Chinese Text Classification Based On LDA and KSVM

Congwei Liang[1, a *], Yong Liu[2,b] and Haiqing Du[3,c]

[1]Beijing University of Posts and Telecommunications, Beijing, China

[a]billie_liang@163.com, [b]liuyo@bupt.edu.cn, [c]duhaiqing@bupt.edu.cn

Keywords: Machine Learning,Text Classification, LDA, KSVM, KNN, SVM-KNN

Abstract. With the rapid development of information technology and social networking, the amount of generated text data has increased enormously. As one of the crucial technologies for information organization and management, text classification has become much more significant in the area of machine learning and natural language processing. According to this paper, we present a text classification system. First, we apply LDA topic model to express the text instead of Boolean model or vector space model. Then, we choose KSVM which combines SVM with KNN as the classification algorithm. Finally, we choose documents with large amount of Chinese news for experiments. Compared with normal language models, these experimental data shows that our system gets higher classification accuracy.

Introduction

With the boosting online information and pervasion of Web pouring in our daily lives, there has been an increasing demand of methods to assist readers in organizing a large number of texts [1]. Text classification plays an important role in data miming. On this occasion, researchers started to pay more attention to text classification and have made surprising achievements in this area. Text classification system consists of text representation, preprocessing, feature dimension reduction and classification algorithm and effect evaluation. Feature dimension reduction and classification algorithm are the keys to the task.

Common feature dimension reduction includes DF(Document Frequency), CHI(Chi-Square), MI(Mutual Information), IG(Information Gain), TS(Term Strength) [2]. A common feature of these methods is the assumption that the words are independent and orthogonal to each other. The words are selected by a specific relationship between the calculation of words and categories, so as to achieve the purpose of dimension reduction. However, these methods ignore the synonym and the meaning of the word, without considering the semantic relation among the words. So, in our system, LDA is used for feature dimension reduction.

Classification algorithm is another key to text classification. Up to now, many popular machine learning algorithms have been applied to text categorization, such as k-Nearest Neighbor [3], Naive Bayes, Support Vector Machine [4], Neural Network [5], Maxent, Decision Tree. Among these methods, k-NN and SVM are the most commonly used algorithms. According to paper [6], we find the relationship between k-NN and SVM, so we put forward a new algorithm which combine SVM and KNN to improve the accuracy of classification algorithm.

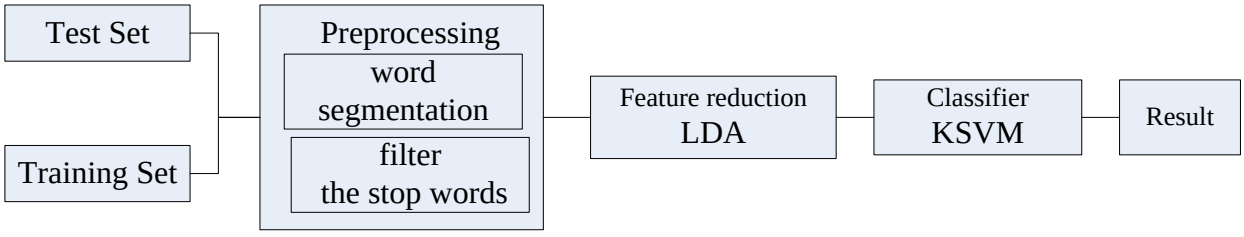

Fig. 1. Our System of text classification

In this paper we propose a text classification framework based on LDA and KSVM, as shown in Fig. 1. Our system consists of the following modules: preprocessing(word segmentation,tilter the stop words), LDA, KSVM. In the following section, we make a brief introduction to the relevant theory, and carry out the experiments and analyze the results.

