A Novel Term Selection Approach in sLDA for Imbalanced Text Categorization

Zhenyan Liu[1,2, a], Dan Meng [2,b], Weiping Wang[2,c] , Yong Wang[1,d] , Chenhao Bai[1,e]
1School of Software, Beijing Institute of Technology, BeiJing, 100081, China
2Institute of Information Engineering , Chinese Academy of Sciences, BeiJing, 100093, China
[a]zhenyanliu@bit.edu.cn, [b]mengdan@iie.ac.cn, [c]wangweiping@iie.ac.cn, [d] wangyong@bit.edu.cn, [e]baichenhao@bit.edu.cn

Abstract.

The supervised Latent Dirichlet Allocation (sLDA) is a probabilistic topic model of labelled documents, which is better than unsupervised LDA for text categorization. But sLDA experiments were based upon this default assumtion that the corpus is balanced, that is, the samples of each class are approximately equal, and chose a vocabulary by *tf-idf*. While the corpus is imbalanced, *tf-idf* tends to choose terms from the majority classes and ignore terms of the minority ones. Thus the performance of text classifier will be degraded severely. Therefore this paper proposed a new term selection approach which can fairly choose more discriminative terms from every category. Experimental results show that using this new approach in sLDA for imbalanced text categorization can greatly impove the recall and precision of the minority classes, and it is superior to *tf-idf*.

Keywords: sLDA, imbalanced dataset, text categorization, topic model.

Introduction

Probabilistic topic model are receiving extensive attention in text mining, natural language processing, information retrieval and so on. Latent Dirichlet Allocation (LDA) [1] is one of the most popular probabilistic topic models, which is utilized to automationally extract latent topics and to represent documents in a semantic

topic space.But LDA is unsupervised: only the words in the documents are modelled and text categories are not efficiently made use of. Furthermore, Blei et al [2] proposed supervised Latent Dirichlet Allocation (sLDA), a probabilistic topic model of labelled documents. And their research showed that sLDA is better than unsupervised LDA for text categorization.

In sLDA experiments, the vocabulary was chosen by *tf-idf* [3]. The *tf-idf* value of a “term” or “word” is computed by using the product of the term frequency (*tf*) and the inverse document frequency (*idf*). The *tf-idf* schema typically starts with a default assumtion that the corpus is balanced, that is, the samples of each class are approximately equal. But many real-world datasets are imbalanced, in which there are many more instances of some classes than others. In such cases, *tf-idf* tends to choose terms from the majority classes and ignore terms of the minority ones. Therefore the classifier will be overwhelmed by the majority classes and ignore the minor ones.

To address this shortcoming, this paper will propose a new term selection approach which can equally choose more discriminative terms from every category.The rest of the paper is organized as follows. In Section 2, we introduce sLDA. To complish imbalanced text categorization, a new term selection approach in sLDA is proposed in Section 3. In Section 4, we present our experiments of this new approach compared to *tf-idf*. Finally, conclusions are addressed in section 5.

sLDA

LDA is a generative probabilistic topic model. The basic idea of LDA is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. In LDA, the topic proportions for a docu-ment are drawn from a Dirichlet distribution. The words in the document are obtained by repeatedly choosing a topic assignment from those proportions, then drawing a word from the corresponding topic [2].

For text classification LDA topics are useful, since they act to reduce data dimension. But LDA draws latent topics without regard to text categorization. So an extended LDA model named sLDA is constructed by adding to LDA a response variable associated with each document. Note that this response variable is the category of a document for text classification. The graphical model of sLDA is depicted in Fig.1.

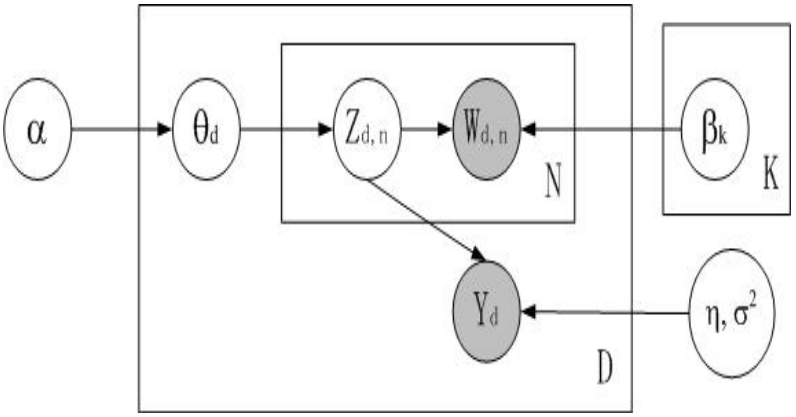

Figure1. A graphical model representation of sLDA

In Fig.1, θ_d refers to a document-topic distribution for each document *d* in a corpus *D*, α is the Dirichlet parameter of θ, $W_{d,n}$ is the *n*th of observed word of document *d* which contains *N* words, and $Z_{d,n}$ is the corresponding latent topic of word $W_{d,n}$, β_k refers to a topic-word distribution for each topic *k*,Y_d is the category of document *d*, η and σ^2 are the response parameters.

