Unsupervised Attention Embedding for Document Clustering

Ji-kang NIE and Zhi-guo ZHANG*

School of Data and Computer Science, Sun Yat-Sen University, Guangzhou 510006, China

*Corresponding author

Keywords: Document clustering, Attention embedding, Deep clustering.

Abstract. Deep clustering algorithms perform learning feature representations and clustering tasks jointly by using neural networks with significantly improved performance over the traditional k-means or spectral clustering. Some groundbreaking proposals extract data spaces directly in “bags of words” approach without considering the semantic information of each document as inputs to deep auto-encoder networks. But these algorithms suffer from inaccurate feature space from the encoder output when dealing with incomprehensible and high-dimensional data. For solving this problem in this paper, an Attention-based Deep Embedded Clustering (ADEC) algorithm is proposed to improve representation of data space. ADEC extracts high quality embedded features and performs clustering jointly with learning embedded features which are suitable for document clustering. The experimental result shows that the performance and accuracy of document clustering is improved significantly using the ADEC clustering framework on two datasets REUTERS-10K and REUTERS.

Introduction

Clustering is a fundamental technique widely used in pattern recognition and machine learning. Traditional clustering algorithms like k-means\cite{1}, Gaussian mixture model(GMM)\cite{2} and spectral clustering\cite{3} group similar data into the same cluster through measuring handcrafted features according to similarity. However, the performance of traditional clustering algorithms are adversely affected when dealing with incomprehensible and high-dimensional data as a result of unreliable similarity metrics extracted with vector space model (VSM)\cite{4} approach. Tackling understanding the meaning and intention of the incomprehensible data, some techniques often combine k-means and Latent Dirichlet Allocation (LDA)\cite{5} to obtain embedded data space which can capture latent semantic information from a large collection of documents. Tackling the dimensional interference, those techniques like Principle Component Analysis (PCA)\cite{6} often initialize raw data into a low-dimensional space, and then cluster the embedded data space.

However, existing clustering approaches suffer from three main issues when dealing with text: (1) using meaningless data space, which lead to a reduced density of similarity between documents and fail to compress all the necessary information of a source document into a fixed-length data space. The distribution of data is not accurately represented; (2) using shallow and linear embedding functions, which cannot capture the non-linear nature of the raw data; (3) lacking of jointly embedding with clustering procedures, which fail to get an optimal embedded subspace for clustering.

In order to solve the problems above, we propose a new algorithm for document clustering, called Attention-based Deep Embedded Clustering (ADEC), which combines the advantages of both clustering methods and deep embedding models. ADEC generally consists of two main components: attention embedding network and feed-forward clustering network. Particularly, we use pre-trained word vectors into the attention embedding network to ensure the quality of the word embedding. Our experimental results show that ADEC achieves better accuracy (ACC) and normalized mutual information (NMI) evaluation compared to many state-of-the-art algorithms on the document benchmark datasets.

The contributions of our work are: (a) using attention embedding to obtain a precise data space which is approximate to raw data distribution; (b) providing a non-linear embedding subspace via the
attention embedding network and optimizing deep embedding and clustering jointly via the clustering
network; (c) achieving better clustering results on incomprehensible and high-dimensional datasets.

Related Works

The performance of document clustering is highly dependent on the quality of the internal document
representations. Those document representations are generated from the document’s text by a
so-called embedding algorithm, which “embeds” the documents as points into a space with a
particular fixed dimensionality.

Traditionally, bag-of-words (BOW) models encode document based on a frequency distribution
(TFIDF algorithm). TFIDF algorithm shares a common problem that each word is only represented
with one vector. The vector clearly fails to capture homonymy and polysemy. For large and complex
document collections, BOW models perform not as good as expected. LDA is a popular probabilistic
clustering method to reduce vector space and captures the latent semantic information from a large
collection of documents.

In recent years, new lights for document clustering have come from deep learning based on deep
neural network. Bengio et al\cite{7} first propose a neural language model for distributed document
representation. Inspired by this pioneering work, many word embedding models\cite{8,9} have been
devised to represent words as real space vectors learned by predicting a probability distribution over
each word, given words used in its context information. Attention mechanism have recently gained
popularity in training neural networks, which aims to focus on the most pertinent information. He et
al\cite{10} propose a sentence level attention embedding. They filter the pre-trained word embeddings
within a sentence using an attention mechanism and use the filtered words to construct aspect
embeddings. Lin et al\cite{11} uses a new self-attention mechanism that allows it to extract different aspects
of the sentence into multiple vector-representations. These models are much more generalized
than BOW models. These models learn similar vectors for words of similar meanings.

