Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1826–1832, DOI 10.12694/scpe.v25i3.2742

TEXT CLASSIFICATION AND CLUSTER ANALYSIS BASED ON DEEP LEARNING AND NATURAL LANGUAGE PROCESSING

HUA HUANG *

Abstract. At present, the commonly used Bag of Words (BOW) expression ignores the semantic information of text and the problems of high dimension and high sparsity of feature extraction. This paper presents a multi-class text representation and classification algorithm. This project is based on the vector expression of keywords and takes the multi-category classification problem as the research object. Then, a hybrid Deep Location network (HDBN) is constructed by combining DBN with Boltzmann (DBM). Then, this paper does a lot of tests on the algorithm and proves the effectiveness of the algorithm. In addition, the 2D visual experiment is carried out with HDBN, and then the high-level text expression based on HDBN is obtained. The expression has strong cohesion and weak coupling.

Key words: Text classification; Deep belief network; Deep learning; Deep Boltzmann machine network

1. Introduction. Under "information overload," managing and screening information effectively is an urgent problem. Text is the primary way for people to get information on the Internet. Using the method of word classification can solve various complicated problems well to help users find the information they need better [1]. Text must first be converted into a readable format to realize automatic recognition of text. Text expression is the most essential part of the whole text recognition process, and its correctness directly affects the whole system's performance. Most existing text expressions are based on lexical packages (BOW) and vector Spaces (VSM). The default words are independent of each other, and the correlation between semantics is ignored. However, due to the diversity of text types and the elaboration of topics, such shallow text expression lacks the semantic meaning of the text itself, and it is difficult to cope with the current complex classification problem. The continuous development of deep learning technology provides a new opportunity for the development of character recognition technology. The project research results in this field will provide new ideas and methods for large-scale data analysis [2]. Deep learning has been widely used in many problems, such as data compression, object detection and tracking, information retrieval, machine translation and speech recognition. Deep learning technology can better relate to specific questions to uncover the complex semantic connections hidden in the text [3]. At the same time, massive data training and processing capabilities have been greatly improved with the expansion of network scale and the rapid development of multimedia networks. This opens up new opportunities for deep learning.

2. Overview of text classification.

2.1. Concept Analysis. Text classification is classifying and marking a text according to a specific system and criteria [4]. The so-called "text characteristics" refers to words closely related to the text and can express the work.

2.2. Development of text classification. Word classification is a typical problem in natural language processing. In the early 1950s, character recognition research mainly used expert judgment. This requires human intervention [5]. This inevitably affects the efficiency of retrieval. With the rapid development of network technology since the 1980s, a large amount of text data has been used for processing. While statistical and computer-aided algorithms have emerged to deal with these problems, these algorithms still stay in the traditional manual processing and single modes. KNN, Naive Bayes, neural networks, decision trees, support

*School of Computer and Artificial Intelligence, Henan Finance University. Zhengzhou, Henan, 450046, China (Corresponding author, hafuhuanghua@163.com)

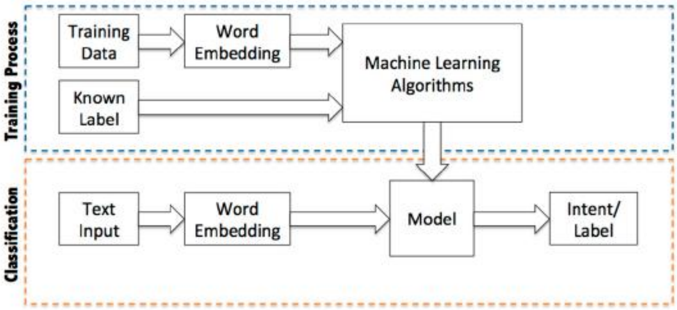

Fig. 2.1: Training flow of training text classifier.

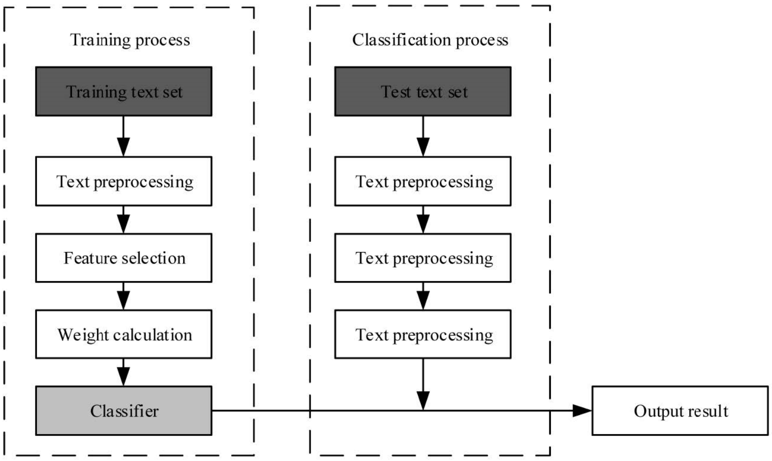

Fig. 2.2: Flow chart of text classification.

