Navy text data set equalization method based on deep learning

Qi Yudong¹, Ding Haiqiang¹*, Zhao Jinchao²
¹Naval aeronautical university, Yantai, 264001, China
²92665 Army, Zhangjiajie, Hunan, 427000, China
*Corresponding author’s e-mail: 1076331817@qq.com

Abstract. Traditional unbalanced data sets, such as artificial set processing method, exists in the complicated and poor universality. It is difficult to apply naval uneven text data set processing. Aiming at this problem, this paper proposes a Navy uneven text data processing model based on biRNN. It learns text sequence features through biRNN model. Then it generates text similar to the original to balance text data. Text classification experiments are carried out on original data set and balanced data set respectively. The experimental results show that balancing the original data set based on biRNN method can effectively improve the performance of text classification.

1. Introduction
With the rapid development of navy informatization construction, the informatization degree of navy equipment management is getting higher and higher. And the business data presents explosive growth. The production of massive data provides data foundation for the research of big data in the navy field. Among them, unstructured data, represented by text-type data, contains large amount of data and rich information. Therefore, the research on mining accurate and valuable information from text-type data has become an important content in the navy's big data research. According to the equipment life cycle phase, the navy text data set is divided into application, allocation, allocate, receiving, storage, use, daily check and maintenance, maintenance, prolong life and scrap consumption. However, each unit informatization construction, construction of varying degrees in the morning and evening, cause early text data loss serious, acquisition work is difficult. The number of ten kinds of text data is uneven, the navy text data set appears imbalance state. Using unbalanced data sets for text data mining will lead to the result tending to the majority class, while the processing error of the minority class text is large, which seriously affects the overall effect of text data mining.

To solve the problem of unbalanced data sets, the current solutions mainly include moving threshold method, weight method and sampling method. Among them, the moving threshold method and the weight method are mainly adjusted at the algorithm level, and the unbalanced data sets are processed by adjusting the classification boundary between different samples and setting different penalty factors for different sample sets. Zhu Xiaogang et al. [1], aiming at the problem of effective identification and analysis of features under the classification of unbalanced data, screened the features of unbalanced features according to the features of power spectral density function of data, and effectively completed the effective identification of data features. Sampling method belongs to the solution of data layer, which mainly achieves the balance of data set by means of down-sampling for most classes or up-sampling for a few classes. According to the probability distribution of samples, Li Guohe et al. [2] adopted rotation method, clustering center selection method and other methods to conduct down-sampling processing for most classes, which showed good performance in the equalization method of seismic data sets. In the aspect of unbalanced text data set processing, the existing processing technology mainly USES machine
learning method to generate a few kinds of text. Cuckoo et al. [3] used k-means algorithm to cluster a few texts, and used genetic crossover and mutation operations to obtain new samples within each cluster, effectively improving the classification effect of minority classes on the text classifier. In this paper, kuai et al. [4] suggested the method of pso-smote to generate difference samples, and iterated to generate minority texts in order to achieve the goal of equalizing the text dataset. In this way, the classification of SVM was optimized.

All the above methods of dealing with unbalanced data sets derive new data sets based on the original data sets by means of artificial construction of features. Although the effect of data set equalization is achieved, feature setting is cumbersome, and one method is only applicable to one or even one data set, which lacks universality. Compared with common text data sets, naval text data sets have the following characteristics: (1) more data types. Naval text data sets contain the data of the whole life cycle of nearly one hundred kinds of equipment, among which each type of equipment contains ten types of text data. (2) there are large differences among various text data, with different life cycle stages, different text source departments of the navy, and large differences in text writing purpose, text format and text style. (3) the imbalance of data sets is more serious. Due to the different degrees of informatization construction of various units and the confidentiality requirements of some units, the number of texts in each stage is significantly different in the actual data collection process. Based on the above reasons, the traditional unbalanced data set processing method is not suitable for the balanced processing of naval text data set. In recent years, deep learning methods have been widely used in text data mining. Deep neural network models, such as convolutional neural network and circular neural network, can reduce the tedious feature engineering and automatically learn the dependence between text sequences. Applying deep learning method to the text data set processing of naval ordnance can effectively solve the problem of unbalanced text data set.

