Chinese Text Categorization Based on Deep Belief Networks

Jia Song
The Faculty of Science and Technology
Communication University of China
Beijing, China
songjia@cuc.edu.cn

Sijun Qin
New Media Institute
Communication University of China
Beijing, China
wingol88888@163.com

Pengzhou Zhang
The Faculty of Science and Technology
Communication University of China
Beijing, China
zhangpengzhou@cuc.edu.cn

Abstract—**With the rapid development of Internet, text categorization becomes a mission-critical technology that organizes and processes large amounts of data in document. Deep belief networks have powerful abilities of learning and can extract highly distinguishable features from the high-dimensional original feature space. So a new Chinese text categorization algorithm based on deep learning structure and semi-supervised deep belief networks is presented in this paper. We extract original feature with TFIDF-ICF, construct the text classification model based on DBN, and select the number of hidden layers and hidden units. Our experimental results indicated that the performance of text categorization algorithm based on deep belief networks is better than support vector machine.**

Keywords-text categorization; restricted boltzmann machine; deep belief networks;

I. INTRODUCTION

The few years have seen significant interest in deep learning algorithms that learn layered, hierarchical representations of high-dimensional data [1] [2] [3]. These deep learning algorithms have been successfully applied to image recognitions [4], handwriting recognition [5] and voice recognition [6], but not extensively to text classification. However, the amount of information now has been growing fast along with rapid development of information technology, and it brings great challenge to information retrieval and data mining. Among them, text categorization is an important basis for information retrieval and data mining. A number of text categorization methods, i.e. naïve bayes algorithm, k-nearest neighbor algorithm, support vector machine algorithm, back propagation neural networks machine, have been applied to text classification, but are facing challenge with the increasing amounts of text data. Hinton et al. proposed deep belief networks.

The DBN can learn more features with hidden layers and get more complex functions to express data. DBN is a generative probabilistic model composed of one visible layer and some hidden layers. Each hidden layer unit learns a statistical relationship between the units in the lower layer, the higher layer representations tend to become more complex. The deep belief network can be efficiently trained using greedy layer-wise training, in which the hidden layers are trained one at a time in a bottom-up fashion[7] [8]. Then the deep belief networks can achieve the approximation of complex functions by learning a deep nonlinear network structure.

In this paper, we propose a machine learning algorithm based on DBNs for text classification and evaluate its performance on Sogou corpus. We took same classification tasks with support vector machine algorithm. By comparing the performance in classification tasks, we proved our method based on DBNs provides similar or even better performance than SVMs in Chinese text classification.

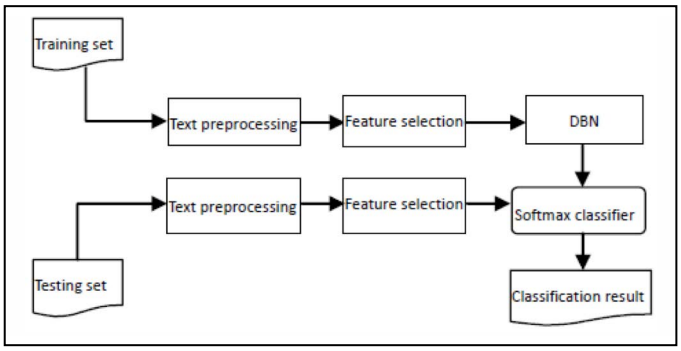


Figure 1. The procedure of text categorization.

II. TEXT CATEGORIZATION PROCEDURE

Because a text document is a collection of a large number of characters and it is unstructured or semi-structured digital information, it cannot be directly recognized by any classifier. For further analysis and processing, it must be converted into a simple, uniform and structural form which can be recognized by classifier or learning algorithm. The procedure of text categorization is shown in Fig. 1. Preprocess the text, represent

978-1-5090-0806-3/16/$31.00 copyright 2016 IEEE
ICIS 2016, June 26-29, 2016, Okayama, Japan

text with vector space model, conduct feature selection, train DBN classifier, and classify the new text.

A. Text Preprocessing

In text classification systems, firstly, it is usually to do text preprocessing and reduce the data noise. Preprocessing has the following steps:

- Do Chinese word segmentation.
- Do stop word filter and stop word generally contains function words and widely used content words.
- Compute term frequency, inverse document frequency and inverse category frequency.

After the above work, we can represent the text in the form of feature vector. Structured text data is convenient to do subsequent processing of the text.

