    Sample a subset of documents $w_d$ from the dataset.
    for Each document $w_d$ in the subset do
        Initialize $\gamma_d$ appropriately
        while unconvergent do
            $\phi_{vk} \propto exp\{E(log\theta_k + E(log\beta_{kv} + \frac{1}{N}\mu_{ck} - (h_v^T \phi_v^{old})^{-1} h_{vk}\}$
            $\gamma_k = \alpha + \sum_{v=1}^{V} \phi_{kv}$
        end while
    end for
    $g(\lambda_{kv}) = -\lambda_{kv}^{(t)} + \eta + \frac{D}{M} \sum_{d=1}^{M} \sum_{v=1}^{V} \phi_{dvk} w_{dv}$
    $g(\mu_{ck}) = \sum_{d=1}^{M} \bar{\phi}_d 1_{c_d=c} - \sum_{d=1}^{M} (\xi_d^{-1} \prod_{n=1}^{N_d} (\sum_{j=1}^{K} \phi_{dnj} exp(\frac{1}{N_d}\mu_{cj}))) \times \sum_{n=1}^{N_d} \frac{\frac{1}{N_d}\phi_{dnk} exp(\frac{1}{N_d}\mu_{ck})}{\sum_{j=1}^{K} \phi_{dnj} exp(\frac{1}{N_d}\mu_{cj})}$
    $\lambda^{t+1} = \lambda^t + \rho_\lambda^{(t)} g(\lambda^{(t)})$
    $\mu^{t+1} = \mu^t + \rho_\mu^{(t)} g(\mu^{(t)})$
end while
```

3. ACTIVE LEARNING

In practice, especially in our case, we can get the documents easily, but understanding the documents and labelling them is very difficult and costly, and this would make traditional supervised learning meaningless. So we introduced active learning, whose key hypothesis is that If the learning algorithm is allowed to choose the data from which it learns, it might perform better with less labelled data.

3.1 Pool-based Active Learning

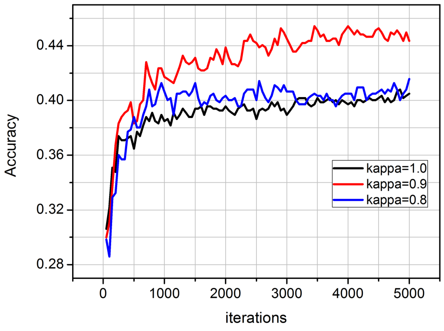


Figure 3: Selection of parameter κ

Figure 2 illustrates a pool-based active learning cycle[4, 5]. Beginning with a small labelled sample set, the machine learning model, which is the sLDA in our case, keeps selecting a certain number of unlabelled documents from the unlabelled dataset pool for the annotator, which would always be human beings, to label. The most important part in the cycle is how to select of the unlabelled documents, which is also called the query strategy.

3.2 Query Strategy

The simplest and most commonly used query strategy is uncertainty sampling[4]. In this strategy, samples that are the least certain how to label are selected. And in most cases, entropy[6] is an excellent tool to measure the uncertainty.

$$x^* = \arg\max_x - \sum_i P(y_i \mid x) \log P(y_i \mid x)$$

4. EXPERIMENTS AND RESULTS

4.1 Dataset

We used US Congressional Bills corpus[8] as our dataset. Each bill text has a label which is the congressional issue discussed. The corpus has 6,528 documents and a vocabulary of 21,007 words. The documents are classified into 19 classes. We randomly select 10 percent of the documents from the corpus as the test set.

4.2 Experiments on SVI

We explored how the SVI parameters, i.e., the batch size M, and the learning rate parameters, impact the performance of sLDA. We first set $M = 10$, $\tau = 10$, and evaluated different κ over the set $\{0.8, 0.9, 1.0\}$. Then we set $M = 10$, $\kappa = 0.9$, and evaluated different τ over the set $\{5, 10, 15, 20\}$. We found that when $\tau = 10$ and $\kappa = 0.9$, sLDA performed the best. We finally fixed $\tau = 10$ and $\kappa = 0.9$ and evaluated different M over the set $\{10, 20, 50\}$. We found that the large size performs slightly better than small size. However, if we enlarge the batch size, the per-iteration training time would also increase. Hence, we believe that training with large batch is not worthy. Figure 3, Figure 4 and Figure 5 illustrates the experimental results.

Furthermore, we compared SVI to Gibbs sampling for sLDA. The results are shown in Figure 6 and Figure 7. SVI obtains a higher accuracy. Besides, SVI is significantly faster than Gibbs sampling.

