A Hybrid Approach for Sparse Data Classification Based on Topic Model

Guangjing Wang, Jie Zhang, Xiaobin Yang, and Li Li(✉)

Faculty of Computer and Information Science, Southwest University, Chongqing 400715, China
lily@swu.edu.cn

Abstract. With an increasing number of short text emerging, sparse text classification is becoming crucial in data mining and information retrieval area. Many efforts have been devoted to improve the efficiency of normal text classification. However, it is still immature in terms of high-dimension and sparse data processing. In this paper, we present a new method which fancifully utilizes Biterm Topic Model (BTM) and Support Vector Machine (SVM). By using BTM, though the dimensionality of training data is reduced significantly, it is still able to keep rich semantic information for the sparse data. We then employ SVM on the generated topics or features. Experiments on 20 Newsgroups and Tencent microblog dataset demonstrate that our approach can achieve excellent classifier performance in terms of precision, recall and F1 measure. Furthermore, it is proved that the proposed method has high efficiency compared with the combination of Latent Dirichlet Allocation (LDA) and SVM. Our method enhances the previous work in this field and establishes the foundation for further studies.

1 Introduction

More and more textual data is unfolding before people's eyes in more diverse forms with the rise of web 2.0. For example, multifarious data is generated from queries and questions in Web search, social networks, various internet news and so on. As a consequence, researchers are urged to solve the problem that internet users sometimes get bored because they are subject to a myriad of turbid information and the restraint of limited message coverage [19].

As an essential topic, lots of methods are put forward for the above problem. Text categorization used in information retrieval, news classification, spam mail filtering to acquire better user experience is studied roundly [10]. However, the applicability of classification for high dimensional and sparse data often becomes a short slab in many models. Like a teeter-board, the efficiency of processing sparse data and performance quality are hard to be fairness considered. On one hand, the classification accuracy would be descending if the dimension was cut down at an efficient level. On the other hand, for sparse and high dimensional datasets, the computing efficiency has to be sacrificed since the dimension will get to thousands or even more [13].

S. Song and Y. Tong (Eds.): WAIM 2016 Workshops, LNCS 9998, pp. 17–28, 2016.
DOI: 10.1007/978-3-319-47121-1_2

Researchers usually characterize sparse data by building semantics association or employing external knowledge base to settle the sparse feature problems. For instance, Wikipedia was used in [15] as an external corpus to rich the corpus. Cataldi et al. [2] used semantics relation rules to build relation rules library, so as to rich feature corpus. Xia et al. [20] introduced topics for multi-granularity, and then discriminative features are generated for sparse data classification. Nevertheless, it is hard to introduce external corpora to sparse text due to specific situations, and appropriate semantic association that can enhance the effect of sparse data classification [23]. What's more, the problem of accuracy and efficiency in classification are difficult to get an optimal solution [8].

A novel way to address the above problem is presented in our paper. To classifying sparse text accurately and fleetly, Biterm Topic Model (BTM) algorithm [21] is used for generating features, so that we can utilize topic information in Vector Space Model (VSM). Then the Support Vector Machine (SVM) is acted on it to obtain better classification result. Through the experiments on 20 Newsgroups datasets and dataset from Tencent Microblogs, we found that the combination of BTM and SVM enhances performance much more than other classification models for sparse data. Moreover, the proposed method provides a novel way to process sparse data.

The rest of the paper is organized as follows: the related work is reviewed in Sect. 2. Section 3 discusses our approach using BTM+SVM, and then the implementation is detailed in Sect. 4. Further discussion is presented experimentally in Sect. 5. Finally, Sect. 6 is the conclusion.

2 Related Work

Text classification is an important task for natural language process, and topic model is popular among researchers to process natural language. Liu et al. [9] devised a semi-supervised learning with Universum algorithm based on boosting technique. In their method, they aims to study a collection of nonexamples that do not belong to any class of interest. Luss et al. [11] developed an analytic center cutting plane method to solve the kernel learning problem efficiently, this method exhibits linear convergence but requires very few gradient evaluations. Lai et al. [7] applied a recurrent structure to capture contextual information as far as possible when learning word representations, and it is said that the proposed method shows better results than the state-of-the-art methods on document-level. By contrast, our method uses the generation of word co-occurrence pattern to keep main information while reducing dimensionality. Landeiro et al. [8] estimated the underlying effect of a text variable on the class variable based on Pearls back-door adjustment.