Related theory

Latent Dirichlet Allocation. Latent Dirichlet Allocation (LDA) is a generative probabilistic model, which was proposed by Blei et al in 2003 [7]. Different from the traditional vector space model which represents text documents as vectors of keywords, LDA posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics, as shown in Fig. 2. The generative process for each document w in a corpus D is as follows:

1. Choose N ~ Possion(ξ), where N is the length of the document.
2. Choose θ_d ~ Dirichlet(α) and Possion(ξ), where θ_d is a multinomial distribution of topics over document d, α is the parameter of Dirichlet distribution.
3. Choose φ_k ~ Dirichlet(β), where φ_k is a multinomial distribution of words over topic k, β is the parameter of Dirichlet distribution.
4. For each word w_i of document d: first, Choose $z_n \sim$ Multinomial(θ_d), where z_n is topic that has been chosen; second choose a word $w_i \sim$ Multinomial(φ_{zn}).

According to the steps we have mentioned, we can obtain the marginal distribution of a document:

$$p(w|\alpha,\beta)= \int p(\theta|\alpha)\left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta)p(w_n|z_n,\beta)\right)d\theta \quad (1)$$

The probability of a corpus D is given by the following formula:

$$p(D|\alpha,\beta)= \prod_{d=1}^{M} \int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right)d\theta_d \quad (2)$$

In learning, the goal is to find the corpus-level parameters α and β to maximize the lob likelihood of the entire database. There are many methods to estimate the parameters such as, variational EM, Expectation Propagation, Gibbs sampling and so on. We choose Gibbs sampling in our system, which is one of the most commonly used and effective algorithm

Compared with other latent variable models, LDA has distinct pros [7]: firstly, LDA model allows documents to exhibit multiple topics to different degrees while each document exhibits exactly one topic in unigram and mixture of unigrams. Moreover, LDA overcomes linear growth in parameters and the difficulty of probability assignment to a previously unseen document in LSI and pLSI.

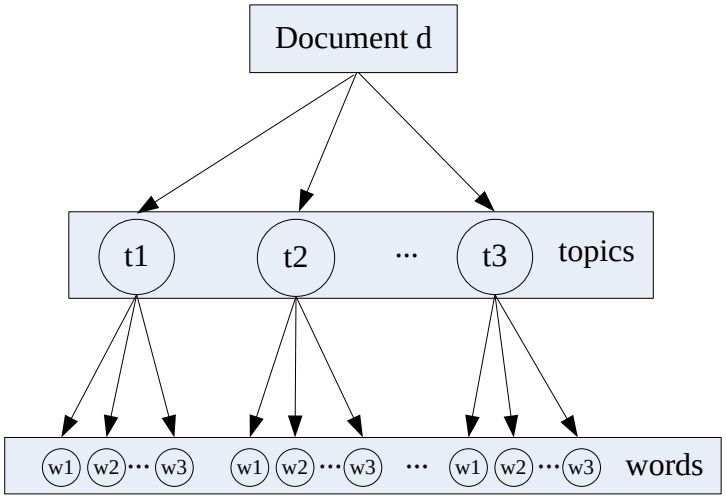

Fig. 2. Topology structure of LDA

The Support Vector Machine. The Support Vector Machine (SVM) which was firstly proposed in 1995 by Vapnik [8] is one of the most import machine learning algorithms widely used in

classification and regression tasks. SVM is based on the structure risk minimization principle and VC theory from the statistical learning theory. The task of SVM is to find a hyper-plane h that has the maximum margin as decision boundary in the linear space. By using kernel function in SVM, the sample space can be mapped into high dimensional feature space so that the linear inseparability problem in the original sample space can be transformed into linearly separable problem.

K-Nearest Neighbor. KNN (K-Nearest Neighbor) classification is a well-known statistical approach that has been intensively studied in pattern recognition for over four decades [9]. The main idea of KNN algorithm is: first calculate the similarity between test sample and all training samples, and find the K nearest neighbors of the test sample in training samples; then according to the categories of these neighbors to determine the category of test sample. KNN has been widely used in text classification, pattern recognition, image and spatial classification due to its simplicity and high accuracy.

KSVM. As mentioned in the [6], SVM is equivalent to 1NN, in which only one representative point is selected for each class. But sometimes the representative point is not a good representative of the class, so KVM has been proposed which is the combination of SVM and KNN for higher accuracy and better performance. Specifically, when an unknown point is in the region 1 or region 2 in Fig. 3, the general SVM can get right result easily for the distance between the point and hyper-plane exceeds the threshold. While the points in region 3 are difficult to classify by general SVM, KNN can get better result.