In deep clustering, the existing algorithms are divided into two categories: (1) learning a
representation first and then applying clustering; and (2) jointly optimizing feature learning and
clustering. The previous category directly takes advantage of existing unsupervised deep learning
algorithms. Tian et al\cite{12} learn a non-linear embedding of the affinity graph using a stacked
auto-encoder, and then obtain the clusters in the embedding subspace via K-means, and then use
k-means to cluster the learned representations. Chen et al\cite{13} train a Deep Belief Network (DBN) first
and then apply non-parametric maximum-margin clustering to learned intermediate representation.
Song et al\cite{14} utilize the deep auto-encoder to obtain good representations by minimizing the data
reconstruction error, and then use k-means to cluster them. However, these methods cannot provide
joint embedding and clustering procedures so that they fail to get an optimal embedding subspace for
document clustering.

The latter category defines a clustering loss and simulating classification error. Employing
de-noising stacked auto-encoder learning approach, Xie et al\cite{15} first pre-trained the model
layer-wisely and then fine-tuned the stacked encoder pathway by a clustering algorithm using
Kullback-Leibler divergence minimization. Guo et al\cite{16} propose the Improved Deep Embedded
Clustering (IDEC) algorithm to take care of data structure preservation. Different with their work, our
ADEC algorithm aims to learn meaningful representation by applying attention mechanism upon the
word embedding layer. Our reconstruction target is document embedding other than VSM raw
document space.

Proposed Framework

Our proposed framework consists of two components: (1) Attention Embedding Network provides a
non-linear and attention mapping function through learning encoder and decoder, where the encoder
is an attention embedding to be trained, and the decoder is required to be capable to reconstruct the
document embedding from those features generated by the encoder; (2) Feed-forward clustering network enables that a distance between different kinds of documents in a space is further distanced and the distance between similar documents is made closer to improve the accuracy of clustering.

**Attention Embedding Network** includes an encoder and a decoder.

**Encoder.** We convert each input document $d_i$ into a vector representation $v_{d_i}$ at first. Then we construct the document $v_{d_i}$ embedding as the weighted summation of word embedding $e_{w_i}$, where $i$ is the word indexes in the document.

$$v_{d_i} = \sum_{j=1}^{n} w_{i_j} e_{w_j}$$  \hspace{1cm} (1)

We compute a weight $w_{i_j}$ for each word $w_j$ in the document in order to capture the most relevant information of the document, where $w_{i_j}$ is regard as the probability that $w_j$ is concerned with the right word. Based on an attention mechanism, $w_{i_j}$ is calculated on the global context of the document and the embedding of the word $e_{w_j}$:

$$w_{i_j} = \frac{\exp(D_{i_j})}{\sum_{j=1}^{n} \exp(D_j)}$$  \hspace{1cm} (2)

$$D_{i_j} = e_{w_i}^T \cdot N \cdot y_{d_i}$$  \hspace{1cm} (3)

$$y_{d_i} = \frac{1}{n} \sum_{i=1}^{n} e_{w_i}$$  \hspace{1cm} (4)

where $y_{d_i}$ is the average of the word embeddings, $N \in R^{D \times D}$ is a matrix mapping between the global context embedding $y_{d_i}$ and word embedding $e_{w_i}$. And both will be learned in the training process.