B. Original Feature Selection

Feature selection is a key process in text categorization. In this paper, we use TF-IDF-ICF [9] as feature selection method and the feature weight formula is as follows:

$$W(t_k,d)=\frac{tf(t_k,d)\times idf(t_k)\times icf(t_k)}{\sqrt{\sum_{i=1}^{P}[tf(t_k,d)\times idf(t_k)\times icf(t_k)]^2}} \qquad (1)$$

Where, $tf(t_k,d)$ is the feature t_k frequency in the document d, the $idf(t_k)$ is a function involved with inverse document frequency, the $icf(t_k)$ is a function involved with inverse category frequency, P is the number of features and the denominator is the normalized factor.

Select top 50000 words as the original features by feature selection method TFIDF-ICF. Then this paper use the deep belief networks to further extract highly distinguishable feature and reduce the dimensionality of original feature space.

III. DBN MODELING

In this paper, we use deep belief networks to construct classification model. Deep belief networks can make full use of a large number of unlabeled data and a small amount of labeled data to classify text by extracting senior features from underlying features. Currently, there are a few or no studies about using deep belief networks to classify Chinese text.

A. RBM Theory

Restricted Boltzmann machine is a typical neural network, and the network is a bipartite graph [10]. Visible units are connected to hidden units. No connection between hidden units or the visible units. Hidden layer can obtain more abstract characteristics. In the RBM, visible or hidden unit has two states: "active" and "inactive", generally represented by 1 and 0. The most important advantages of RBM is that the activation state of each of hidden units are conditionally independent when given the state of visible units, and the activation state of each of visible units are conditionally independent when given the state of hidden units.

B. RBM Energy Model

RBM is an energy-based model, a RBM consists of n visible units and m hidden units, vector v and h represent the state of visible and hidden units respectively. Given a set of state (v,h), the energy of a RBM system is defined as:

$$E(v,h|\theta)=-\sum_{i=1}^{n}b_i v_i-\sum_{j=1}^{m}c_j h_j-\sum_{i=1}^{n}\sum_{j=1}^{m}v_i W_{ij} h_j \qquad (2)$$

Where v_i represents the state of i-th visible unit, h_j represents the state of j-th hidden unit. $\theta=(W_{ij}, b_i, c_j)$ represents all the parameters of RBM. W_{ij} represents connection weights of visible units and hidden units. There is bias bi for each visible unit and bias c_j for each hidden unit. When the parameter determined, the joint and probability distributions based on energy are defined as:

$$P(v,h|\theta)=\frac{e^{-E(v,h|\theta)}}{Z(\theta)} \qquad (3)$$

$$Z(\theta)=\sum_{v,h}e^{-E(v,h|\theta)} \qquad (4)$$

Where $Z(\theta)$ is the normalization factor (also known as the partition function).When given the state of visible units, the activation state of each of hidden units are conditionally independent. At this point, the activation probability of j-th hidden units is defined as:

$$P(h_j|v,\theta)=\sigma(c_j+\sum_i W_{ij}v_i) \qquad (5)$$

Where $\sigma(x)=1/(1+e^{-x})$ is the sigmoid activation function. When given the state of hidden units, the activation state of each of visible units is also conditionally independent. The activation probability of i-th visible units is defined as:

$$P(v_i|h,\theta)=\sigma(b_i+\sum_j W_{ij}h_j) \qquad (6)$$

C. DBN Training

Next we describe how to train an RBM and how it is used in the construction of a DBN. First of all we must emphasize that RBM training is unsupervised. By maximizing log-likelihood of observation data (T samples of training set), we obtain RBM parameter θ.

$$\theta^*=argmaxL(\theta)=argmax\sum_{t=1}^{T}\log P(v^{(t)}|\theta) \qquad (7)$$

Hinton presented a fast learning algorithm that called contrastive divergence [11]. Contrastive divergence is an approximation of the log-likelihood gradient that has been found to be a successful update rule for training RBM. The procedure of contrastive divergence is showed as follow:

- Initialize the state of visible units with training sample.
- Compute the state of the hidden units according to the conditional distribution specified in equation (5).
- Reconstruct the state of the visible units according to the conditional distribution specified in equation (6).
- Reconstruct the state of the hidden units according to

the conditional distribution specified in equation (5).