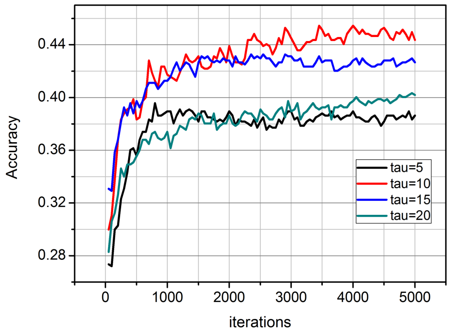


Figure 4: Selection of parameter τ

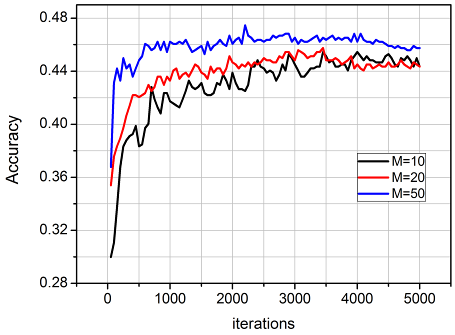


Figure 5: Selection of parameter M

4.3 Experiments on Active Learning

We conducted active learning experiment using uncertainty query strategy. We measured the uncertainty of every document by Entropy. Our startup training data set contains 310 labelled documents, and our unlabeled document pool contains 5284 documents. In each iteration, we query 50 documents from the pool to label. Figure 8 shows the performance of active learning. The more documents lablelled, the higher performance. And in the beginning ,there is a rapid growth of the performance, and this is the effect of active learning.

4.4 Experiments on Feature Engineering

In order to capture the contextual information more precisely, we implement three typical feature engineering methods: stemming, bigrams and trigrams. This results in a much larger feature space, we selected the most important 20000 features. We found that by adding these features into our model, not only the accuracy is increased, but also the tendency by iteration is more stable . This may due to the fact that we selected more representative features than the original dataset. Results of this part are shown in Figure 9.

5. CONCLUSION

Our project is basically about improving the performance of sLDA. Experiments have shown that the three approaches we introduced have improved sLDA in different aspects. SVI improves the efficiency of training sLDA model. Feature engineering improves the accuracy performance, and active learning deals with actual situation and makes up a framework for reducing the labelling cost.
All of our codes are open sourced at this address.

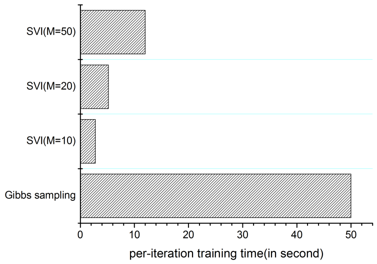
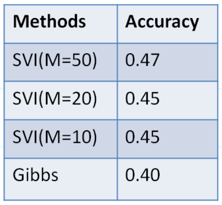


Figure 6: Comparison of time complexity



Methods	Accuracy
SVI(M=50)	0.47
SVI(M=20)	0.45
SVI(M=10)	0.45
Gibbs	0.40

Figure 7: Comparison of accuracy performance

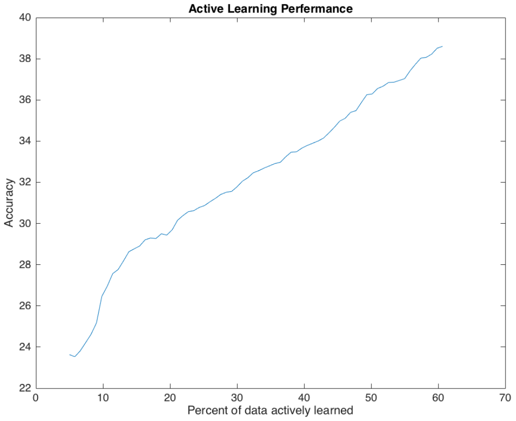


Figure 8: Active Learning Performance

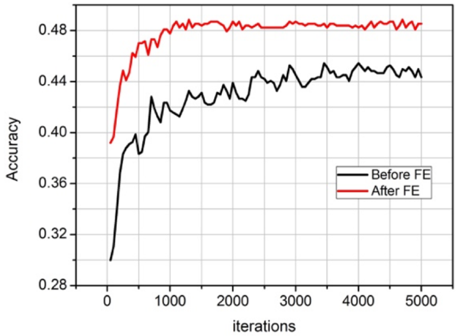


Figure 9: Comparison of accuracy performance of sLDA with and without Feature Engineering

6. REFERENCES

[1] D. M. Blei and J. D. Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35, 2007.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[3] M. D. Hoffman and D. M. Blei. Structured stochastic variational inference. *arXiv preprint arXiv:1404.4114*, 2014.

[4] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12. Springer-Verlag New York, Inc., 1994.

[5] B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[6] C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[7] C. Wang, D. Blei, and F.-F. Li. Simultaneous image classification and annotation. *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, pages 1903–1910, 2009.

[8] T. Yano, N. A. Smith, and J. D. Wilkerson. Textual predictors of bill survival in congressional committees. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 793–802. Association for Computational Linguistics, 2012.

[9] J. Zhu, A. Ahmed, and E. P. Xing. Medlda: maximum margin supervised topic models. *the Journal of machine Learning research*, 13(1):2237–2278, 2012.