Use of Semantics in Topic Based Classification

Madhav Ram Nimishakavi and M. Narasimha Murty

Computer Science and Automation,
Indian Institute of Science, Bangalore, India
{madhav,mnm}@csa.iisc.ernet.in

Abstract. Topic models, like Latent Dirichlet Allocation, discover underlying low-dimensional topic spaces from huge data collections. The lower dimensional representation can be used for classification. Topics learnt from standard topic models do not consider correlations among words, which are useful for making the topics more sensible. Moreover, standard models are bag of words models. We propose an approach where the correlations among words are used for learning topics and improve the topic-word probabilities of rare yet important words. We show through classification results that our models learn a better topic structure compared to standard models, specifically LDA and DiscLDA. Many classification models ignore words which are not part of the vocabulary, we propose an approach to use such words for better classification.

1 Introduction

Topic models assume that a hidden topic structure is responsible for generating any document collection . For example, Latent Dirichlet Allocation (LDA) [2] models each document as a mixture of topics and each word is associated with one or more topics, however it is a bag of words model. Since the modeling of topics is based on frequency of occurrence of words, less frequent words have poor estimates within topics. Some words may be less frequent but are very relevant to some of the topics, such words should have proper estimates within the topics.

Topics obtained from LDA are not always refined [8], furthermore the quality of topics decreases with increase in number of topics [7]. Coherence and interpretability of learned topics can be improved by using *regularized* learning of topic models [8]. We propose an approach to improve the topic estimates of words which are less frequent but are relevant to some topics by using a variant of the regularizer described by [8], our regularizer uses the correlations among words for creating a structured prior on topic-word probabilities. We show that the topics learned by our implementation of LDA are better than those of standard LDA by classifying documents in topic space. DiscLDA(Discriminative Learning for Dimensionality Reduction and Classification) [5] is a discriminative variation on LDA. We introduce a structured prior on topic-word probabilities of DiscLDA, and show that this regularized version of DiscLDA performs better than the standard DiscLDA in classification.

Generally classification models do not consider the words in test documents, which are not part of the vocabulary, for classification. In some cases the words in test documents which are not part of the vocabulary convey more information about the document than those words which are in vocabulary. We propose an approach, base on topics, which takes those words into consideration while classifying the documents, thus improving the results.

2 Background

2.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation [2] is a generative probabilistic model for discrete data collections. Each document in the corpus is modeled as a mixture of finite number of topics and each word in vocabulary is associated with one or more of these topics. LDA tries to find out the set of topics in the document collection, where each topic is considered to be a distribution of valid words and each document is a distribution of topics.

Each document is modeled as a multinomial distribution (θ) over the topics and each topic is modeled as a multinomial distribution (ϕ) over words. The parameters θ and ϕ should be estimated in order to find the above distributions. A symmetric Dirichlet prior (α) is placed on θ in LDA making it a generative model. The prior parameter α determines the smoothing of the topic distribution. A symmetric Dirichlet prior (β) is also placed on ϕ which determines smoothing of word distribution in every topic. The choice of hyper-parameters α and β will depend on the number of topics and vocabulary size. [9] recommends the values of $\alpha = 50/T$ and $\beta = 0.01$ for better results, where T is the number of topics.

2.2 DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification

DiscLDA [5] is a discriminative variation on LDA in which a class dependent linear transformation is introduced on the topic proportions. This parameter is estimated by maximizing the conditional likelihood. By using the transformed topic mixture proportions as a new representation of documents, a supervised dimensionality reduction algorithm is obtained, which uncovers the latent structure in a document collection while preserving predictive power for the task of classification. DiscLDA is a supervised form of LDA. Side information such as class labels are incorporated into LDA, while retaining its favourable unsupervised dimensionality reduction abilities. The goal is to develop parameter estimation methods that yield LDA topics that characterize the corpus and maximally exploit the predictive power of the side information.

2.3 Support Vector Machines

A Support Vector Machine is a supervised learning algorithm for binary classification and regression analysis [10]. The main computational problem underlying the SVM methodology is optimizing a quadratic cost function with linear constraints. SVM is based on Structural Risk Minimization (SRM) [3]. Given a training set of examples, each marked as belonging to one of the classes, SVM builds a model by representing each example as a point in real space based on which the classification is done. SVM constructs a separating hyperplane and two parallel hyperplanes on each side of it. The aim is to maximize distance between the two parallel hyperplanes, since in general the larger the margin the lower the generalization error of the classifier [10].