- Update the parameters of RBM with stochastic gradient descent as formula (8) to (10) below.

$$\Delta W_{ij} = \eta(< v_i h_j >_{data} - < v_i h_j >_{recon}) \quad (8)$$

$$\Delta a_i = \eta(< v_i >_{data} - < v_i >_{recon}) \quad (9)$$

$$\Delta b_i = \eta(< h_j >_{data} - < h_j >_{recon}) \quad (10)$$

Where $\eta>0$ is the learning rate and $<v_i h_j>_{data}$ denotes the fraction of times that visible unit i and hidden unit j are on together when the original data is propagated through the RBM. Similar is the meaning of the notation for the rest of the parameters' update formulas. We can repeat this method for a defined number of epochs or until the reconstruction error of the original data becomes small, i.e. the confabulation is very similar to the data.DBN is a deep neural network that consists of a sequence of RBMs and a layer of back propagation network [12-13].

Figure 2. DBN architecture with l hidden layers.

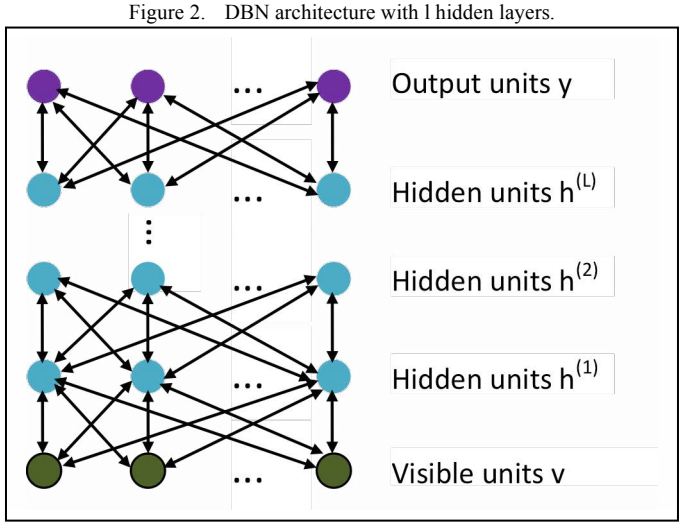
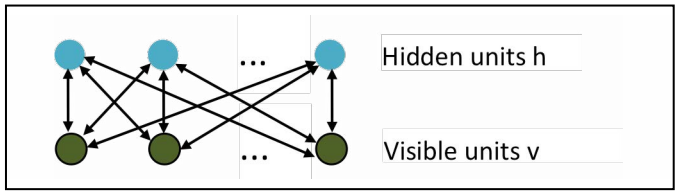


Figure 3. RBM architecture.



The learning process of training DBN model can be classifyed into two stages: pre-training and fine-tuning. More on this in the next section. DBN and RBM architecture are shown in Fig. 2 and Fig. 3.

D. Network Tuning

Pre-training stage uses a greedy layer-wise learning procedure to train each of RBMs [14]. Two adjacent layers are regarded as a RBM in the network. Train RBM layer by layer, output data of lower layer server as input data of higher layer in the network and initialize network parameter of DBN. Train RBM layer by layer can only make the parameters of this layer to achieve optimal and the errors of the lower RBM will be passed to the higher RBM. So the whole network is not optimal. But the errors will be amended at fine-tuning stage. At fine-tuning stage, the output data of DBN is regarded as the input data of softmax regression classifier and so far a whole neural network has been constructed. Next use global supervised back propagation algorithm to further optimize and adapt correlated parameters of network in a top-down direction. Finally a complete and optimal network of DBN has been built.

IV. EXPERIMENTAL EVALUATION

A. Testing Corpora

Data set of text classification is the premise and basis for this experiment, we collected the open text classification data sets to verify the feasibility of text categorization based on deep belief networks. But there is not a standard corpus for Chinese text categorization. In this paper we adopted the Chinese categorization Sogou corpus after analysis and comparison [15]. We randomly selected 6 categories from the corpus and deleted some error documents. Finally the corpus contains 6 categories and 10800 documents. The document number of 6 categories is same, and the ratio of pre-training set, validation set and test set of each category is about 4:1:1. Detailed distribution of data set is shown in Tab. 1.