Comprehensive the above research, this paper presents a bidirectional circular Neural Network (biRNN) navy uneven text data processing method, using biRNN model active text extraction sequence characteristics, reduce the dependence on the characteristics of engineering. Through the existing minority class text data, biRNN model generates text. In this paper, a few samples are generated by text extension technique to achieve data set equalization. The original data set, the balanced data set and the text classification experiment were respectively carried out. The experimental results show that the balanced data set and the expanded data set can effectively improve the data set quality and improve the text classification performance.

2. Text data extension method
Recurrent Neural Network (RNN) is a kind of deep Neural Network which carries out modeling of sequence data and can learn the sequential relations among input sequences[5]. Suppose the input text sequence is \((w_1, w_2, ..., w_t)\), and the output at time \(t\) is \(y_t\). In the circular neural network, the text sequence at \(t\) time \(w_t\) and the output sequence at \(t-1\) time \(y_{t-1}\) are combined as the input at \(t\) time, and the formula is expressed as

\[ y_t = \tanh(W[w_t; y_{t-1}]) \]

Where, \(W\) represents weight matrix and \([w_t; y_{t-1}]\) represents \(w_t, y_{t-1}\) splicing operation.

The circulatory neural network iterates the above operation process and finally outputs a sequence \(y_{t+1} = (y_1, y_2, ..., y_t)\), where \(y_t\) represents the output at \(t\) time.

In order to solve this problem, a bidirectional Recurrent Neural Network (biRNN) is introduced to explore the hidden rules of the positive and negative directions of text sequences. The operation of the forward Recurrent Neural Network is denoted by

\[ \bar{y}_t = \tanh(W^f[w_t; y_{t-1}]) \]

Where, \(\bar{y}_t\) is the output at the \(t\) moment of forward circulatory neural network, and \(W^f\) is the weight matrix of forward circulatory neural network.

Operation of reverse circulation neural network is denoted as
\( y_t = \tanh(W^b[w_t; y_{t-1}]) \)

Where, \( y_t \) is the output at the \( t \) moment of the reverse circulatory neural network, and \( W^b \) is the weight matrix of the forward circulatory neural network.

In fact, text dataset extension can turn into a problem of generating similar text sequences from known text sequences[6-9]. In this paper, biRNN model is used to train neural network model with known text sequence of naval ordnance, and automatically learn sequence features of text sequence. By inputting core feature words, biRNN model can automatically predict surrounding text and output complete text sequence, so as to realize the purpose of text data set expansion. In addition, compared with traditional RNN model, biRNN model can extend text sequence from both positive and negative directions simultaneously, which improves the efficiency of text extension.

3. Text data set extension model design with birnn

The text data extension model based on biRNN is divided into four parts, text preprocessing layer, embedding layer, biRNN layer and output layer.

3.1. Text preprocessing layer

The original naval text data structure is not uniform, which requires text preprocessing to be converted into a form that is easy to be processed by computer, including text segmentation, word vector dictionary generation and data set division.

In the aspect of Chinese word segmentation, Chinese usually adopts the form of concatenation of characters and words, which requires the reasonable segmentation of the whole sentence through the technology of text word segmentation. In this paper, the stuttering word segmentation tool is used for word segmentation. In order to retain grammatical and semantic information as much as possible, Numbers, punctuation marks, English equipment models, etc are classified as separate words. In terms of word vector dictionary generation, word2vec, an open source word vector tool released by Google in 2013, is used for word vector training. The corpus after word segmentation is trained by word vector alone, and the word vector dimension is set as 150 dimensions to generate word vector dictionary. In terms of data set division, the whole naval ordnance text data set is divided into training set, cross validation set and test set in proportion to 6:2:2.