vector machines (SVM), etc. The problem is then decomposed into two main steps: one is featuring extraction, and the other is classifier design [6]. The process of training the text classifier is shown in Figure 2.1. The core of text classification lies in the selection of feature values and the design of algorithms, in which the feature quantities used include information gain, text frequency, mutual information, CHI ($\chi 2$) and so on. It is compared with the data to be tested to determine the text category. These algorithms are prone to problems such as small samples, local overfitting and local optimality, and need dimensionality reduction. This leads to the loss of text and the reduction of recognition accuracy.

2.3. Process of text classification. Text classification is divided into two stages: one is the learning stage, and the other is the classification stage. A machine learning algorithm based on a deep neural network is proposed. The text classification process is shown in Figure 2.2. The input text is standardized data. It is a computer-expressible form, that is, the text vector. A text-based automatic recognition method is proposed [7]. The recognition model is learned and trained to get the model parameters using the text training set method. Experimental results show that this method has good learning performance. An adaptive learning algorithm based on a neural network is proposed and dynamically adjusted to improve the learning effect of the classifier.

3. Natural language text processing model based on deep learning. Firstly, ICTCLAS software is used to slice the original text and eliminate invalid words to obtain the characters needed for the experiment. The traditional TF-IDF algorithm is used to solve the weight of each characteristic word [8]. Construct the original feature matrix of text. Assume that each text has n properties. In this way, an n-dimensional vector space is formed, and a characteristic vector of n-dimensions can represent each literal s:

$$U(s) = (R_1, E_1(s); R_2, E_2(s); \cdots\cdots; R_n, E_n(s)) \tag{3.1}$$

R_i is a segmented word of text. Where$E_i(s)$ is the weight of R_i in the text D. Use the TF-IDF formula to calculate the weight of text segmentation:

$$e_i(s) = \frac{TF(t_i) \times IDF(t_i)}{\sqrt{\sum\limits_{i=1}^{n} (TF(t_i) \times IDF(t_i))^2}} = \frac{TF(t_i) \times \log(\frac{N}{n_i} + D)}{\sqrt{\sum\limits_{i=1}^{n} (TF(t_i) \times \log(\frac{N}{n_i} + D))^2}} \tag{3.2}$$

where $e_i(s)$ is the weight of eigen term R_i. $TF(t_i)$is the frequency with which the eigen b is used in the s sentence. Where N represents the total number of samples. n_i is the number of samples that occur in R_i.

3.1. Automatic recognition of text. The existing SVM algorithm and BP neural network algorithm have significant differences in the recognition accuracy of different samples due to the interference of sampling data. Text recognition based on a deep confidence network can be divided into two stages: pre-training artificial neural network and network adjustment [9]. Most existing classification methods use dimensionality reduction to avoid dimensionality disaster, while deep belief networks (DBN) can extract low-dimensional features with strong discrimination ability from massive original features. In this way, the classification model can be built directly without dimensionality reduction. Meanwhile, it fully uses the rich information in the text. The weights of each BP neural network level are initialized using DBN network weights [10]. This method does not need to initialize any initial value of DBN, nor does it need to extend the BP neural network. BP neural network is used for global optimization to solve the local extreme value problem caused by DBN's randomness of weight parameters.

3.2. DBN Pre-Learning. A deep confidence Network (DNN) is a deep nonlinear network. The underlying information is fused by constructing the learning mode of multiple implicit levels. This creates more abstract high-level features to recognize text effectively. Suppose $F$is a system that includes n layer $(F_1, F_2, \cdots, F_n)$, if G is used to represent the input and P is used to represent the output. It can be expressed in $G \geq F_1 \geq F_2 \geq \cdots \geq F_n \geq P$ to continuously adjust the parameters in the system [11]. The result of the system is still input G, and then we can automatically obtain the hierarchical property of input G, which is $F_1, F_2, \cdots, F_n$. DBN is a probability-based modeling method that assigns observed samples and tags jointly. The DBN is formed by stacking layer upon layer of constrained Boltzmann machines (RBMS). RBM is a representative neural network (Figure 3.1 cited in Deep neural Networks (Part IV). Creating, training and testing a model of Neural Networks).