3 Related Work

There has been some work to use domain knowledge from external sources in topic modeling. For example [1] incorporates domain knowledge into topic modeling using Dirichlet Forest Priors. Their method is based on replacing the symmetric Dirichlet prior over topic-word probabilities, but their method differs from ours in choosing the correlations between words. [8] replaces the symmetric Dirichlet prior over topic-word probabilities, but our method differs from their's in the way the covariance matrix is constructed. They focus on improving the quality of topic models in noisy environment but our method focuses on improving the topic-word probabilities of rare yet important words. Use of asymmetric priors is investigated by [11], they show that use of an asymmetric prior over document-topic distributions has more advantages than a symmetric prior.

4 Regularization

Regularization in topic models involves creating structured prior over words such that it reflects the association between them. As described earlier, LDA is a generative process, which is given by:

$$\theta_{t|d} \sim Dir(\alpha) \qquad \phi_{w|t} \sim Dir(\beta)$$

$$z_{id} \sim Multi(\theta_{t|d}) \qquad x_{id} \sim Multi(\phi_{w|z_{id}}).$$

The following Gibbs sampling [6] update for the posterior distribution of topic assignments is obtained by marginalizing out θ and ϕ from the joint probability.

$$p(z_{id} = t|x_{id} = w, z^{-i}) \propto \frac{N_{wt}^{-i}+\beta}{N_{t}^{-i}+W\beta}(N_{td}^{-i} + \alpha).$$

where z^{-i} denotes the set of topic assignment variables except the i^{th} variable, N_{wt} is the number of times word w has been assigned to to topic t, N_{td} is the number of times topic t has been assigned to document d and $N_t = \sum_{w=1}^{W} N_{wt}$.

We introduce structured prior on ϕ_t (vector of word probabilities for a given topic t), which has a regularization effect on LDA. Prior on ϕ makes use of a $W \times W$ covariance matrix C. The entries of the covariance matrix are similarity values among words. Difference between the structured prior of [8] and our structured prior is the way in which covariance matrix C is built. We consider those words which are relatively frequent in the external data, but [8] takes only those words into consideration which are highly frequent in the dataset. $C[i][j]$ value of the covariance matrix is the similarity measure between i^{th} and j^{th} word, which is defined as

$$sim(w_i, w_j) = log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}$$

where $P(w_i, w_j)$ is fraction of the external documents in which both the words w_i and w_j appear and $P(w_n)$ is the fraction of the external documents in which the word w_n appears.

4.1 Quadratic Regularizer

We use a standard quadratic form with a trade-off factor. For a given matrix C, the prior will be :

$$p(\phi_t|C) \propto (\phi_t^T C \phi_t)^v$$

for some power v. The normalization constant is not needed for MAP estimation, the log posterior is given by:

$$\mathcal{L}_{MAP} = \sum_{i=1}^{W} N_{it}\ log\phi_{i|t} + v\ log(\phi_t^T C \phi_t)$$

Optimizing the above equation with respect to $\phi_{w|t}$ subject to constraints $\sum_{i=1}^{W} \phi_{i|t} = 1$, the following fixed point update for $\phi_{w|t}$ is obtained:

$$\phi_{w|t} \leftarrow \frac{1}{N_t + 2v}(N_{wt} + 2v\frac{\phi_{w|t} \sum_{i=1}^{W} C_{iw}\phi_{i|t}}{\phi_t^T C \phi_t})$$

Therefore, now the update equation for the posterior distribution of topic assignments is :

$$p(z_{id} = t|x_{id} = w, z^{-i}, \phi_{w|t})\ \propto\ \phi_{w|t}(N_{td}^{-i} + \alpha)$$

5 Experiments and Results

We present results on datasets obtained from Wikipedia documents through JWPL [12]. Specifically we deal with documents related to the field *Computer Science*. Four different datasets are obtained from these documents for the purpose of binary classification, they are *Computer Engineering & Software Engineering, Computer Scientists & Theoretical Computer Science, Computational Science & Computer Graphics and Software Engineering & Theoretical Computer Science* . We use the entire collection of Wikipedia documents related to the field *Computer Science* as the *External Data* [8] for our purposes.

