A TEXT CLASSIFICATION METHOD BASED ON LATENT TOPICS

Yanshan Wang and In-Chan Choi
Department of Industrial Management Engineering, Korea University, Anamdong, Seoul, Republic of Korea

Keywords: Text Classification, Latent Topic, Indexing by LDA.

Abstract: Latent Dirichlet Allocation (LDA) is a generative model, which exhibits superiority over other topic modelling algorithms on latent topics of text data. Indexing by LDA is a new method in the context of LDA to provide a new definition of document probability vectors that can be applied as feature vectors. In this paper, we propose a joint process of text classification that combines DBSCAN, indexing with LDA and Support Vector Machine (SVM). DBSCAN algorithm is applied as a pre-processing for LDA to determine the number of topics, and then LDA document indexing features are employed for text classifier SVM.

1 INTRODUCTION

For decades, the information on the Internet has been increasing explosively. Amongst many other forms of information representation, text data holds a crucial position. Text data help people obtain valuable information with reduced network data flow. The task of categorizing natural language texts into topical categories has become one of the key factors in organizing online information.

Many machine learning and statistical language processing methods have been applied to text classification. Relavent approaches include decision trees, k-nearest neighbors (kNN) (Hart, 1967), neural networks (Wiener, Pedersen, and Weigend, 1995), Naïve Bayes and support vector machines (SVM) (Cortes and Vapnik, 1995). Joachims (1988) was the first to apply SVM to text classification. He used the inverse document frequency as a feature vector and reported the computational performance of SVM.

However, a big problem in the text classification is that the captured text data often lies in a high-dimensional feature space. Thus, efficient dimension reduction remains to be a big challenge in classification methods. One of the efficient dimension reduction techniques is Latent Dirichlet Allocation (LDA) (Blei et al., 2003), which is originally designed for topic and document modelling. It is a newly designed sequential generative model that aims to represent text documents based on latent topics.

2 OVERVIEW OF OUR METHOD

In this paper, indexing by LDA (Choi and Lee, 2010) is used for document modelling. Because the document probability proposed by Choi and Lee (2010) is easy to obtain, it is utilised to reduce any document to a fixed set of real-valued features.

The number of latent topics k plays a crucial role when LDA generates latent topics for a corpus. It is necessary to find a method that is effective for determining k. Here in this paper, the DBSCAN clustering idea is considered.

Our proposed method for the text classification consists of the following three steps:

1. Topic number determination step. DBSCAN is used to acquire the number of clusters k of the feature words.
2. Document modelling step. The indexing with LDA model is used with the number of topics k to obtain the feature of each document.
3. Classification step. Using the document probability as the feature for each document, SVM classifier is utilised to generate a model for the corpus.

This paper is organized as follows: the components in the proposed model, including DBSCAN algorithms, indexing by LDA and the SVM algorithm, are presented in the next section. Section 3 reports preliminary computational results on a benchmark data set. The last section presents concluding remarks with some future work.

Wang Y. and Choi I..
A TEXT CLASSIFICATION METHOD BASED ON LATENT TOPICS.
DOI: 10.5220/0003740902120214
In *Proceedings of the 1st International Conference on Operations Research and Enterprise Systems* (ICORES-2012), pages 212-214
ISBN: 978-989-8425-97-3

3 MODEL COMPONENTS

3.1 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) (Ester *et al*, 1996) is a density-based clustering algorithm designed to discover clusters. As in many document clustering algorithms, in which documents and words are represented by the vector space model (Van, 1979), their method considers each document as a vector in the term-space. The vector is represented by the term frequency (TF) vector: $d_j = \{tf_{1j}, tf_{2j}, \dots, tf_{Nj}\}$, where tf_{ij} is the frequency of term *i* in the document *j*. Just as each document can be represented by a vector of *tf-idf* weight, each word is denoted inversely as:

$$w_i = \{tf_{i1} \times idf_i, tf_{i2} \times idf_i, \dots, tf_{iN} \times idf_i\} \quad (1)$$

where $idf_i = \log \frac{N}{df_i}$ and df_i is the number of documents that contain word *i*.

The distance of two words is defined by the cosine similarity measure, i.e.

$$dis(w_i, w_j) = \cos(w_i, w_j) = \frac{w_i \cdot w_j}{\|w_i\| \|w_j\|}. \quad (2)$$

The cosine value is 1 when two words are identical, and 0 if there is nothing in common between them.

3.2 Indexing by LDA

The Latent Dirichlet Allocation (LDA) method is based on the assumption of "*bag-of-words*" that the order of words in a document can be neglected and the documents in a corpus share some latent topics (Blei *et al*, 2003).

In LDA, a document of size N is defined as a sequence of N words denoted by $w = \{w_1, w_2, \dots, w_N\}$, where $w_n \in \{w^1, w^2, \dots, w^K\}$ denotes n^{th} word in the sequence. A corpus is defined as a collection of M documents denoted by $C = \{w_1, w_2, \dots, w_M\}$. LDA generates topic model from given documents, which is obtained by inferring the topic mixture θ in document-level, where $\theta = (\theta_1, \theta_2, \dots, \theta_N)$, and $\theta_n \in \{\theta^1, \theta^2, \dots, \theta^K\}$, which is a K-dimensional vector. A set of N topics in word-level is defined as z, where $z = (z_1, z_2, \dots, z_N)$, $z_n \in \{z^1, z^2, \dots, z^K\}$.