3.2. Embedded layer

The input text sequence is transformed into a numerical vector by embedding layer, which is the text vectorization process. For input text sequences \( X = (x_1, x_2, ..., x_i, ..., x_n) \), where \( x_i \) represents the \( i \) first participle. Using the word vector dictionary generated in the text preprocessing layer, each word segmentation corresponds to a word vector, and the whole input text sequence is converted into a word vector form, denoted as \( w = (w_1, w_2, ..., w_i, ..., w_n) \), where \( w_i \) represents the word vector representation corresponding to the \( i \) first word segmentation.

3.3. biRNN layer

For input text sequence \( y_{1:n} = \text{biRNN}(w_{1:n}) \), output \( y_{1:n} \) vector is generated by bidirectional circular neural network model, denoted as \( y_{1:n} = \text{biRNN}(w_{1:n}) \), where \( w_{1:n} = (w_1, w_2, ..., w_i, ..., w_n) \). Then each output vector is coded by MultilayerPerceptron (MLP), and softmax function is used to predict the output vector

\[ \bar{w}_{1:n} = \text{softmax}(\text{MLP}(\text{biRNN}(w_{1:n}))) \]

Where, \( \bar{w}_{1:n} \) represents the text vector of the predicted output.

In this paper, three word segmentation units are used as input units for surrounding word prediction. The length of predictive words on both sides is set as 30 and the input unit is assumed to be

\[ w_{j:j+2} = (w_j, w_{j+1}, w_{j+2}) \]
forward biRNN operation is denoted as biRNN\(^f\)\( (w_{ln}) \), backward biRNN operation is denoted as biRNN\(^b\)\( (w_{ln}) \). Then the text sequence generated by the text data extension method based on biRNN can be expressed as

\[
\overline{w}_{j-30:j-1} = \text{softmax}(\text{MLP}(\text{biRNN}^b(w_{j+j+2})))
\]

\[
\overline{w}_{j+3:j+32} = \text{softmax}(\text{MLP}(\text{biRNN}^f(w_{j+j+2})))
\]

\[
\overline{w} = [\overline{w}_{j-30:j-1}; \overline{w}_{j+j+2}; \overline{w}_{j+3:j+32}]
\]

Where, \(\overline{w}_{j-30:j-1}\) represents the text vector predicted by the forward circular neural network, \(\overline{w}_{j+3:j+32}\) represents the text vector predicted by the backward circular neural network, and \([x; y; z]\) represents the splicing processing of vectors.

3.4. output layer
After text preprocessing layer, text embedding layer and biRNN layer, the extended text vector sequence \(1:n = w_{ln}\) is obtained. According to word vector dictionary, the extended text vector sequence \(\overline{x} = \overline{x}_{ln}\) is inversely quest-generated to output the extended text sequence. This process can be expressed as \(\overline{x}_{ln} = W^{-1}(1:n)\).

3.5. model parameter
BiRNN model adopts GRU unit. The hidden layer is set as 3. The first layer is 256 neural network units, and the second and third layers are 512 neural network units. AdamOptimizer was used for model training. Softmax cross entropy loss function was used for training loss function. Learning rate was set to 0.001. Dropout was set to 0.5. Word vector input was set to 150 dimensions, and iteration round was set to 100.

4. Experiment and result analysis
In order to verify the effect of the biRNN model-based method for the expansion of unbalanced text datasets of the navy, the original naval ordnance text datasets and the balanced naval text datasets were used in the text classification experiments respectively. The text classification algorithm used the traditional convolutional neural network algorithm.

Data sets all contain ten types of text data including application, allocation, allocation, receiving, storage, using, daily inspection and maintenance, maintenance, life extension, consumption and scrap. Among them, the original data set has not done any data balancing and expansion processing, and the text data set presents an unbalanced state, denoted as data set 1;Balanced data set 1. Data set 2 is obtained by using the method of literature [3] and using genetic crossover and mutation operations to balance data within a few types of texts. The balanced data set 2 extends a few kinds of text by biRNN text data extension method, making the number of ten kinds of text data roughly equal, which is denoted as data set 3.