SVM is widely uesed in text classification. Yin et al. [22] used semi-supervised learning and SVM to improve the traditional method and it can classify a large number of short texts to mine the useful massage from the short text, however

the efficiency is not satisfactory. Song et al. [18] illustrated Chinese text feature selection method based on category distinction and feature location information, while this method has boundedness that location information is not easy to obtain. Nguyen et al. [14] proposed the improving multi-class text classification method combined the SVM classifier with OAO and DDAG strategies. In Seetha et al. [16], nearest neighbour and SVM classifiers are chosen as text classifiers for their good classification accuracy. Luo et al. [10] presented a method which combines the Latent Dirichlet Allocation (LDA) algorithm and SVM. However, the method is not good at deal with sparse text data according to our experiments. Altinel et al. [1] proposed a novel semantic smoothing kernel for SVM based on a meaning measure.

3 Problem Formalization

Motivated by researches on classification models, this study first formalizes the data collection to meet the prerequisites in algorithms. As usual, we use a vector to represent a document, and the whole text data can be regarded as a matrix. The problem is formalized technically as follows.

Every document and extracted term are supposed to be mapped into a vector [5] to represent text documents as a document-term matrix according to VSM.

$$d_j = (w_{1j}, w_{2j}, \ldots, w_{tj}) \tag{1}$$

Each dimension related to a separate term, where the value corresponds to the term is usually computed by term frequency-inverse document frequency model (TF-IDF). The weight vector for document d is

$$v_d = [w_{1,d}, w_{2,d}, \ldots, w_{N,d}]^T \tag{2}$$

where $w_{t,d} = tf_{t,d} \cdot log\dfrac{|D|}{|\{d\prime \in D|t \in d\prime\}|}$ and tf_{td} is the term frequency of term t in document d, $|D|$ is the total number of documents in the set; $|\{d\prime \in D|t \in d\prime\}|$ is the number of documents containing the term t.

For dimension reduction, there are two general ways to apply. One is feature extraction, large data is transformed into a reduced features vector, so that the desired task can be solved using the reduced representation [13]. The data transformation model can be nonlinear like kernel principal component analysis, linear like latent semantic indexing, linear discriminant analysis and so on. The other one is known as feature selection, such as χ^2 statistic, document frequency and so forth, those are selecting a subset of relevant features for use in model construction.

4 Novel Method for Sparse Data Classification

In this section, we will illustrate our method for sparse data classification carefully. To begin with, an overview of BTM and SVM model is presented.

After that we will elaborate how to employ BTM to generate the document topic matrix, and then explain how to utilize the SVM to classify and predict the category of sparse data.

4.1 Matrix of Topic Distribution

BTM is a probabilistic model that learns topics over short texts by directly using the generation of biterms in the whole corpus [21]. The notation of "biterm" refers to an instance of unordered word pair occurrence, and any two distinct words in a document compose a biterm. The model in graph is showed in Fig. 1. The key point is that two words are more likely to be in the same topic if they co-occur more frequently.

Given a corpus with N_D documents, we can utilize a K-dimensional multinomial distribution $\theta = \{\theta_k\}_{k=1}^{K}$ with $\theta_k = P(z = k)$ and $\sum_{k=1}^{K} \theta_k = 1$ to show the prevalence of topics. Suppose each biterm is drawn from a specific topic independently, the specific generative process of the corpus in BTM can be shown as follows [4]. The notations used in BTM are listed in Table 1.

1. For each topic z, draw a topic-specific word distribution $\phi_z \sim Dir(\beta)$.
2. Extracting a topic distribution $\theta \sim Dir(\alpha)$ for the whole collection.