The document generating mechanism assumed by LDA consists of three steps:

1) Choose the number of words N ~ Poisson
2) Choose θ ~ Dirichlet(α)
3) For each of the N words w_n:
   a) Choose a topic z_n ~ Multinomial(θ).
   b) Choose a word w_n ~ Multinomial($\theta|z_n$), a multinomial distribution conditioned on the topic z_n.

Choi and Lee (2010) make a direct use of the beta matrix in LDA to get the document probability:

$$\begin{aligned} D_i^k &\cong \widetilde{D}_i^k \\ &= \sum_{j=1}^{V} p(z^k|w^j, d_i) p(w^j|d_i) \\ &= \frac{\sum_{j=1}^{V} W_j^k n_{ij}}{N_{d_i}}, \end{aligned} \quad (3)$$

Where W_j^k represents the probability of the word w^j embodying the *k*th concept z^k, n_{ij} denotes the number of occurrence of word w^j in document d_i and N_{d_i} denotes the number of words in the document d_i, i.e. $N_{d_i} = \sum_j^V n_{ij}$.

3.3 Support Vector Machines

Support Vector Machines (SVM) is a type of learning system developed by Vapnik (1995) based on the structural risk minimization principle from the statistical learning theory. The document probability matrix obtained from the indexing by LDA model is used as an input.

Here, the RBF kernel function (Aizerman et al., 1964) is utilized. The classification decision is obtained by the following equation:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{n} y_i \alpha_i^* \exp\left(\frac{-\|x - x'\|^2}{\sigma^2}\right) + b^*\right). \quad (4)$$

4 PRELIMINARY RESULT

In this section, the results on a preliminary experiment on real data are provided. The proposed method was tested on the first four categories of a subset of 20news-bydate-matlab corpus (Available at: http://people.csail.mit.edu/jrennie/20Newsgroups/20news-bydate-matlab.tgz). Table 1 shows the numbers of training and testing documents in four categories. The corpus contains 3710 documents, 56222 words and 4 categories. The first 2220 documents were trained to build the classifier and the categories of the remaining 1490 documents were predicted to test the proposed model.

Table 1: Number of Training and Testing Items.

Category Name	Num Train	Num Test
alt.atheism	480	318
comp.graphics	581	391
comp.os.ms-windows.misc	572	391
comp.sys.ibm.pc.hardware	587	390
All	2220	1490

The DBSCAN clustering algorithm resulted in 15 clusters. In order to see the effectiveness of the algorithm, the proposed method without DBSCAN was tested by varying the number of latent topics and the test result is shown in Figure 1.

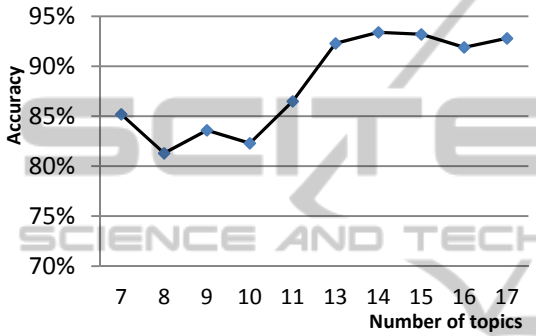

Figure 1: Accuracy according to different numbers of latent topics.

DBSCAN algorithm has obtained a satisfying result, even though the highest accuracy is acquired when the number of topics is 14. The precision and recall with the proposed method is shown in Table 2.

Table 2: Precision and Recall.

Category Name	Precision	Recall
alt.atheism	93.7%	99.3%
comp.graphics	97.1%	94.3%
comp.os.ms-windows.misc	92.5%	92.1%
comp.sys.ibm.pc.hardware	92.2%	88.2%
Accuracy	93.2%	

Computational comparison with other classification algorithms remains as a future study.

5 CONCLUSIONS

This paper presented a joint project based on a generative model LDA and Support Vector machine to categorize text. DBSCAN clustering algorithm is used to obtain the number of latent topics. A preliminary experiment was carried out to a small-scale data corpus. Further research on applications to other large-scale data is necessary. Moreover, other classification algorithms should also be tested on this corpus in comparison with the proposed method.

Furthermore, the potential use of the proposed method to the area of information systems is subject to a further study.

ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (2009-0083893).

REFERENCES

Aizerman, M. A., Braverman, E. M., and Rozono'er, L. I., 1964. Theoretical foundations of the potential function method in pattern recognition learning. *Automat. Rem. Control*, 25, pp.824-837.

Blei, D. M., Ng, A. Y., and Jordan, M. I., 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, pp.993-1022.

Choi, I.C., and Lee, J. S., 2010. Document indexing by latent dirichlet allocation. *Proceedings of The 2010 International Conference on Data Mining*, pp.409-414.

Cortes, C., & Vapnik, V., 1995. Support vector networks. *Machine Learning*, 20(3), pp.273-297.

Ester, M., Kriegel, H. P., Sander, J., and Xu, X., 1996. A density based algorithm for discovering clusters in large spatial databases. *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*, pp.226-231.

Joachims, T., 1988. Text categorization with support vector machines: Learning with many relevant features. *Proceedings of the 10th European conference on machine learning*, pp.137-142.

Hart, P., 1967. Nearest neighbor pattern classification. *IEEE Transaction on Information Theory*, 13(1), pp.21-27.

Van Rijsbergen, C. J., 1979. *Information Retrieval*, Buttersworth. London, 2th edition.

Vapnik, V., 1995. *The nature of statistical learning theory*, Springer. New York.

Wiener, E., Pedersen, J. O., Weigend, A. S., 1995. A Neural Network Approach to Topic Spotting. *SDAIR*, pp.317-332.