The RBM model is divided into two levels: one is the visual layer, usually the input layer, and the second is the implicit layer, usually called feature extraction [12]. Learning the neurons of the hidden layer in the visual-hidden layer can capture the higher-order association information presented by the video layer. Where is the weight of the visible layer and the hidden layer. is the displacement of the nodes of the visible layer. is the displacement of the node of the hidden layer. Where is the state vector of the node of the visual layer. Where is the state vector of the node of the hidden layer. In the process of BP network learning, the greedy algorithm is used for hierarchical learning of each layer of RBM. After learning the RBM of the previous layer, it is used to learn the RBM of the next layer, and so on, finally forming a complete DBN network (Figure 3.2).

RBM is an energy-based model that combines the visible layer variable u and the hidden layer variable l in RBM. Its energy expression is shown in Figure 3.1:

$$Q(u, l|\beta) = \frac{1}{2}(u^T el + \varepsilon^T u + \sigma^T l) \tag{3.3}$$

$\beta = (e, \varepsilon, \sigma)$ is a parameter combination. After the values of each parameter are given, the standardized coefficient of RBM is $C(\beta) = \sum_{u,l} e^{-Q(u,l|\beta)}$. According to this energy equation, the joint probability distribution of (u, l) can be obtained as follows:

$$p(u, l|\beta) = \frac{e^{-Q(u,l|\beta)}}{C(\beta)} \tag{3.4}$$

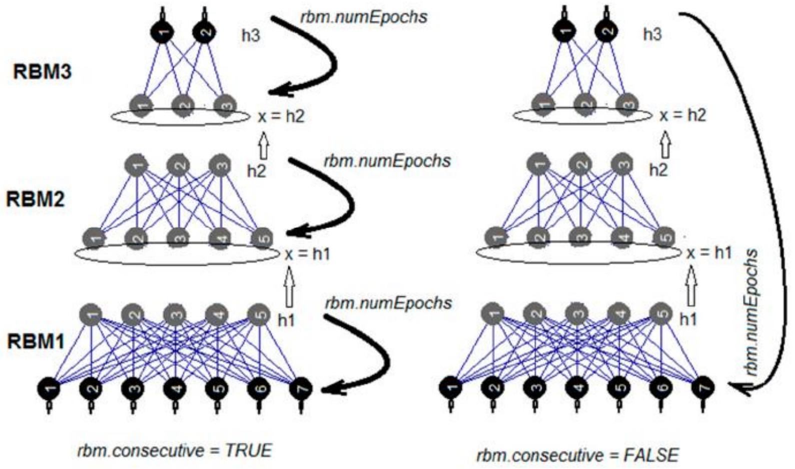

Fig. 3.1: Neural structure of RBM.

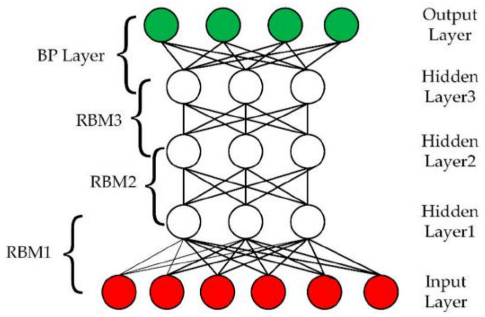

Fig. 3.2: DBN network structure.

There are several possibilities for the nodes of the hidden layer:

$$p(l_j = 1|u) = \varphi(\sigma_j + \sum{}_i u_i e_{ij}) \tag{3.5}$$

There are the following possibilities for visual layer nodes:

$$p(u_i = 1|l) = \varphi(\varepsilon_i + \sum{}_j e_{ij} l_j) \tag{3.6}$$

The learning essence of the RBM method is to find a probability distribution that can generate training samples to the greatest extent [13]. In other words, you need a distribution that produces the most significant number of possibilities. Because weight A is the key to influencing the probability distribution, we will learn weights based on probability graphs to learn the underlying model. This paper presents a fast algorithm called "contrast branching." This method only repeats the cycle in B cycles and gets a model estimate of 1. CD method first uses training samples to initialize the visual layer and then uses the conditional probability method to find the hidden layer [14]. The visual layer is obtained from the hidden layer by the conditional distribution. The result is a reconstruction of the input. The visual layer generates a vector C and transmits the value to the hidden layer through this vector. Inputs at the corresponding visual level are randomly selected to recover the original input. Finally, the visualized neuron reconstructs the neuronal activity unit l in the network through forward conduction [15]. The adjustment of weights is determined according to the degree of correlation between the hidden layer's active cells and the visible layer's input end. The model is solved according to the CD algorithm

$$\Delta e_{ij} = \zeta(\langle u_i l_j \rangle_{data} - \langle u_i l_j \rangle_{recon}) \tag{3.7}$$

ζ is how much students learn. $\langle u_i l_j \rangle_{data}$ represents the expected value of the sampled data. $\langle u_i l_j \rangle_{recon}$ represents the expected value of the reconstructed visualized data. The pre-training procedure for DBN follows the following steps:

1) The greedy algorithm is used to learn the first RBM.