Under the sLDA model, each document and category arises from the following generative process [2]:

1. Draw topic proportion $\theta \mid \alpha \sim Dir(\alpha)$.
2. For each word
   a) Draw topic assignment $z_n \mid \theta \sim Mult(\alpha)$.
   b) Draw word $w_n \mid z_n, \beta_{1:K} \sim Mult(\beta z_n)$
3. Draw category variable $y \mid z_{1:N}, \eta, \sigma^2 \sim N(\eta^T, \bar{z}, \sigma^2)$

Note that α, $\beta_{1:K}$, η and σ^2 are treated as unknown constants to be estimated, rather than random variables.

A New Term Selection Approach in sLDA

In sLDA experiments, the vocabulary was chosen by *tf-idf*. However the term weights computed by *tf-idf* can only reflect the document difference, not the category difference. As a result, especially for an imbalanced corpus, most of terms chosen by *tf-idf* may be come from a majority class, which will tend to degrade the performance of classifier directly.

So we will propose a new term selection approach to address this shortcoming. The new appoach introduced a new factor — Information Gain (IG), to replace the *idf* factor of *tf-idf*. IG is a sophisticated measure of term relevance which takes into account the relations between terms and categories. IG measures the number of bits of information obtained for the prediction of categories by knowing the presence or absence in a document of a term[6].

The information gain of a term t is defined to be:

$$\mathrm{IG}(t) = -\sum_{i=1}^{m} P(c_i)\,\lg P(c_i) + P(t)\sum_{i=1}^{m} P(c_i|t)\,\lg P(c_i|t) + P(\bar{t})\sum_{i=1}^{m} P(c_i|\bar{t})\,\lg P(c_i|\bar{t}) \quad (1)$$

In eq.1, $c_1, \ldots, c_m$ denote the set of possible categories in the current collection. $P(c_i) = \frac{D_{c_i}}{D}$, here D denotes the total number of docments and D_{C_i}denotes the number of documents that belongs to class c_i. Moreover, $P(t) = \frac{D_t}{D}$, here D_t denotes the number of documents in which the term t occurs. $P(c_i|t) = \frac{D_{c_i} \cap D_t}{D_t}$, here $D_{c_i} \cap D_t$ denotes the number of documents from class c_i that have at least one occurence of term t. $P(\bar{t}) = \frac{D_{\bar{t}}}{D}$, here $D_{\bar{t}}$ denotes the number of documents in

which the term t doesn't occur. $P(c_i|\bar{t}) = \frac{D_{c_i} \cap D_{\bar{t}}}{D_{\bar{t}}}$, here $D_{c_i} \cap D_{\bar{t}}$ denotes the number of documents from class c_i that does not contain term t.

The information gain is computed for each term of the collection, and the terms whose information gain is less than some predetermined threshold are removed.

Experiments

In the experimental evaluation, we focused on a comparison between *tf-idf* and *tf-IG* in sLDA. We ran experiments on a subset of Reuters R8 dataset from Ana [7], which had been pre-processed that includes tokenization and stop word removal. The experiment dataset contained three categories: "earn", "trade", "grains"; 3,132 training documents and 1,168 test documents. There were 2,840 training documents and 1,083 test documents in "earn" category; 251 and 75 in "trade"; 41 and 10 in "grains".

With this imbalanced dataset sLDA model was trained. An 8000-term vocabulary of sLDA was chosen by *tf-idf* or *tf-IG*. The documents were represented in latent topic space drawn by sLDA. We built SVM classifier with LIBSVM development kit [8], in which linear kernel function is used.

Commonly the evaluation metrics for imbalanced text categorization are macro-averaged precision, macro-averaged recall, macro-averaged F1. Macro-averaged scores are averaged values over the number of categories. Let P be the pecision, R be recall, and *m* denotes the number of categroies, then macro-avgraged precision is $\frac{1}{m}\sum_{i=1}^{m} P_i$, macro-avgraged recall is $\frac{1}{m}\sum_{i=1}^{m} R_i$, macro-averaged F1 is $\frac{1}{m}\sum_{i=1}^{m} F1_i$, where *F*1 is $\frac{2PR}{P+R}$.

Table1 summarizes the results for *tf-idf* and *tf-IG* in sLDA on the Reuters R8 dataset. "Macro" stands for macro-averaged perfomance. From Table1 we can see the minority categories benefit most significantly. For example, the recall, precision and F1measure of "grain" are increased by more than 13%, 15% and 14%

respectively, and macro-averaged recall, precision and F1measure are increased by more than 7%, 7% and 8% respectively.