5.1 Topic Based Classification

We did classification in the topic space for evaluating the model. Topics are obtained from applying Regularized LDA on the documents in train set, each document is represented as a point in the topic space. Since the test documents are not used for obtaining the topics, the following method is used for representing them in the topic space.

Initially, each test document d_{test} is represented as an N-dimensional point $d_{test} = (|w_1|, |w_2|, ..., |w_N|)$, where $|w_i|$ is the frequency of the i^{th} word of the vocabulary $V = \{w_1, w_2, ..., w_N\}$. Note that vocabulary is the set of highly frequent N words in the *External Data* and also appear in the documents of the dataset. Let $D^{test}_{1\times N}$ be the matrix representation of the test document d_{test} and $\mathcal{D}^{test}_{1\times|T|}$ be the representation of test document in topic space, it is obtained as follows:

$$\mathcal{D}^{test}_{1\times|T|} = \Upsilon_{|T|\times N}\ (D^{test}_{1\times N})^T$$

where $\Upsilon_{|T|\times N}$ is the topic-word matrix obtained from LDA.

We used SVM in the topic space for classification, specifically we used *libsvm* [4]. Classification accuracies in Table 1, show that Regularized version of LDA outperformed LDA in terms of classification.

Classification accuracies have varied with change in the v parameter used in the structured prior. Variation of the accuracies for the datasets *Computer Engineering & Software Engineering*, *Computer Scientists & Theoretical Computer Science*, *Computational Science & Computer Graphics* and *Software Engineering & Theoretical Computer Science* is shown in Figures 1, 2, 3 and 4 respectively. On the X-axis is the value of v varying between 0 and 1 and on the Y-axis is the value of the classification accuracy of *RegLDA+SVM*. This behaviour is dataset dependent

Table 1. Classification results

Dataset	**LDA+SVM**	**RegLDA+SVM**
Computer Engineering & Software Engineering	62.9186%	74.5624%
Computer Scientists & Theoretical Computer Science	66.8277%	71.6586%
Computational Science & Computer Graphics	64.1529%	75%
Software Engineering & Theoretical Computer Science	53.9718%	80.8949%

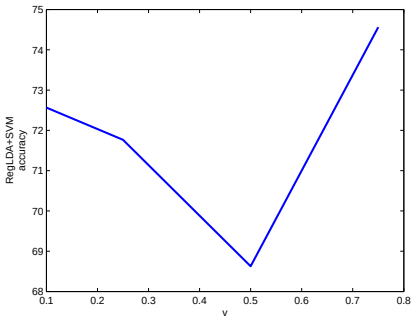

Fig. 1. RegLDA+SVM accuracy variation with v for Computer Engineering & Software Engineering

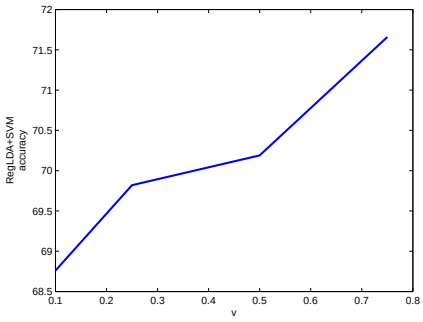

Fig. 2. RegLDA+SVM accuracy variation with v for Computer Scientists & Theoretical Computer Science

5.2 Classification using DiscLDA

We have applied the structured prior on the topic-word probabilities of DiscLDA. Table 2 shows that the regularized DiscLDA performs better than the standard DiscLDA in classification.

6 A Method to Use Important Non-vocabulary words in Test Documents for Better Classification

Generally any classification algorithm has a fixed set of features, using which documents or patterns are represented. Whenever a class label has to be assigned to a new document or a pattern, the fixed set of features are only used. However, the assumption that the training set provides sufficient information for classification is not always valid. For instance, consider a binary classification scenario where the classes are *Politics* and *Cricket*. If a new document containing information about the controversies in *Indian Premier League 5* should be classified as one

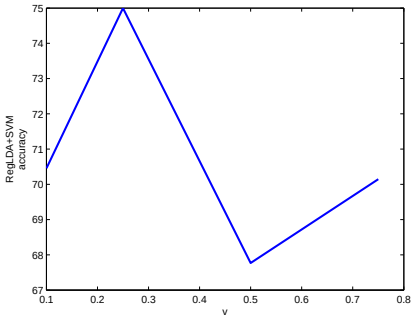

Fig. 3. RegLDA+SVM accuracy variation with v for Computational Science & Computer Graphics