2) Determine the weight and bias of the first RBM and use the calculated results as input to the upper RBM;

3) Repeat the above process several times until the reconstruction errors are minimized. The hidden layer can then become input to the visual layer.

E, a, b The specific steps of the DBN pre-training algorithm are as follows:

The input training: sample x_0, number of visible layer and hidden layer units n, m, learning rate ζ, and maximum training cycle R.

The output training: weight matrix e, visible layer bias a and hidden layer bias b.

Step 1. Initialize the initial state$u_1 = x_0$ of the visibility unit. E, a, b is a small arbitrary number.

Step 2. The iterative training period is t.

Step 3. The hidden layer l_1 is calculated from the visible layer u_1. The value of $P(l_{1j} = 1|u_1)$ is periodic, and probability is used as the probability of hiding the j cell of the layer.

Step 4. The visible layer u_2 is calculated from the hidden layer l_1. The value of $P(u_{2i} = 1|l_1)$ is computed cyclically, and this possibility is given as the possibility that the i unit of the visible layer is set to 1.

Step 5. The visible layer l_2 is calculated from the hidden layer u_2. The value of $P(l_{2j} = 1|u_2)$ is computed cyclically, and this possibility is given as the possibility that the i unit of the visible layer is set to 1.

Step 6. Update parameters

$$E \leftarrow E + \zeta(P(l_1 = 1|u_1)u_1^T - P(l_2 = 1|u_2)u_2^T) \tag{3.8}$$

$$a \leftarrow a + \zeta(u_1 + u_2) \tag{3.9}$$

$$b \leftarrow b + \zeta(P(l_1 = 1|u_1) - P(l_2 = 1|u_2)) \tag{3.10}$$

Step 7. Confirm that the number of iterations has reached the 8th step, not the 2nd step. Step 8: Output parameter

$$e, a, b$$

End.

3.3. Network Tuning. BP neural network is used to achieve one-step training based on each given weight. This process is called optimizing the deep trust net (Figure 3.3).

Set up the BP network at the last layer of the DBN. The feature vector is used as its input for guided learning. Each level of the BP neural network only ensures that the weight of this layer corresponds to the characteristic vector of this layer [16]. BP algorithm adopts a backward neural network to transmit the error message to each RBM layer from top to bottom to adjust the whole DBN. This improves the classification effect of the neural network.

4. Chinese news text classification experiment.

4.1. Introduction to Data Sets. The author obtains information from the Internet and corporate information by sifting financial information on a website. Divide the database into 1000 categories, each representing a business. The performance of the model will be tested with some mainstream classification algorithms.

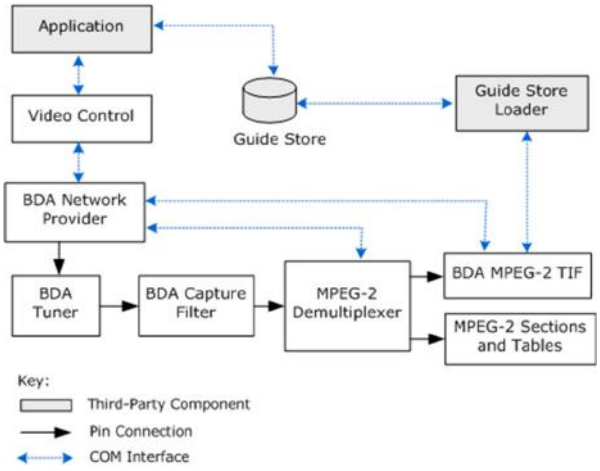

Fig. 3.3: Network tuning.

Table 4.1: Experimental results I.

Classification algorithm	Accuracy rate (%)	Recall rate (%)
HDBN	92.19	92.19
BP	88.54	87.50
SVM	90.63	91.67
ELM	90.10	89.58

Table 4.2: Experimental results II.