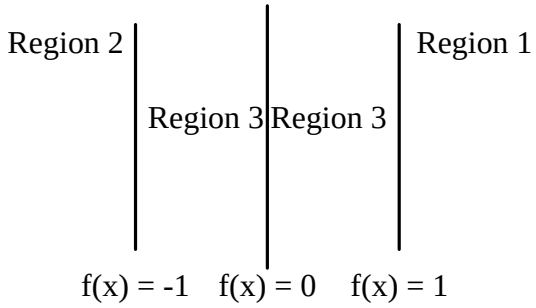

Fig. 3.

In paper [10] , the process of KSVM is as follows:

1. Train the data with SVM using appropriate parameter, get T_{sv}(the set of support vector), the parameter α_i, b of SVM, and set training set is T.
2. If $T \neq \Phi$, choose $x \in T$, if $T = \Phi$, stop;
3. Get $g(x) = \sum y_i \alpha_i K(x_i, x) - b$;
4. If $|g(x)| > \varepsilon$, $f(x) = sgn(g(x))$, if $|g(x)| < \varepsilon$, $f(x) = KNN(k, x, T_{sv})$;
5. $T = T - \{x\}$, go to Step2.

The process between the KNN in step2 and general KNN is almost the same, except training set to compute KNN and the way to compute the distance between two points. First, as mentioned above, the general KNN is time consuming, so we select every support vector as representative points in SVM-KNN. Second, we use Eq. 3 to compute the distance between two points rather than Euclidean distance which is the most commonly used algorithm.

$$d(x,x_i) = |\phi(x) - \phi(x_i)|^2 = k(x,x) - 2k(x,x_i) + k(x_i,x_i),\ x_i \in Tsv \quad (3)$$

Experiments and Results

Data. The dataset we use for evaluating our proposed system is sougou News data. It consists of 7 categories: car, finance, health, sports, tourism, education, and military. We isolate 90% data to train our model, the remaining data is testing data.

Experiment setup. In this section, we introduce the tool we use and the related parameters in our system. Unlike English, Chinese has no obvious boundaries between words, so Chinese word automatic segmentation is the most basic and important step in Chinese text preprocessing. In our

system, we adopt Jieba [11] as our Chinese word segmentation. We can add additional words to user dictionary in Jieba for a better performance.

We use the open source project JGibbLDA [12] to achieve the training of LDA. When the topic number k is 80, $\alpha = 50/k = 0.625$, $\beta = 0.01$, we can get the best result. We use LIBSVM [13] to train SVM model and choose RBF (radial basis function) as kernel function, c = 5.6, g = 0.2. After many experiments, the best parameter K in KSVM is 15.

Results. In our experiments, we compare the performance of our approach LDA-KSVM with the performance of LDA-SVM and LDA-KNN. For the evaluation of our system, the measurement includes precision, recall, F-measure [14] .The results are shown in Table.1 and Fig. 4. According to Table 1, LDA-KSVM performs better than LDA-KNN or LDA-SVM in all 7 categories of Sougou News. In Fig. 4, we can see recall and F-measure of LDA-KSVM also get much better result than the other algorithms.

Summary

In this paper, we present an automatic text classification system, in which we use LDA as feature dimension reduction and use KSVM as lassification algorithm. Compared with LDA-SVM and LDA-KNN, our system gets a better performance.

Table 1. Precision of LDA-KNN, LDA-SVM, LDA-KSVM of Sougou News

Algorithms / Categories	LDA-KNN	LDA-SVM	LDA-KSVM
Car	0.83	0.84	0.86
finance	0.88	0.89	0.98
health	0.76	0.79	0.83
sports	0.83	0.86	0.87
tourism	0.77	0.80	0.85
education	0.80	0.81	0.84
military	0.85	0.90	0.93
Average	0.81	0.84	0.88

