Text Retrieval analysis based on Deep Learning

Kai LIU[a], Limin Zhang[b], Yongwei Sun[c]

Dept. of Electronic and Information Engineering, Naval Aeronautical and Astronautical Univ., Yantai, Shangdong, China, 264001

[a]wendao_2008@163.com, [b]imzhanglimin@163.com, [c]wendao_2012@126.com

Keywords: Text Retrieval; Replicate Softmax Models; Deep Learning; Deep Boltzmann Machines

Abstract. In view of the advantages of deep learning model in the extraction of abstract concept, a new text clustering algorithm is designed based on Deep Boltzmann Machines. Based on Replicate Softmax Model and new Deep Boltzmann Machine, energy function of this model is proposed and the detail learning algorithm is introduced. The learning can be made more efficient by using a layer-by-layer "pre-training" phase that allows variation inference to be initialized with a single bottom up pass. The values of the latent variables in the deepest layer are easy to infer and give a much better representation of each document than low learning. The 20-newsgroups document sets experiment results illustrated that the novel algorithm learn good generative models, get the better competence of a shallow model- Replicate Softmax Model in handling with an extract abstract concept and has good feasibility in large scale text clustering analysis.

Introduction

Text Clustering is mainly based on the famous retrieval hypothesis: the same kind of document similarity greater without the same kind of document similarity small [1]. As an unsupervised machine learning methods, retrieval does not require advance documentation manual annotation categories because of no training process, so it has a higher degree of flexibility and automation capabilities, text messaging has become effectively an important means of organization, summary and navigation, as more and more researchers are concerned [2].

Deep Learning networks are a new type of neural network that can discovers more important object features [3]. These networks determine features adept at learning high level abstractions about their datasets without supervision. It has been successfully applied in various application domains [4]. Given the development of deep learning model, based on a number of distribution and use of text analysis model -RSM (Replicated Softmax Model, RSM) [5], this paper designs based on its depth Boltzmann machine (Deep Boltzmann Machine, DBM) [6] for text feature extraction, combined with K-mean clustering algorithm to text clustering.

Background: DBMs and K-means

Deep Boltzmann Machines. A deep Boltzmann machine (DBM) is a probabilistic model consisting of many layers of random variables, most of which are latent binary units and also can be seen as a network of symmetrically coupled stochastic binary units. There are connections only between hidden units in adjacent layers. For example, we design a DBM in two hidden layers that contains a set of visible units $\mathbf{v} \in \{0,1\}^D$, and a sequence of layer of hidden units $\mathbf{h}^{(1)} \in \{0,1\}^{F_1}$, $\mathbf{h}^{(2)} \in \{0,1\}^{F_2}$, as shown in Fig.2. The energy of the joint configuration $\{\mathbf{v}, \mathbf{h}\}$ is defined as (ignoring bias terms):

$$E(\mathbf{v}, \mathbf{h}; \theta) = -\mathbf{v}^{\mathbf{T}}\mathbf{W}^{(1)}\mathbf{h}^{(1)} - \mathbf{h}^{(1)\mathbf{T}}\mathbf{W}^{(2)}\mathbf{h}^{(2)} \qquad (1)$$

Where $\mathbf{h} = \{\mathbf{h}^{(1)}, \mathbf{h}^{(2)}\}$ are the set of hidden units, and $\theta = \{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}\}$ are the model parameters, representing visible-to-hidden and hidden-to-hidden symmetric interaction terms. The probability

© 2015. The authors - Published by Atlantis Press

that the model assigns to a visible vector $\mathbf{v}$ is:

$$P(\mathbf{v};\theta)=\frac{P^{*}(\mathbf{v};\theta)}{Z(\theta)}=\frac{1}{Z(\theta)}\sum_{\mathbf{h}}\exp\left(-E\left(\mathbf{v},\mathbf{h}^{1},\mathbf{h}^{2};\theta\right)\right) \tag{2}$$

Where $Z(\theta)=\sum_{\mathbf{v},\mathbf{h}}\exp\left(-E\left(\mathbf{v},\mathbf{h}^{1},\mathbf{h}^{2};\theta\right)\right)$ represents the partition function.

We use a greedy layer-wise pre-training strategy to initialize the model parameters to good value and employ Stochastic Approximation to fine-tune the parameters after the previous stage.

