Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 15 (2011) 1963 – 1967

Procedia Engineering

www.elsevier.com/locate/procedia

Advanced in Control Engineering and Information Science

Multi-class text categorization based on LDA and SVM

Kunlun Li[a*], Jing Xie[a], Xue Sun[b], Yinghui Ma[a], Hui Bai[c]

[a]*College of Electronic and Information Engineering, Hebei University, Baoding, 071002, China*
[b]*Industrial&Commercial College, Hebei University, Baoding, 071002, China*
[c]*Software School, Fudan University, Shanghai, 201203, China*

Abstract

When dealing with the high dimensions and large-scale multi-class textual data, it is commonly to ignore the semantic relation between words with the traditional feature selection method. In order to solve the problem, we introduce the categories information into the existing LDA model feature selection algorithm, and construct SVM multi-class classifier on the implicit topic-text matrix. Experimental results show that this method can improve classification accuracy and the dimensionality is reduced availably, the value of F1, Macro-F1, and Micro-F1 are obtained improvement.

Keywords: text categorization, feature selection, LDA model, SVM;

1. Introduction

Text categorization is a supervised learning task in which documents are assigned to categories based on the training. It is the research focus and core technology in the field of information retrieval and data mining since the amount of electronic text information has been rapidly expanded. Text representation and characteristic dimensionality reduction are important problems which should not be neglected problems to improve the overall classification performance. One of the common methods is Vector Space Model-VSM, It is a critical problem for most traditional classifier, and many texts cannot be classified correctly because of the problem of data sparseness caused by high dimensional characteristic. In recent years, the application of Latent Semantic Indexing-LSI and PLSI is obtained in-depth research. The above

* Kunlun Li. Tel.: +86-312-5079368;
E-mail address: likunlun@hbu.edu.cn, jinggurusi@163.com

1877-7058 © 2011 Published by Elsevier Ltd.
doi:10.1016/j.proeng.2011.08.366

methods have remarkable effect on dimensionality reduction, but the parameters space is proportional to the training data, and the effect of modeling is not obvious on the dynamic growth or large-scale corpus. Feature filtering (DF, Chi, MI etc.) and feature extraction (LSI, PCA etc.) are broadly used on textual data for the purpose of reducing the dimensionality [2]. The methods ignore synonymous, polysemous words and the semantic ties of words, they easily overlook the important feature of rare category and the classification performance suffers bad influence. Recently, A probability growth model called LDA (latent dirichlet allocation) [3] can be used to feature selection for the purpose of discovering the underlying semantic structures. It has more outstanding characteristic compared with the relatively simple latent variables models.

In this paper, we introduce LDA model into feature selection and integrate into the relations between feature words and categories and establish their generative model according to different types of textual data. Then the multi-class SVM classifier is trained based on the implicit topics–text matrix, which is obtained from the feature words probabilistic distributions of categories LDA model. It combines the outstanding feature dimensionality reduction and text representation capabilities of LDA with the powerful classification ability of SVM to improve the text classification performance.

2. Topics Modeling and Parameter Estimation

2.1 The LDA model

LDA (Latent Dirichlet Allocation) is a completely generating probabilistic model on discrete data set (such as documents). Assuming that there are K independent implicit topics in the text set which include M documents, each topic is the polynomial probabilistic distribution of words, and each document is randomly generated by the K implicit topics. The key problem of building and employing LDA model is the inference of implied variable, in others words, it is important to obtain the composition information (θ, z) of hidden topic. If given the Dirichlet parameter α and β, the simultaneous distribution of the random variables θ, z and w in document d is computed as:

$$P(\theta, z, w / \alpha, \beta) = p(\theta / \alpha) \prod_{i=1}^{N_w} p(z_i / \theta) p(w_i / z_i, \beta) \qquad (1)$$

As there are multiple connotative variables simultaneously, it is hard to compute the value directly, and the approximate solution is demanded [3]. There are several similar reasoning algorithms to acquire the estimated parameter values, such as Variational Bayes Inference, Gibbs sampling approach, Laplace approximation and Expectation-Propagation algorithm.