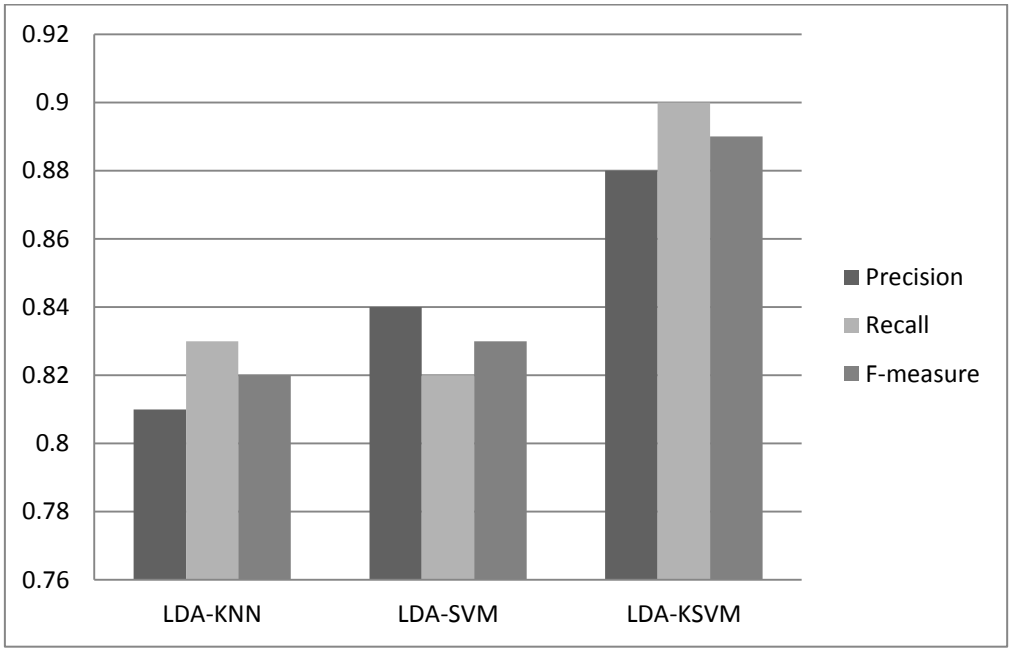

Fig. 4. Performances of LDA-kNN, LDA-SVM, LDA-KSVM on Sougou News

References

[1] Yi, Guo, Automatic text categorization based on content analysis with cognitive situation models, J. Information Sciences. 180 (2010) 613-630.

[2] WU, Jian-jun, KANG, Yao-hong, A Study on Feature Dimension Reduction in Text Categorization, J. Journal of Hainan University, (2007) 62-66.

[3] R. Duwairi, An eager k-nearest-neighbor classifier for Arabic text categorization, Proceedings of the International Conference on Data Mining, Nevada, USA, 2005, pp.187-192.

[4] J. Cervantes, X. Li, W. Yu, and K. Li, Support vector machine classification for large data sets via minimum enclosing ball clustering, Neurocomputing, 71 (2008) pp 611-619.

[5] Y. S. Xia and J. Wang, A one-layer recurrent neural network for support vector machine learning, IEEE Trans, Syst. Man Cybern. 2004, pp.1261-1269.

[6] LI, Rong, YE, Shi-wei, SHI, Zhong-zhi. SVM-KNN Classifier — A New Method of Improving the Accuracy of SVM Classifier, J. Acta Electronica Sinica, 30(2002) 745-748.

[7] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, J. Journal of Machine Learning,

Research. 3 (2003) 993-1022.

[8] Vladimir Cherkassky, Filip Mulier, Learning from Data: Concepts, Theory, and Methods, Wiley, New York, NY, 1998.

[9] B. V. Dasarathy, Nearest neighbor (NN) norms: NN pattern classification techniques, Los Alamitos: IEEE Computer Society Press, 1990.

[10] Li, Chengxiong, Application of SVM-KNN Combination Improvement Algorithm on Patent Text Classification, J. Computer engineering and Applications. 20 (2006) 193-196.

[11] Information on http://www.oschina.net/p/jieba.

[12] Information on http://jgibblda.sourceforge.net/.

[13] Information on http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[14] YAO Quanzhu, SONG Zhili, PENG Cheng, Research on text categorization based on LDA, J. Computer Engineering and Applications, 47 (2011) 150-153.