Replicated Softmax Model. The Replicated Softmax Model is useful for modeling word count vectors in a document. Let $\mathbf{v}\in\mathbb{N}^{K}$ be a vector of visible units where v_k is the number of times word k occurs in the document with the vocabulary of size K and $\mathbf{h}\in\{0,1\}^{F}$ seemed as be binary stochastic hidden topic features. The energy of the state $\{\mathbf{v},\mathbf{h}\}$ is defined as follows

$$E(v,h;\theta)=-\sum_{k=1}^{K}\sum_{j=1}^{F}v_k W_{kj}h_j-\sum_{k=1}^{K}b_k v_k-M\sum_{j=1}^{F}a_j h_j \tag{3}$$

Where $\theta=\{\mathbf{a},\mathbf{b},\mathbf{W}\}$ are the model parameters and $M=\sum_k v_k$ is the total number of words in a document.

K-means. K-means clustering [7, 8] is the most famous division clustering algorithm and popular for cluster analysis in data mining. Given a set of data points and the required number of clusters K, this algorithm aims to partition into K clusters in which each observation belongs to the nearest mean. We first select K objects as the initial cluster centers randomly, then calculate the distance of each object and each seed cluster centers, assign each object to its distance from the nearest cluster center. Once all objects have been assigned, the center of each cluster will be recalculated based on the existing cluster objects. This process is repeated until a termination condition is satisfied. The termination condition may be any of the following:

(1) No (or minimal number) of an object to be re-assigned to different clusters.

(2) There is no (or minimal number) of the cluster centers vary.

(3) Local minimum squared error.

Algorithm Design

Through the introduction of K-means and mechanisms for DBM structure analysis, we exchange the lowest unit for RSM visible cells so to constitute Deep Boltzmann machine model based on Replicated Softmax Model, the principle mechanism for this model are as follows, we make the underlying characteristics as the top RBM input, to enhance the ability of the DBM feature extraction by fine-tuning parameters and weights biased to achieve automatic feature extraction purposes. In a given sample $\{v^1,v^2,...,v^K\}$, the probability of the entire model, such as follows

$$P(\mathbf{v};\theta)=\sum_{\mathbf{h}^{(1)},\mathbf{h}^{(2)}}P\left(\mathbf{h}^{(1)},\mathbf{h}^{(2)}\right)\left(\sum_{\mathbf{h}^{(1)}}P\left(\mathbf{v},\mathbf{h}^{(1)}\right)\right) \tag{4}$$

Since the lowest unit and the upper unit is connected only to the intermediate layer unit, so we list the activation of the posterior probability of the intermediate layer unit as follows

$$P\left(h_j^{(1)}=1\,|\,\mathbf{v},\mathbf{h}^{(2)}\right)=sigm\left(\sum_{i=1}^{D}\sum_{k=1}^{K}v_i^k W_{ij}^k+\sum_{m}W_{jm}^{(2)}h_j^{(2)}\right) \tag{5}$$

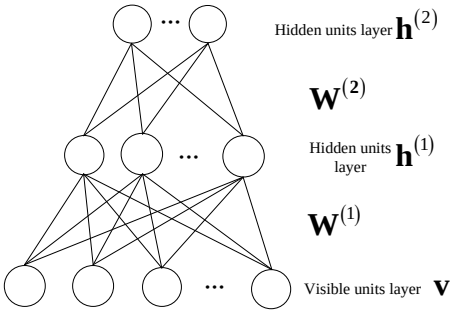

Fig.1 The structure of Deep Boltzmann Machine for text clustering

We show the structure of Deep Boltzmann Machine used for text clustering as Fig. 1. After the extraction of highest feature, we cluster the set into K group based on K-mean clustering Algorithm.

Experiments

20-newgroups document set [9, 10] is a popular dataset in the community for text retrieval and is comprised of 18,845 testing documents. This entire document collection is divided into 20 different themes newsgroups. For test the performance of this retrieval model, we further randomly split the training set into 11,314 training and 7,531 validation documents. To speed up learning, we divided datasets into mini-batches, each containing 100 cases and updated the weights after each min-batch.

Experimental preparation steps: First, we remove the text stop words and no words, then extract the maximum information gain before 5000 and integrate it as a dictionary word database, transform each text into a one-dimensional vector according to the dictionary in the library vocabulary finally.

In this experiment, DBM is composed by two levels, RSM for intermediate extraction and RBM for extraction of more abstract concept. According to the experimental preparation steps, we set the dictionary vocabulary $K = 5000$, all RBM learning rate as $\eta = 0.01$, and the CD epoch is no more than 2000 cycles. After the extraction of text feature, we retrieval the feature set according to K-means retrieval approach for text classification. We used text retrieval indicators such as recall - precision curves (RPC) as a measure of the effectiveness of document retrieval.

We use desktop which installed Matlab (2013a) and clocked at 2.4GHz as experimental platform. This experiment is designed to test the effectively of Deep Boltzmann machine used in text retrieval. So we design three models for text retrieval analysis, which include RSM and DBM, as shown below

(1) Direct K-means retrieval, feature dimension for 5000, K value is 10.