![Figure 1. The network structure of ADEC. It consists of two components: Attention Embedding Network (left) and Feed-forward Clustering Network (right).](image)

We reduce $v_{d_i}$ from D dimensions to a lower dimensions K and then apply a softmax non-linearity to produce normalized non-negative weights to obtain $R_i$:

$$R_i = \text{softmax}(W \cdot v_{d_i} + b)$$  \hspace{1cm} (5)

where $W$ is the parameter of weighted matrix, $b$ is the bias vector. They are both need to be learned in the training process. $R_i$ is the attention-based embedded point which we think can represent the most relevant information of the document.
**Decoder.** As shown on the left of Figure 1. The \( v_{id} \) reconstruction process consists of two steps of transitions. We have completed the first step in the decoder part and obtained \( R_t \), now we just regard the reconstructed vector representation \( u_{id} \) as a linear combination of attention embeddings from \( \mathcal{T} \):

\[
u_{id} = T^T \cdot R_t
\]

**Feed-forward Clustering Network** concerns the distances between different kinds of documents. After the process of attention embedding network, we remove the decoder and optimize attention embedding via encoder output jointly with clustering. The objective function of clustering network is to minimize the difference between the soft assignment \( q_{ij} \) and the auxiliary target distribution \( p_{ij} \) using KL divergence.

\[
L = KL(P \parallel Q) = \sum_{i} \sum_{j} p_{ij} \log \frac{p_{ij}}{q_{ij}}
\]

where \( q_{ij} \) can be interpreted as the similarity between the attention-based embedded point \( R_t \) and the cluster centroid \( \mu_j \):

\[
q_{ij} = \frac{(1 + \| R_t - \mu_j \| ^2)^{-1}}{\sum_j (1 + \| R_t + \mu_j \| ^2)^{-1}}
\]

\( q_{ij} \) are soft assignments, which can be seen as the probability of assigning the attention-based embedded \( R_t \) to the cluster \( j \):

\[
p_{ij} = \frac{q_{ij}^2}{\sum_j (q_{ij}^2 / \sum_i q_{ij})}
\]

As we can see, \( Q \) defines the target distribution \( P \), and \( q_{ij} \) can be refined repeatedly by \( p_{ij} \) to increase clustering coupling in the training process.

**Experiment**

We evaluate the proposed method (ADEC) on Reuters and Reuters-10K. Reuters contains about 810000 English news stories labeled with a category tree. We use the four root categories: economics, corporate/industrial, government/social, and markets, as labels. And we further prune all documents labeled by multiple root categories to get 685071 articles before we compute attention embedding features. Since some algorithms do not scale to the full Reuters dataset, we also sample a random subset of 10000 examples, which is referred as to REUTERS-10k, for comparison purposes.

**Comparing Methods.** In order to evaluate the performance of our clustering algorithm in dealing with incomprehensible and high dimensional data, we compare the clustering results of ADEC with its competing algorithms: TFIDF+IDEC, TFIDF+DEC, attention embedding network+k-means, TFIDF+auto-encoder+k-means, and TFIDF+k-means.

**Parameters Setting.** For the sake of completeness, traditional and classic clustering algorithm, k-means is also included in comparison. k-means runs 20 times with different initialization and the result with best objective value is chosen. We also use the publicly available code released by the author to report the performance of DEC. For the ADEC model, we initialize the word embedding matrix with word vectors trained by word2vec with negative sampling on each dataset. And we set the embedding size to 300, window size to 10, and negative sample size to 5. During the training process, we fix the word embedding matrix and optimize other parameters using Adam. After pre-training, the coefficient \( \gamma \) of clustering loss is set to 0.1 and batch size to 256 for all datasets, the clustering
network uses SGD as optimizer with learning rate $\lambda=0.1$ and momentum $\beta=0.99$. The convergence threshold is set to $\delta=0.1\%$.

**Evaluation Metric.** Generally, we use two of the most popular evaluation criteria widely used for unsupervised learning clustering algorithms which are accuracy (ACC) and normalized mutual information (NMI).

We report the result of all compared algorithms on two datasets in Table 1. From Table 1, we can see that our proposed algorithm outperforms all other methods in the two datasets. In most cases, ADEC significantly improves the performance of the document clustering compared with traditional methods. ADEC also performs better than the current state-of-art deep clustering algorithms.

| models                        | datasets       | REUTERS-10k | REUTERS     |
|-------------------------------|----------------|-------------|-------------|
|                               |                | ACC         | NMI         | ACC          | NMI          |
| TFIDF+k-means                 |                | 52.42       | 31.83       | 53.29        | 32.51        |
| TFIDF+auto-encoder+k-means    |                | 70.52       | 39.19       | 71.84        | 41.63        |
| attention embedding network+k-means |             | 76.65       | 45.74       | 78.29        | 47.16        |
| TFIDF+DEC                     |                | 73.68       | 49.76       | 76.72        | 52.37        |
| TFIDF+IDEC                    |                | 75.64       | 49.81       | 78.86        | 53.85        |
| ADEC                          |                | **80.07**   | **58.43**   | **82.18**    | **59.94**    |