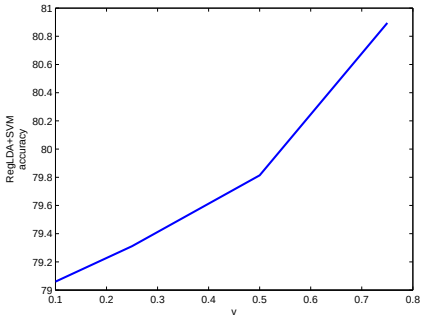

Fig. 4. RegLDA+SVM accuracy variation with v for Software Engineering & Theoretical Computer Science

Table 2. DiscLDA Results

Dataset	**DiscLDA**	**RegDiscLDA**
Computer Engineering & Software Engineering	90.44%	91.55%
Computer Scientists & Theoretical Computer Science	90.18%	91.42%
Computational Science & Computer Graphics	90.12%	92.43%
Software Engineering & Theoretical Computer Science	82.89%	83.81%

belonging to the class *Cricket*, since it is about controversies it is likely to be classified as a document belonging to *politics*. In such situation the frequency of the words in test document which belong to vocabulary, that are more relevant to the class *Cricket*, should be high for the document to be classified as a *Cricket* document.

If a word is highly frequent with in a document and is not a stop word, then it is highly likely that the word is important for the document. The method we propose makes use of such words in test documents for improving topic based classification. For every such word in a test document, we increment the count of each word which occurs along with it in the same sentence and belongs to vocabulary, therefore the test document $d_{test} = (|w_1|, |w_2|, ..., |w_N|)$ becomes $d^{'}_{test} = (|w_1|', |w_2|', ..., |w_N|')$ where $|w_i|'$ is the modified count of i^{th} word of vocabulary. Now d'_{test} is represented in topic space by following the method described in section 4.1.

Table 3. Change in classification accuracies from considering non-vocabulary words in test documents

Dataset	**RegLDA+SVM ignoring non-vocabulary words in test documents**	**RegLDA+SVM considering non-vocabulary words in test documents**
Computer Engineering & Software Engineering	74.5624%	82.8957%
Computer Scientists & Theoretical Computer Science	71.6586%	83.9775%
Computational Science & Computer Graphics	75%	81.5083%
Software Engineering & Theoretical Computer Science	80.8494%	77.9286%

Table 3 shows that the classification accuracies have increased for all the datasets except for *Software Engineering & Theoretical Computer Science*, this is due to the class imbalance of the dataset.

7 Conclusion

In this paper we showed that the regularized version of LDA brings out better topics and also produces better results for classification when compared to the standard LDA. We also showed that DiscLDA is improved with a structured asymmetric prior over the topic-word probabilities. We addressed the problem of dealing with words not belonging to vocabulary in test documents and we gave a method for this problem. We showed that our method performs better in topic based classification.

Acknowledgment

We would like to thank Dr. Indrajit Bhattacharya, Assistant Professor, Computer Science and Automation, Indian Institute of Science for his help.

References

1. D. Andrzejewski, X. Zhu, , and M. Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *ICML*, 2009.
2. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
3. Christopher J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(121-167), 1998.

4. Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.
5. Simon Lacoste-Julien, Fei Sha, and Michael I. Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. In *Proceedings of Neural Information Processing Systems*, 2008.
6. Jun S. Liu. The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(958-966), 1994.
7. David Mimmo, Hanna M. Wallach, Edmund Tally Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Empirical Methods in Natural Language Processing*, 2011.
8. David Newman, Edwin V. Bonilla, and Wray Buntime. Improving topic coherence with regularized topic models. In *Neural Information Processing Systems*, 2011.
9. Mark Steyvers and Tom Grifiths. *In Latent Semantic Analysis: A Road to Meaning*, chapter Probabilistic Topic Models. 2006.
10. V. Vapnik. *The Nature of Statistical Learning Theory*. NY: Springer-Verlag, 1995.
11. Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In *Neural Information Processing Systems(NIPS)*, 2009.
12. Torsten Zesch, Christof Müller, and Iryna Gurevych. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, 2008.