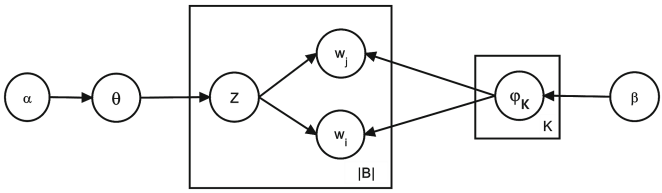

Fig. 1. BTM: a generative graphical model

Table 1. Notations in BTM

N_D	The number of documents
K	The number of latent topics
W	The number of unique words
$\|B\|$	The number of biterms
$\mathbf{B} = \{b_i\}_{i=1}^{\|B\|}$	The collection of biterms
$b_i = w_{i,1}, w_{i,2}$	The i-th biterm
$\theta = \{\theta_k\}_{k=1}^{K}$	A K-dimensional multinomial distribution
$\theta_k = P(z = k)$	The prevalence of topic k where $\sum_{k=1}^{K} \theta_k = 1$
Φ	A $K \times W$ matrix
Φ_k	A W-dimensional multinomial distribution in k-th row
α, β	Dirichlet hyperparameters

3. For each biterm b in the biterm set B, draw a topic assignment: $z \sim Multi(\theta)$, and draw two words: $w_i, w_j \sim Multi(\phi_z)$.

The joint probability of a biterm $b = (w_i, w_j)$ over topic z can be written as:

$$P(b) = \sum_z P(w_i|z)P(w_j|z) = \sum_z \theta_z \phi_{i|z} \phi_{j|z} \tag{3}$$

Similar as LDA, Gibbs sampling can be adopted to perform approximate inference. In the process, the topic-word distribution ϕ and global topic distribution θ can be generated as:

$$\phi_{w|z} = \frac{n_{w|z} + \beta}{\sum_w n_{w|z} + M\beta} \tag{4}$$

$$\theta_z = \frac{nz + \alpha}{|B| + K\alpha} \tag{5}$$

where $|B|$ is the aggregated number of biterms. The matrix θ is an essential part of our method as the matrix of topic distribution.

4.2 Support Vector Machine (SVM)

SVM plays an important part in lots of domains, and hyperplanes are constructed when it performs classification tasks in a multidimensional space. It is reported that SVM can generate better results than other learning algorithms in classification [6]. The basic theory of SVM is elaborated next:

When the training dataset of n points in the form of $(\boldsymbol{x_1}, \boldsymbol{y_1}), \ldots, (\boldsymbol{x_n}, \boldsymbol{y_n})$ is known, where y_i is either 1 or -1, the optimization problem is defined as:

$$min\ \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i \ \ s.t.\ \ y(w^T \phi(x_i) + b) \geq 1 - \zeta_i, \zeta_i \geq 0 \tag{6}$$

where function ϕ can map training vectors x_i into a higher dimensional space b. $C > 0$ is the penalty parameter of the error instances, which should be chosen with care to avoid over fitting. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables. On the basis of Mercer theorem [12], there always exists an equation $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ called the kernel function. The problem 6 can be derived as:

$$f(x) = \sum_{i=1}^{l} a_i y_i K(x_i, x_j) + b \tag{7}$$

By solving the optimization, parameters of the maximum-margin hyperplane are derived specifically. Note that the core of SVM which is good at processing high dimensional data is that the number of dimensions can be turned from $\phi(x_i)$ to x_i. What's more, LIBSVM [3] has some attractive training time properties. Each convergence iteration takes linear time to read the training data and the iterations also have a Q-Linear Convergence property, which makes the algorithm extremely fast [17].

4.3 Experimental Procedure for Enhancement

For less complexity and higher performance, our method retrieves optimal set of features, which reflects the original data distribution. The steps in document classification are listed as follows.

Step 1. Making a document-term matrix according to the vector support model.
Step 2. Analysing the topic distribution and building a matrix about topic distribution for documents.
Step 3. Acquiring the weight of vector support model by using the topic distribution values.
Step 4. Testing documents by building the classifier.