2.2 Gibbs Sampling

Markov Chain Monte Carlo (MCMC) provides the approximate iteration method of extracting sample value from the complex probabilistic distribution [7]. The Gibbs sampling of MCMC is a straightforward realization method, and it aims to structure the Markov chain which converges to a target probabilistic distribution and extracts the specimen with probabilistic approximation. Consequently, composing the objective probabilistic distribution function is the key point of using the Gibbs sampling algorithm. In order to obtain the probabilistic distribution of the word layer, we consider the posteriori probability $p(w / z)$ and take advantage of Gibbs sampling to gain the value of the posterior parameters φ and ψ indirectly. In LDA model, it only needs to sample the topic variables z_i. its posteriori probability is computed as follows:

$$P(\mathrm{z}_i = j / z_{-i.w_i}) = \frac{n_{-i,j}^{w_i} + x}{n_{-i,j}^{(.)} + Wx} \cdot \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,}^{(d_i)} + T\alpha} / \sum_{j=1}^{T} \frac{n_{-i,j}^{(w_i)} + x}{n_{-i,j}^{(.)} + Wx} \cdot \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,}^{(d_i)} + T\alpha} \qquad (2)$$

3. Feature Selection Based on the Categories LDA Model

In this paper, we adopt double feature selection algorithm. The frequency of words and documents is filtered in the text preconditioning. Assuming that a document is generated by several topics, then the different types of document are randomly composed by mixture themes which have different kinds of probability, and the same type of document has similar themes probability distribution. We consider that the content of subjects which is expounded by different types of document is disparate, and each kind of document has its own expression and organization structure. The main steps of the multi-class classification algorithm which combines the feature selection method based on the categories LDA model with SVM is as follows:

- (1) Assuming that the c category document sets include n documents, the known fixed parameters K, M and N_m are imported.
- (2) Deducing the valid information $P(w/K)$ and confirming the optimal topics K with the standard method of Bayesian statistical theory. $P(\omega/K)$ approximately is a series harmonic average of $\mathrm{P}(\omega/\mathrm{z})$, the value is calculated according to the following formula:

  $$\frac{1}{P(\omega/K)}=\frac{1}{M}\sum_{m=1}^{M}\frac{1}{P(\omega/z^{(m)})} \tag{3}$$

  The optimal topics K is confirmed when the value best fit the effective information of corpora.
- (3) The model parameters are estimated with the Gibbs sampling algorithms for the purpose of confirming the value of parameters α and β after enough iterative times. Then the other model parameters are computed with the known α and β

  $$P(\theta/\alpha)=\frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha-1_1}\ldots\theta_k^{\alpha_k-1} \tag{4}$$

  The conditional probability variable could be calculated by the formula (1) and (5).

  $$P(D/\alpha,\beta)=\prod_{m=1}^{M}\int p(\theta_m/\alpha)(\prod_{n=1}^{N_m}\sum_{z_{\max}}p(z_{\max}/\theta_m)p(w_{mn}/z_{mn},\beta))d\theta_m \tag{5}$$
- (4) Merging the same feature items in different types of documents, the probability distribution of feature words is obtained, namely, the implicit topics-documents matrix.
- (5) Training support vectors on the implicit topics-documents matrix, and constructing the multi-class text sorter, and then we can get the multi-class classification model based on SVM algorithm.
- (6) The Gibbs sampling algorithm will be used after preprocessing the test documents, and the probability distributions could be obtained after few iterative times. Then we could predict the categories of the unknown documents with the above multi-class SVM classification model.

4. Experiments and Results

For this work, we use the Chinese text classification corpora of Fudan University as experimental data; it is a typical unbalanced corpus. We choose 5 classes' documents and 2000s documents are randomly extracted from each classes for the purpose of eliminating the influence from the imbalanced data. We equally separate the corpora into ten groups, and each group is divided into train set and test set according to the proportion of one to one.

4.1. Confirming the optimal topics number

First of all, we adopt the maximal positive matching participle method based on Sogou dictionary to preprocess the corpus. We select the experience value $\alpha = 50/k$ and $\beta = 0.1$.

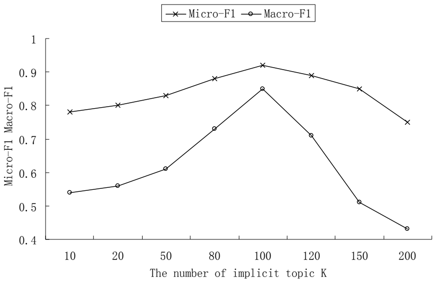

Figure 1 the value of Micro-F$_1$ and Macro-F$_1$ under different numbers of implied topic

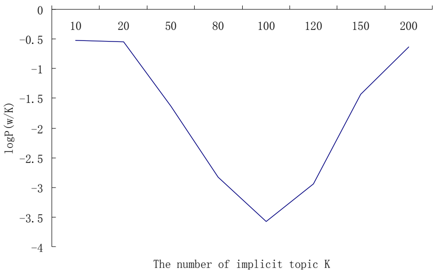

Figure 2 the relation between the connotative topics number K and log $P(\omega / K)$

As the above two figures show, the value of Micro- and Macro-averaged F$_1$-measures are the maximums and the value of log $P(\omega / K)$ is the minimum when the number of hidden topic is 100. This consequence means that it achieves the best fitting effect when the model fit the effective information at the hidden topics number 100. Therefore, we choose 100 for the number of hidden topics at the follow-up experiments.