TABLE I. DISTRIBUTION OF SOGOU DATA SET

Category	Pre-training set	Validation set	Test set
Sports	1200	300	300
Health	1200	300	300
IT	1200	300	300
Culture	1200	300	300
Military	1200	300	300
Finance	1200	300	300

B. Performance Measures

For evaluating the performance of a text classifier, the standard measures – precision, recall and F1, as well as those used in conventional information retrieval, is used. From the perspective of probability, precision is defined as the conditional probability that given a category c, the probability that assign the category to a test document d is correct. The recall is also defined as a conditional probability that if d ought to be assigned c, this decision is taken [16]. Given the contingency table of category C_i as shown in Tab. 2. In this table, a is the number of documents correctly assigned to C_i, b is the number of documents incorrectly assigned to C_i, c is the number of documents incorrectly rejected by C_i, d is the number of documents corrected rejected by C_i. The precision (P_i), recall (R_i), and F1 measure ($F1_i$) of category C_i are calculated as follows:

$$P_i = \frac{a}{a+b}, R_i = \frac{a}{a+c}, F1_i = \frac{2 \times P_i \times R_i}{P_i + R_i} \quad (11)$$

TABLE II. THE CONTINGENCY TABLE FOR CATEGORY CI

Category c_i		Expert Judgement	
		Yes	*No*
Classifier Judgement	*Yes*	a	b
	No	c	d

C. Experimental Setting and Results Analysis

To evaluate and compare the performance of DBN classification algorithm and SVM classification algorithm, we conducted some experiments on the same datasets. For DBN, it consists of 4 layers, the unit number of each layer are 20000-10000-2500-6, it runs for 1000 pre-training epochs, and we use an unsupervised learning rate of 0.01, with supervised learning rate of 0.1. For SVM, we use the LIBSVM tool, a library for support vector machines(http://www.csie.ntu.edu.tw/~cjlin/lib-svm) [17].

In the procedure of performing experiments, we set different value for two classifiers parameters. Finally we choose the best ones from these results. The best classification results are shown in Fig. 4-6. Fig. 4 shows the recall of DBN and SVM. We see that the recall of DBN classification algorithm is higher and more stable than that of SVM algorithm. Fig. 5 and Fig. 6 show the precision and f1score of DBN and SVM. We also see that the precision and f1score of DBN classification algorithm is significantly higher than that of SVM algorithm. From the results, we find that DBN can extract abstract features which can greatly improve the performance of the classifier. So the results show that DBN is an accurate and efficient classification algorithm.

V. CONCLUSION AND DISCUSSION

In this paper we focused on the use of Deep Belief Networks for text categorization. We studied the details of DBN training, evaluated the performance of our approach on Sogou corpus, and compared the effeteness of categorization between DBN classification algorithm and SVM classification algorithm. Experimental results show that the performance of DBN classification algorithm is significantly better than that of SVM algorithm.

At the same time we know that there are still some problems need to solve. Next, we will take the change of feature dimensions into consideration and expand the data sets to obtain more accurate and valuable results. In addition, optimizing the structure parameters of deep belief network to improve its performance is future work.

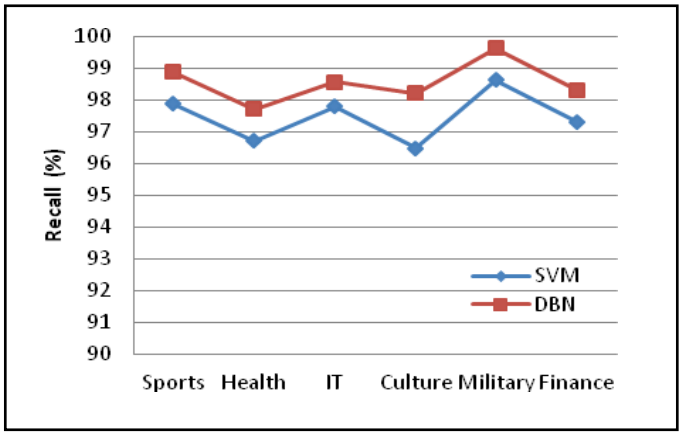


Figure 4. The Recall of SVM and DBN.

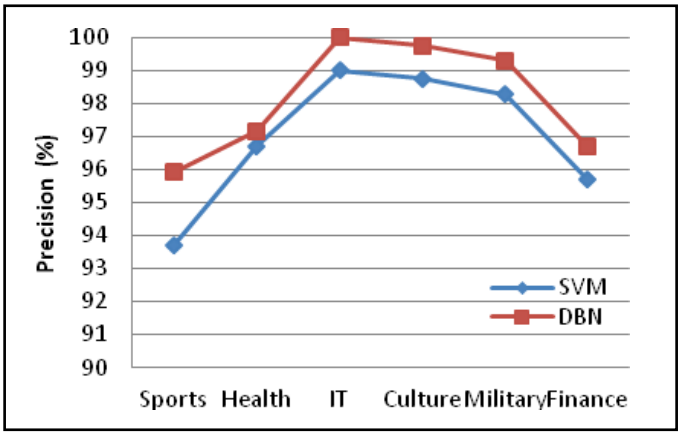


Figure 5. The Precision of SVM and DBN.