We firstly formalize the data collection in order that it can be used in SVM, so a document-term matrix must be built in Step 1. Since Step 2 utilizes matrix θ to indicate the relationship between texts and topics, we need to generate it by BTM estimation with Gibbs sampling first. In Step 4, SVM is used to build upon the characteristics identified in Step 2.

5 Experimental Evaluation

In this section, we conduct several experiments to show the great superiority of our method, results are presented below followed by discussion.

5.1 Data Preparation

We evaluate our method on two popular datasets used in large scale and sparse text classification study. One is Tencent microblogs, which contains 11,285,538 messages from seven different micro-channels posted by users from July 2 to July 14 in 2013 [19] on Tencent microblog platform (http://t.qq.com/). The other dataset

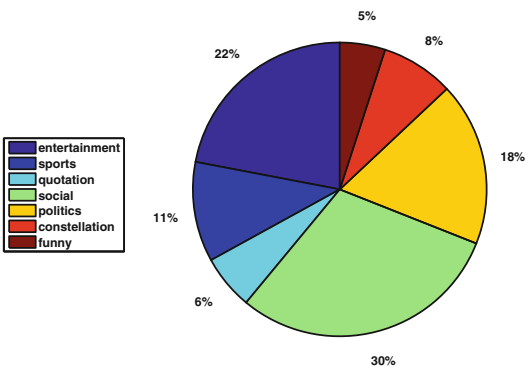

Fig. 2. Category distribution of Tencent messages

Table 2. Data description for 20 Newsgroups

Dataset	Category	Training data	Test data
20 Newsgroups	alt.atheism	480	319
	comp.graphics	584	389
	comp.os.ms-windows.misc	591	394
	comp.sys.ibm.pc.hardware	590	392
	comp.sys.mac.hardware	578	385
	comp.windows.x	593	395
	misc.forsale	585	390
	rec.auto	594	396
	rec.motorcycles	598	398
	rec.sport.baseball	597	397
	rec.sport.hockey	600	399
	sci.cypt	595	396
	sci.electronics	591	393
	sci.med	594	396
	sci.space	593	394
	sci.religion.christian	599	398
	talk.politics.guns	546	364
	talk.politics.mideast	564	376
	talk.politics.misc	465	310
	talk.religion.misc	377	251

is 20 Newsgroups (http://qwone.com/~jason/20Newsgroups/), which has 20 categories and is widely used in text classification.

The raw data of these collections is very noisy. For preprocessing, the terms like the punctuation marks, stop words, links and other non-words in the raw microblogging datasets are removed in data preparation using a punctuation list and a stop words dictionary. Specifically, for the process of word segmentation, the ICTCLAS (http://www.ictclas.org/) is used in this paper.

To further describe the datasets for classification, Fig. 2 is showed for category distribution of Tencent messages, and Table 2 illustrates the classical data proportion on 20 Newsgroups.

5.2 Evaluation Criteria

In our experiment, the $Macro/Micro - precision$, $Macro/Micro - Recall$ and $Macro/Micro - F1$ criteria are employed to evaluate the method. The definitions

are showed below.

$$Micro - Precision = \frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m} TP_i + FP_i} \tag{8}$$

$$Micro - Recall = \frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m} TP_i + FN_i} \tag{9}$$

$$Micro - F1 = \frac{Micro - Precision \times Micro - Recall \times 2}{Micro - Precision + Micro - Recall} \tag{10}$$

$$Macro - Precision = \frac{1}{m} \sum_{i=1}^{m} P_i \tag{11}$$

$$sMacro - Recall = \frac{1}{m} \sum_{i=1}^{m} R_i \tag{12}$$

$$Micro - F1 = \frac{Macro - Precison \times Macro - Recall \times 2}{Macro - Precision + Macro - Recall} \tag{13}$$

5.3 Results and Analysis

We choose two other methods PCA+SVM and LDA+SVM as baselines to verify the advantage of our approach. Documents used in our experiments are mapped into document-term matrix firstly. Considering topic model as a method of dimensionality reduction firstly, we then trained the document vectors by LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html), and we then predicted the categories of new documents. Unlike the PCA method which treats terms as features of document vector, the LDA and BTM methods use the topics as features of documents vectors. In order to obtain document-topic matrix, the widely used LDA tool GibbsLDA++ (http://gibbslda.sourceforge.net/) was employed in our experiments. BTM (http://shortext.org/) is first used to acquire the matrix of topic distribution for documents. The number of Gibbs sampling iterations in the following experiment is set to 1000 to insure the classification accuracy.