According to the performance ACC and NMI in the two datasets. We find that the larger the data set, the better the performance of the algorithm. Clustering based on our ADEC algorithm achieves evident improvement than traditional TFIDF+k-means clustering in ACC and NMI. At the same time, our proposed method has better performance compared to current state-of-art methods.

Our attention embedding network introduces an analogous training process like the auto-encoder in DEC. Unlike the pre-training category of DEC, we apply attention mechanism upon the word embedding layer to obtain a document embedding. The attention mechanism can capture the global information of the document as our encoder. Our decoder reconstruction target is document embedding other than VSM raw document space. The experimental result shows that the decoder is effective in the design of document embedding reconstruction. ADEC achieves a higher quality low-dimensional feature space by different pre-training strategies than DEC. Particularly, ADEC algorithm gains a significant improvement in the NMI evaluation. The reason is that the attention embeddings of encoder output enhance the mutual dependence between documents. The implementation of our clustering network, which is same as DEC, performs the fine-tuning process on the output of the encoder jointly with the clustering centroid.

**Conclusion and Future Work**

In this paper, we propose a new document clustering method named Attention-based Deep Embedded Clustering (ADEC) algorithm. ADEC applies attention mechanism in mapping high quality embedding features and performs clustering jointly with learning embedded features that are suitable for document clustering.

Generally, ADEC consists of an attention embedding network and a clustering network. Attention embedding network has two main functions: (1) providing attention mechanism to capture the global information of the document for better approximation to the raw data distribution than traditional VSM approach; (2) providing non-linear mapping function through an variant auto-encoder, where the encoder is a attention mapping function to be trained and the decoder is required to be capable to reconstruct document embedding from those features generated by the encoder. Clustering network runs by optimizing clustering loss based on KL divergence with a self-training target distribution. Experimental result shows that ADEC performs with high performance and accuracy for document clustering tasks.
In the future, we will try more ways to reconstruct the document embedding effectively and apply a better loss function in our clustering network training process.

Acknowledgement
This research was financially supported by the National Key Research and Development Program of China (2017YFB0202400, 2017YFB0202403).

References
[1] MacQueen J. Some methods for classification and analysis of multivariate observations[C]//Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. 1967, 1(14): 281-297.
[2] Bishop, Christopher M. Pattern recognition and machine learning. springer New York, 2006.
[3] Von Luxburg U. A tutorial on spectral clustering[J]. Statistics and computing, 2007, 17(4): 395-416.
[4] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.
[5] Blei D. M., Ng A. Y., and Jordan M. I., “Latent Dirichlet Allocation,” Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.
[6] Wold S, Esbensen K, Geladi P. Principal component analysis[J]. Chemometrics and intelligent laboratory systems, 1987, 2(1-3): 37-52.
[7] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. Journal of machine learning research, 2003, 3(Feb): 1137-1155.
[8] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
[9] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.
[10] He R, Lee W S, Ng H T, et al. An unsupervised neural attention model for aspect extraction[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017, 1:388-397.
[11] Lin Z, Feng M, Santos C N, et al. A structured self-attentive sentence embedding[J]. arXiv preprint arXiv:1703.03130, 2017.
[12] Tian F, Gao B, Cui Q, et al. Learning deep representations for graph clustering[C]//AAAI. 2014: 1293-1299.
[13] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
[14] Song C, Liu F, Huang Y, et al. Auto-encoder based data clustering[C]//Iberoamerican Congress on Pattern Recognition. Springer, Berlin, Heidelberg, 2013: 117-124.
[15] Xie J, Girshick R, Farhadi A. Unsupervised deep embedding for clustering analysis[C]//International conference on machine learning. 2016: 478-487.
[16] Guo X, Gao L, Liu X, et al. Improved deep embedded clustering with local structure preservation[C]//International Joint Conference on Artificial Intelligence (IJCAI-17). 2017: 1753-1759.