4.2. The compare of feature selection approach

In the above LDA model, we continue to infer the parameters with the Gibbs sampling algorithms on the initialized data sets. Enough frequencies are iterated until convergence. Each document is signified as the polynomial probability distribution of topics set with 100s themes, in other words, the implied topics-documents matrix. And we compose the SVM classifier on this matrix.

As contrast experiment 1, we use the conventional method such as MI, Chi, TE, WET and ECE as feature selection methods, and the vector space model-VSM is adopted for text representation. The accuracy is used for the evaluation index, and table 1 shows the experimental results.

Table 1 the accuracy of the comparison among various feature selection methods

feature selection method	MI	Chi	TE	Wet	ECE	LDA	c-LDA
accuracy	77.8%	89.1%	87.6%	90.3%	86.9%	91.3%	93.2%

After the course of participle and stop words collation, there is 36,057 characteristics in the candidate feature set. We combine the VSM text representation with the MI feature selection and construct the SVM classifier in the contrastive experiment 2. The experimental result is showed in figure 7 and table 2.

Table 2 the comparison of the value of Micro- and Macro-averaged F1-measures

	Macro-P	Macro-R	Macro-F$_1$	Micro-F$_1$
MI+SVM	0.863596	0.865328	0.86463	0.872053
LDA+SVM	0.89279	0.889472	0.889523	0.890751
C-LDA+SVM	0.903726	0.894257	0.897205	0.909427

The candidate set includes 3000 dimension after the contrast experiment 2 and there are only 100 dimensions after the method of this paper. The degree of dimensionality reduction is 91.68% and 99.72% respectively.

5. Conclusions and the Next Work

In this paper, we have blended into the category information based on the existing LDA model feature selection algorithm for the purpose of discovery the difference of underlying topics among the disparate class documents, and the simplified multi-class textual data have been sorted with the SVM classifier. LDA model is an emergent probability model and it has the incomparable modeling strengths, SVM classification algorithm has the unique excellent properties on text categorization. We have combined the good text representation performance of the former with the powerful classification ability of the latter. The experimental results confirm the effectiveness and superiority of this method. However, the experiment is based on the hypothesis cases of balanced data set, but the distribution of each category is uneven in the actual data. So the theme modeling and the classification algorithm for unbalance database become the further step unfolding study. LDA model ignores the relationship between topics because of the hypothesis of independence and random exchange based on documents and words, but there are related themes in real data. The topic number K exert a tremendous influence on fitting documents sets. So how to determine the optimal topics number K is also an on-going research issue.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant No. 61073121 and the Natural Science Foundation of Hebei Province under Grant No. F2009000215.

References

[1] Zhang Xiao-Ping, Zhou Xue-Zhong, etc. A topic model based on CRP and word similarity. *Pattern Recognition and Artificial Intelligence*; 2010, p. 72-76.(in chinese)

[2] Bill B. Wang, R. I. Bob Mckay, etc. A comparative study for domain ontolory guided feature extraction. *ACSC `06*, Vol 16, Darlinghurst: Australian Computer Society, Inc; 2003, p. 69-78.

[3] D. M. Blei, A. Y. Ng, etc. Latent dirichlet allocation. *The Journal of Machine Learning Research*, Vol.3; 2003, p. 993-1022.

[4] D. M. Blei, T. L. Griffiths, etc. Hierarchical topic models and the nested Chinese restaurant rocess. In: Thrun S, Saul L K, SchiSlkopf B, eds. *Advances in Neural Information Processing Systems.* Cambridge, USA: MIT Press; 2004, p. 17-24.

[5] Wei Li, Andrew McCallum. Pachinko allocation: DAG-structured mixture models of topic correlations. *ICMI `06*. NY, USA: ACM; 2006, p. 577-584

[6] D. M. Blei, J. D. Lafferty. Correlated topic models. *Advances in Neural Information Processing Systems*; 2006, p. l47-154.

[7] B. Walsh. Markov chain monte carlo and gibbs sampling. *Lecture Notes for EEB 581*, version 26; 2004.

[8] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, Vol 2. No.3. NY, USA, ACM; 2011.