We use $Macro - Precision$, $Macro - Recall$, $Macro - F1$ and $Micro - F1$ to evaluate the classifiers PCA+SVM, LDA+SVM and BTM+SVM based on 20 Newsgroups which are depicted in Figs. 3 and 4, respectively. What need to mention is that Micro-Precision and Micro-Recall are the same as Micro-F1 since we suppose each instance has exactly one correct label. From the result, we can see that the values in Fig. 3 reach peak value after the dimensionality is brought down at 400. By contrast, as we can see from Fig. 4, when the number of topics is merely set to 180 for BTM+SVM, the $Macro - Precision$, $Macro - Recall$, $Macro - F1$ and $Micro - F1$ undulate slightly around 0.87,0.86,0.87,0.90, respectively. It can be seen from that the values of those criteria for BTM+SVM are relatively higher than those of PCA+SVM and LDA+SVM, respectively.

Comparison experiments were made in order to verify the high performance of BTM for feature selection, we estimated the number of iterations needed to obtain high accuracy by spending less time on topic-matrix generation.

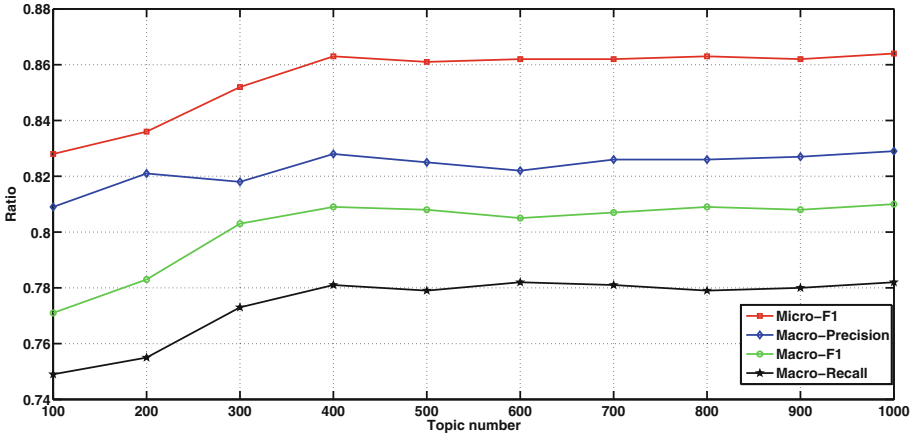

Fig. 3. The values of evaluation criteria under diverse number of features reduced by PCA+SVM method on 20 newsgroups collection

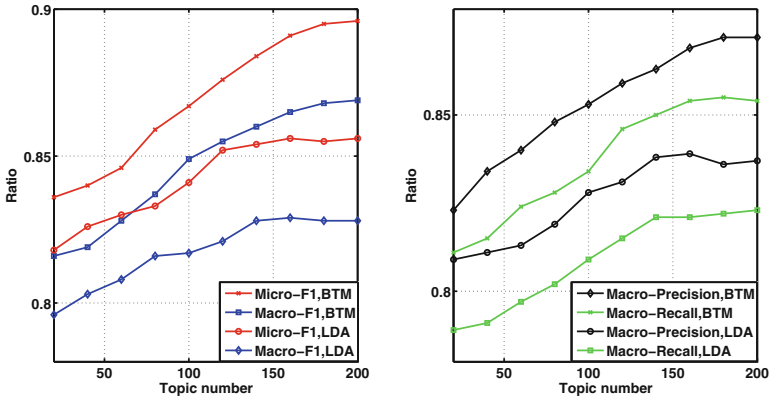

Fig. 4. The values of evaluation criteria under diverse number of features reduced by LDA+SVM, BTM+SVM methods on 20 newsgroups collection