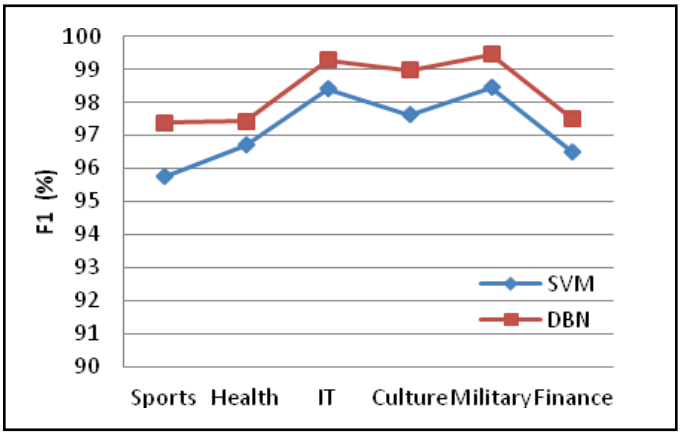


Figure 6. The F1 of SVM and DBN.

ACKNOWLEDGMENT

The work was supported by the project of National Key Te chnology R&D Program (2014BAK10B01).

REFERENCES

[1] A. R. Mohamed, G. Dahl and G. E. Hinton, "Deep belief networks for phone recognition", NIPS 22 workshop on deep learning for speech recognition, 2009.

[2] I. Goodfellow, Q. Le, A. Saxe and A.Ng, "Measuring invariances in deep networks", Advances in Neural Information Processing Systems, 2009, vol. 22, pp. 646-654.

[3] A. K. Noulas and BJ.A. Krose, "Deep Belief Networks for Dimensionality Reduction", Belgian-Dutch Conference on Artificial Intelligence 2008, Netherland, 2008.

[4] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[5] Yuan A, Bai G, Yang P, et al. Handwritten English Word Recognition Based on Convolutional Neural Networks[C]//IEEE International Conference on Frontiers in Handwriting Recognition (ICFHR), 2012: 207-212.

[6] Hannun A, Case C, Casper J, et al. DeepSpeech: Scaling up end-to-end speech recognition [J]. arXiv preprint arXiv: 1412.5567, 2014.

[7] G. E. Hinton, S. Osindero and Y.-W. Teh"A fast learning algorithm for deep belief nets"Neural computation, vol. 18, no. 7, pp. 1527-1554, 2006.

[8] Y. Bengio“Learning deep architectures for AI”Foundations and trends in Machine Learning, vol. 2, no. 1, pp. 1-127, 2009.

[9] J. Song, S. Qin and P. Zhang, An Improved Feature Weighting Strategy in Chinese Text Categorization, Proceeding of the International Conference on Manufacturing Science and Engineering (ICMSE 2015),2015,pp.202-208.

[10] C. Zhang, N. Ji, and G. Wang, Restricted Boltzmann Machines, Chinese Journal of Engineering Mathematics[J], 2015,(2):159-173.

[11] Hinton G .A practical guide to training restricted boltzmann machines [J]. Momentum, 2010,9(1):926-947.

[12] Ackley D H, Hinton G E, Sejnowski T J. A learning algorithm for Boltzmann machines*[J]. Cognitive science, 1985, 9(1): 147-169.

[13] Neukart F, Moraru S A. A Machine Learning Approach for Abstraction based on the Idea of Deep Belief Artificial Neural Networks[J]. ProcediaEngineering, 2014, 69: 1499-1508.

[14] Hinton G E, Dayan P, Frey B J, et al. The "wake-sleep" algorithm for unsupervised neural networks[J]. Science, 1995, 268(5214): 1158-1161.

[15] Information on http://www.sogou.com/labs/dl/c.html.

[16] F. Sebastiani, Machine learning in automated text categorization, ACM Computing Surveys, vol. 34, no.1, pp.1-47 2002.

[17] R.E.Fan, K.W.Chang, C.J.Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A Library for Large Linear Classification, Journal of Machine Learning Research 9(2008), 1871-1874.Software available at http://www.csie.ntu.edu.tw/~cjlin/liblinear.