Table 1. Results for *tf-idf* and *tf-IG* in sLDA on Reuters R8

	tf-idf			tf-IG		
	R	P	F1	R	P	F1
earn	95.12%	94.49%	94.81%	96.25%	94.32%	95.27%
trade	79.03%	85.56%	82.17%	90.48%	93.86%	92.14%
grain	74.54%	78.29%	76.37%	88.19%	93.67%	90.85%
Macro	82.90%	86.11%	84.45%	91.64%	93.95%	92.75%

Fig.2 summarizes the comparison between *tf-idf* and *tf-IG* in sLDA in chart form. As can be seen from Fig.2, the use of *tf-IG* in sLDA can greatly improve the performance of imbalanced text classifier compared with *tf-idf*.

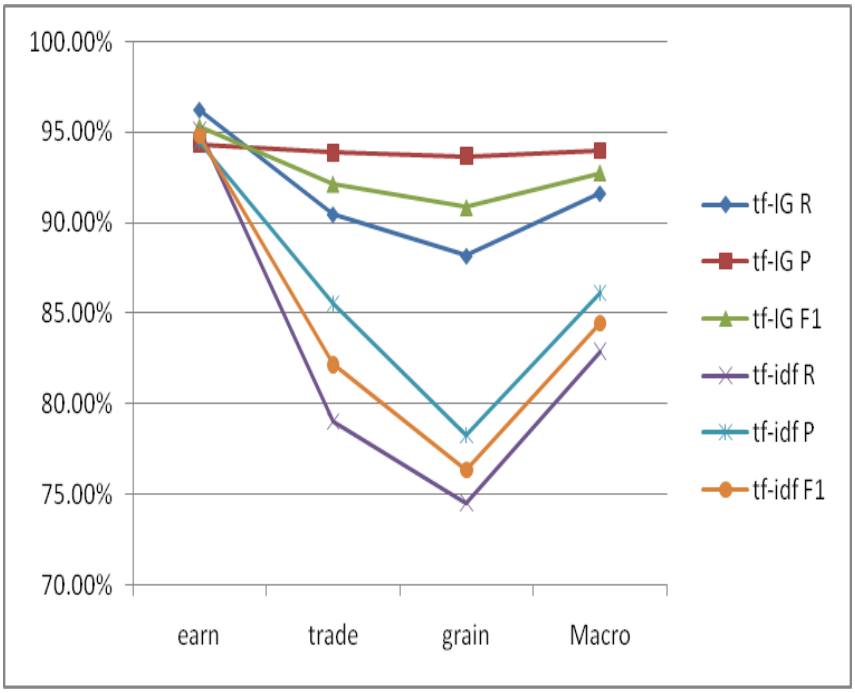

Figure 1. A comparison between *tf-idf* and *tf-IG*

Conclusions

The *tf-IG* is a superior term selection approach especially for imbalanced text categorization. The term weights computed by *tf-IG* can not only reflect the document diffrence but also the category difference, while *tf-idf* can only reflect the document diffrence. By *tf-IG* more discriminative terms fairly can be chosen from every category. Compared with *tf-idf* the vocabulary of sLDA chosen by *tf-IG* is more efficient for imbalanced text categorization.

Acknowledgements

This work was financially supported by National Natural Science Foundation of China (61272361), also supported by Key Project of National Defense Basic Research Program of China (B11201320), National HeGaoJi Key Project (2013ZX01039 -002-001-001), National High-Tech Research and Development Program of China (2012AA011002).

References

[1] D.Blei, A.Ng, M.Jordan. Latent Dirichlet Allocation. Machine Learning Research, 3(3), pp.993-1022 (2003).

[2] D.Blei, J.McAuliffe. Supervised topic models. Advances in Neural Infomation Processing Systems (2008).

[3] N.Chawla, N.Japkowicz, A.Kotcz. Editorial: Special issue on learning from imbalanced data sets. ACM SIGKDD Explorations Newsletter, 6(1), pp.1-6 (2004).

[4] G.Salton, C.Buckley. Term-Weighting Approaches in Automatic Text Retrieval.Journal of Information Processing & Management, 24(5), pp. 513–523 (1988).

[5] B.C.How, N.K. An Empirical Study of Feature Selection for Text Categorization based on Term weightage. The IEEE/WIC/ACM International Conference on Web Intelligence, pp. 599-602 (2004).

[6] K. Aas, L. Eikvil.Text categorisation: a survey. Technical Report, Norwegian Computing Center (1999).

[7] http://web.ist.utl.pt/~acardoso/datasets/.

[8] http://www.csie.ntu.edu.tw/~cjlin/.