The accuracy on 5-fold cross validation is reported in Fig. 5. It can be seen that 900 iterations is a relatively better choice on Tencent Dataset, and accuracy keeps around 90 % with 60 features generated. From Fig. 5(b), we can see that all the methods work better with training data size grows. It suggests that the LDA+SVM method is not able to overcome the sparsity problem, while BTM+SVM can achieve better performance than LDA+SVM, which also shows the superiority of our method.

BTM+SVM can resolve the over-fitting and feature redundancy problem, and yields better classification results than others. Utilizing the topical model is able to accelerate the process of classification. What's more, for sparsity problem in conventional topical model, BTM is better at capturing the topics by using

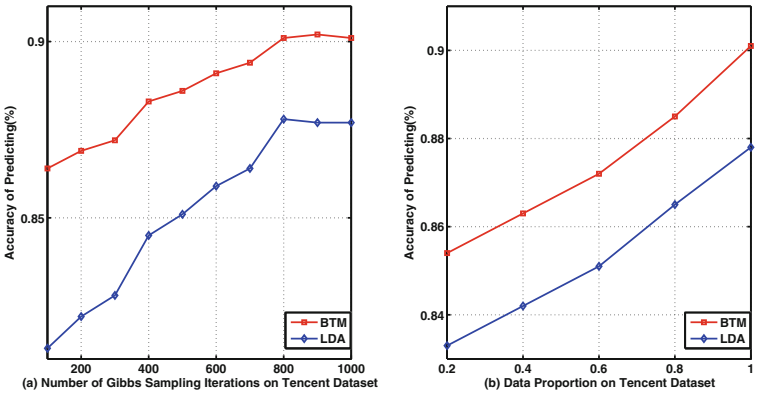

Fig. 5. Comparision of classification performance in different aspects between LDA+SVM and BTM+SVM on Tencent Dataset

Table 3. Time cost for dimensionality generated on 20 Newsgroups by three following methods using 3.0 GHz CPU, 2G memory

Methods	File quantity	Time consumed	Dimensionality generated
PCA+SVM	18846	Roughly 250 min	100
LDA+SVM	18846	Roughly 80 min	100
BTM+SVM	18846	Roughly 50 min	100

word co-occurrence patterns in the whole corpus [21]. The Table 3 presents information about training speed of three provided methods, which also shows the high efficiency of BTM+SVM by comparison. It only takes 50 min to generate a topic matrix by GibbsLDA++ with 100 topics and 1000 iterations, which saves about 30 min than LDA+SVM and is only one fifth of the time PCA+SVM consumed.

6 Conclusion

In this paper, we proposed a hybrid approach called BTM+SVM for sparse data classification. We explored the difference among BTM+SVM, PCA+SVM and LDA+SVM, and the results showed that our method has superiority over accuracy and efficiency when sparse text is processed. We figured out the number of topics to use when approximating the matrix properly. Comparing with traditional methods, we improved the classification accuracy and tested the training speed over the experiments. Overall, our method is able to cope with sparse problem properly, which is promising and can be used extensively in real applications.

Acknowledgments. This work is supported by Natural Science Foundations of China (No. 61170192), National High-tech R&D Program of China (No. 2013AA013801), Fundamental Research Funds for the Central Universities (No. XDJK2016E064).