(2) RSM-K-means retrieval, the structure of 5000-300, the feature dimension is 300, K value is 10.

(3) DBM-K-means retrieval, structure 5000-200-100 feature dimension is 300, K value is 10.

Results. We show the results by Fig.2 and Table 1, which is the RPC for text retrieval and the retrieval error of 20-newgroups document set.

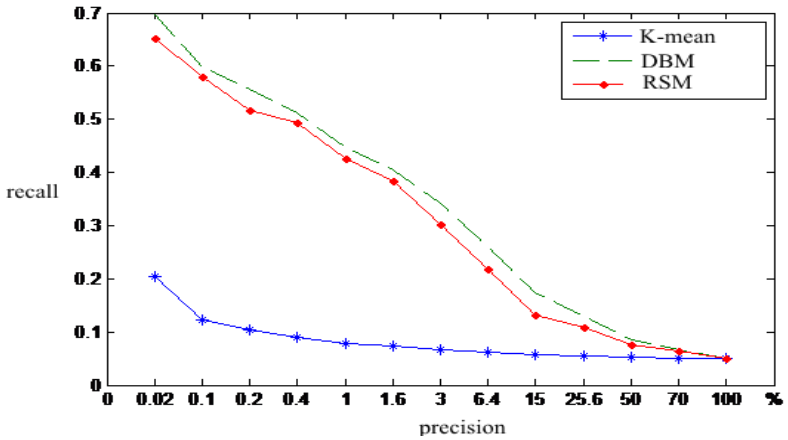

Fig.2 The RPC of different models in 20-newgroups document set

Fig.2 shows that through the highest feature based on Deep Boltzmann Machines, we can get more accurate feature for text retrieval and better than the shallow model RSM. This result illustrate Deep Boltzmann Machines is more adapt for text feature extraction.

Table 1 20-newgroups retrieval error rate

model	retrieval error rate	size of feature vector
K-mean	20.20%	5000
RSM-K-mean	6.62%	300
DBM-K-mean	3.07%	300

Table 1 shows the value of the error in three different retrieval models. As can be seen from the table, DBM-K-mean gets the best performance in three models, and decreases the retrieval error at least a 3%. These also prove the effectiveness of the proposed algorithm and illustrate the advantages of deep learning model in dealing with text features and text retrieval.

Conclusions

This paper presents a Deep Boltzmann machine text retrieval approach and its effect experimentally. Through tested with shallow model RSM on 20-newgroups documentation set, this new retrieval method is more effective. We find this text extraction mechanism based on Deep Boltzmann machine can not only improve the accuracy of feature representation, and more abstract, may be suitable for large-scale text analysis, classification study. Future we should continue to study how to improve training efficiency deep learning model and try more areas through deep learning model.

References

[1] R. Salakhutdinov, G. E. Hinton. Semantic hashing [J]. International Journal of Approximate Reasoning, 2009 50 (7) 969–978.

[2] Andriy Mnih, G E. Hinton. A Scalable Hierarchical Distributed Language Model [C]. Vancouver: In Advances in Neural Information Processing System, 2008: 1081-1088.

[3] Asja Fischer, Christian Igel. An Introduction to Restricted Boltzmann Machines [C]. Vancouver: In Iberoamerican Congress on Pattern Recognition, 2012: 14-36.

[4] G. E. Hinton, R. Salakhutdinov. Reducing the dimensionality of data with neural networks [J]. Science, 2006 313 (3) 504–507.

[5] Ruslan Salakhutdinov, G.E.Hinton. Replicated Softmax: an Undirected Topic Model [C]. Vancouver: Advances in neural information processing systems. 2009: 1607-1614.

[6] R. Salakhutdinov, G. E. Hinton. Deep Boltzmann machines [C]. New York: In Proceedings of the International Conference on Artificial Intelligence and Statistics, 2009: 14-36.

[7] C. W. Chen, J. Luo, K. J. Parker. Image segmentation via adaptive K-mean clustering and knowledge-based morphological operations with biomedical applications [J]. Image Processing, IEEE Transactions on, 1998 7(12) 1673-1683.

[8] S. Tokdar, R. Kass. Importance sampling: a review Wiley Interdisciplinary Reviews [J]. Computational Statistics, 2010 (1) 54-60.

[9] J. Yang, Y. Liu, X. Zhu, et al. A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization [J]. Information Processing & Management, 2012 48(4) 741-754.

[10] Kumar M A, Gopal M. A comparison study on multiple binary-class SVM methods for unlabel text categorization [J]. Pattern Recognition Letters, 2010 31(11) 1437-1444.