Classification algorithm	Accuracy rate (%)	Recall rate (%)
HDBN	90.10	90.10
BP	86.88	84.58
SVM	91.56	92.19
ELM	90.31	90.10

4.2. Test Results. First, the test sample's recognition accuracy should be evaluated. If the results significantly differ from the expected results, returning to the feature screening process and re-screening until the recognition value is in the appropriate range is necessary. The accuracy rate reflects the accuracy of text classification [17]. Only a high accuracy and a low recall rate mean that the label categories that should be predicted are not predicted. In particular, unbalanced samples tend to turn smaller categories into larger ones. Some other multilabel classification methods have problems, such as over-matching between samples. All these problems are worthy of attention. This paper uses HDBN, BP neural network algorithm, support vector machine, ELM and other algorithms to test it. The experimental classification accuracy and recall rate were evaluated (Table 4.1).

THUC News verifies the algorithm. The results are shown in Table 4.2.

5. Conclusion. This paper uses TF-IDF to weigh the text features and obtain the original text feature matrix. The classifier is built and optimized by using the DBN network. Finally, the accurate and fast classification of the text is achieved. Experiments show that the accuracy of using deep neural networks for text classification is significantly higher than BP, SVM, ELM and other classification methods.

REFERENCES

[1] Wu, H., Qin, S., Nie, R., Cao, J., & Gorbachev, S. (2021). Effective collaborative representation learning for multilabel text categorization. IEEE Transactions on Neural Networks and Learning Systems, 33(10), 5200-5214.

[2] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning–based text classification: a comprehensive review. ACM computing surveys (CSUR), 54(3), 1-40.

[3] Edara, D. C., Vanukuri, L. P., Sistla, V., & Kolli, V. K. K. (2023). Sentiment analysis and text categorization of cancer medical records with LSTM. Journal of Ambient Intelligence and Humanized Computing, 14(5), 5309-5325.

[4] Srilakshmi, V., Anuradha, K., & Shoba Bindu, C. (2021). Incremental text categorization based on hybrid optimization-based deep belief neural network. Journal of High Speed Networks, 27(2), 183-202.

[5] El Rifai, H., Al Qadi, L., & Elnagar, A. (2022). Arabic text classification: the need for multi-labeling systems. Neural Computing and Applications, 34(2), 1135-1159.

[6] Kumar, Y., Koul, A., & Mahajan, S. (2022). A deep learning approaches and fastai text classification to predict 25 medical diseases from medical speech utterances, transcription and intent. Soft computing, 26(17), 8253-8272.

[7] Luo, X. (2021). Efficient English text classification using selected machine learning techniques. Alexandria Engineering Journal, 60(3), 3401-3409.

[8] Moon, S., Kim, M. Y., & Iacobucci, D. (2021). Content analysis of fake consumer reviews by survey-based text categorization. International Journal of Research in Marketing, 38(2), 343-364.

[9] El-Alami, F. Z., El Alaoui, S. O., & Nahnahi, N. E. (2022). Contextual semantic embeddings based on fine-tuned AraBERT model for Arabic text multi-class categorization. Journal of King Saud University-Computer and Information Sciences, 34(10), 8422-8428.

[10] Ibrahim, M. F., Alhakeem, M. A., & Fadhil, N. A. (2021). Evaluation of Naïve Bayes classification in Arabic short text classification. Al-Mustansiriyah J. Sci, 32(4), 42-50.

[11] Wang, Z., Wang, L., Huang, C., Sun, S., & Luo, X. (2023). BERT-based Chinese text classification for emergency management with a novel loss function. Applied Intelligence, 53(9), 10417-10428.

[12] Gurcan, F., & Cagiltay, N. E. (2023). Research trends on distance learning: A text mining-based literature review from 2008 to 2018. Interactive Learning Environments, 31(2), 1007-1028.

[13] Kalra, V., Kashyap, I., & Kaur, H. (2022). Improving document classification using domain-specific vocabulary: hybridization of deep learning approach with TFIDF. International Journal of Information Technology, 14(5), 2451-2457.

[14] Lagrari, F. E., & Elkettani, Y. (2021). Traditional and deep learning approaches for sentiment analysis: A survey. Advances in Science, Technology and Engineering Systems Journal, 6(4), 1-7.

[15] Pintas, J. T., Fernandes, L. A., & Garcia, A. C. B. (2021). Feature selection methods for text classification: a systematic literature review. Artificial Intelligence Review, 54(8), 6149-6200.

[16] Sharma, S., Princy, K. B., & Sharma, R. (2023). A Study on Image Categorization Techniques. International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET), 6(5), 1147-1152.

[17] El-Alami, F. Z., El Alaoui, S. O., & Nahnahi, N. E. (2022). A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model. Journal of King Saud University-Computer and Information Sciences, 34(8), 6048-6056.

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Oct 28, 2023
Accepted: Nov 15, 2023