References

1. Altınel, B., Ganiz, M.C., Diri, B.: A corpus-based semantic kernel for text classification by using meaning values of terms. Eng. Appl. Artif. Intell. **43**, 54–66 (2015)
2. Cataldi, M., Di Caro, L., Schifanella, C.: Emerging topic detection on twitter based on temporal and social terms evaluation. In: Proceedings of the Tenth International Workshop on Multimedia Data Mining, p. 4. ACM (2010)
3. Chang, C.-C., Lin, C.-J.: Libsvm: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) **2**(3), 27 (2011)
4. Cheng, X., Yan, X., Lan, Y., Guo, J.: BTM: topic modeling over short texts. IEEE Trans. Knowl. Data Eng. **26**(12), 2928–2941 (2014)
5. Dhillon, I.S., Modha, D.S.: Concept decompositions for large sparse text data using clustering. Mach. Learn. **42**(1–2), 143–175 (2001)
6. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J.: Liblinear: a library for large linear classification. J. Mach. Learn. Res. **9**, 1871–1874 (2008)
7. Lai, S., Xu, L., Liu, K., Zhao, J.: Recurrent convolutional neural networks for text classification. In: AAAI, pp. 2267–2273 (2015)
8. Landeiro, V., Culotta, A.: Robust text classification in the presence of confounding bias (2016)
9. Liu, C.-L., Hsaio, W.-H., Lee, C.-H., Chang, T.-H., Kuo, T.-H.: Semi-supervised text classification with universum learning. IEEE Trans. Cybern. **46**(2), 462–473 (2015)
10. Luo, L., Li, L.: Defining and evaluating classification algorithm for high-dimensional data based on latent topics. PloS one **9**(1), e82119 (2014)
11. Luss, R., d'Aspremont, A.: Predicting abnormal returns from news using text classification. Quant. Financ. **15**(6), 999–1012 (2015)
12. Minh, H.Q., Niyogi, P., Yao, Y.: Mercer's theorem, feature maps, and smoothing. In: Lugosi, G., Simon, H.U. (eds.) COLT 2006. LNCS (LNAI), vol. 4005, pp. 154–168. Springer, Heidelberg (2006). doi:10.1007/11776420_14
13. Moura, S., Partalas, I., Amini, M.-R.: Sparsification of linear models for large-scale text classification. In: Conférence sur l'APprentissage automatique (CAp 2015) (2015)
14. Nguyen, V.T., Huy, H.N.K., Tai, P.T., Hung, H.A.: Improving multi-class text classification method combined the svm classifier with oao and ddag strategies. J. Convergence Inf. Technol. **10**(2), 62–70 (2015)
15. Phan, X.-H., Nguyen, L.-M., Horiguchi, S.: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: Proceedings of the 17th International Conference on World Wide Web, pp. 91–100. ACM (2008)
16. Seetha, H., Murty, M.N., Saravanan, R.: Effective feature selection technique for text classification. Int. J. Data Min. Model. Manag. **7**(3), 165–184 (2015)
17. Shalev-Shwartz, S., Singer, Y., Srebro, N., Cotter, A.: Pegasos: primal estimated sub-gradient solver for svm. Math. Program. **127**(1), 3–30 (2011)

18. Song J., Zhang P., Qin S., Gong, J.: A method of the feature selection in hierarchical text classification based on the category discrimination and position information. In: 2015 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII), pp. 132–135. IEEE (2015)
19. Wang, J., Li, L., Tan, F., Zhu, Y., Feng, W.: Detecting hotspot information using multi-attribute based topic model. PloS one **10**(10), e0140539 (2015)
20. Xia, C.-Y., Wang, Z., Sanz, J., Meloni, S., Moreno, Y.: Effects of delayed recovery and nonuniform transmission on the spreading of diseases in complex networks. Phys. A: Stat. Mech. Appl. **392**(7), 1577–1585 (2013)
21. Yan, X., Guo, J., Lan, Y., Cheng, X.: A biterm topic model for short texts. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 1445–1456. International World Wide Web Conferences Steering Committee (2013)
22. Yin, C., Xiang, J., Zhang, H., Wang, J., Yin, Z., Kim, J.-U.: A new svm method for short text classification based on semi-supervised learning. In: 2015 4th International Conference on Advanced Information Technology and Sensor Application (AITS), pp. 100–103. IEEE (2015)
23. Zhang, H., Zhong, G.: Improving short text classification by learning vector representations of both words and hidden topics. Knowl.-Based Syst. **102**, 76–86 (2016)