In the experiment, F1 value, a comprehensive evaluation index, was used to evaluate the classification results, and the calculation formula was shown in (1).

\[
F = \frac{2PR}{P + R}
\]

Where, \(P\) represents accuracy, \(R\) represents recall rate.

According to the experimental results, the text classification F value of the three data sets is statistically analyzed according to the text category. And the results are shown in table 1.
Table 1. Experimental results.

| Text category                        | DS 1    | DS 2    | DS 3    |
|--------------------------------------|---------|---------|---------|
| apply for                            | 95.4%   | 96.9%   | 95.6%   |
| distribution                         | 76.2%   | 82.4%   | 80.2%   |
| transfers                            | 94.4%   | 94.0%   | 94.7%   |
| receive                              | 85.6%   | 91.5%   | 87.4%   |
| storage                              | 79.7%   | 89.6%   | 88.7%   |
| use of                               | 92.5%   | 92.9%   | 92.8%   |
| daily inspection and maintenance     | 87.5%   | 88.3%   | 87.0%   |
| maintenance                          | 72.7%   | 87.0%   | 86.3%   |
| prolong life                         | 71.5%   | 81.9%   | 78.5%   |
| consumption of scrapped              | 83.4%   | 90.4%   | 89.9%   |

In data set 1, namely, the original data set, there is a big gap in the classification effect of texts of different categories. Among them, the classification F value of application, allocation and utilization of 3 types of texts reached more than 90%. The classification F value of receiving, daily checking, maintaining and consuming scrapped 3 types of texts reaches more than 80%. The classification F value of distribution, storage, maintenance and life extension is less than 80%. Among the ten categories of texts, the maximum F value of the application category is 95.4%. The lowest F value of the classification of longevity text is 71.5%. The above results indicate that in the unbalance text data set of naval ordnance, the text classification model results tend to be the majority text, while the text classification of the minority text is poor.

Compared with data set 1, the classification results of 6 categories of texts under data set 2, data set 3, including distribution, reception, storage, maintenance, life extension and consumption and scrap are significantly improved. It is shown that after the equalization of naval text data set, the classification performance of a few types of text can be significantly improved, and even achieve the classification effect close to that of most types of text.

Compared with the data set 2 under the effect of text categorization, the distribution of the under the three data sets, receiving, storage, maintenance, prolong life, scrap consumption of 6 class text text classification F value is higher. It shows that compared with the traditional SMOTE sampling method, using biRNN naval ordnance text data set equilibrium model the final text categorization effect is better.

5. TAG
In this paper, a method of navy unbalanced data set processing based on biRNN model is proposed. It uses biRNN model to automatically learn the features of text sequence, and uses feature words as the center to carry out two-way text expansion, which avoids the defects of traditional unbalanced data set expansion method. The experimental results show that using the method of naval unbalanced data set expansion based on biRNN model to equalize and expand the original data set can solve the negative influence of a few kinds of texts on the overall classification. It improves the quality of text data set as a whole. And it improves the accuracy of text classification model.

References
[1] Zhu xiaogang, ZHANG jidong. Effective identification and analysis of characteristics under unbalanced data classification [J]. Computer simulation, 2008,35(04):145-148.
[2] Li guohe, Zhang teng, Wu weijiang, Hong yunfeng, et al. Training data set equalization method for machine learning [J]. Computer engineering and design, 2019, 40(03): 812-818.

[3] Du Juan, Jiang lili, Chen Hongli. Research on sample generation methods for a few classes in text classification of unbalanced data sets [J]. Computer application research, 2009, 26(10): 3731-3734.

[4] Xu hanlin, Xu hao-lin, et al. Support vector machine-based method for unbalanced text classification [J]. Modern electronic technology, 2008, 41(15): 183-186.

[5] Yoav Goldberg. Deep learning-based natural language processing [M]. Che wanxiang, Guo jiang, Zhang weinan, Liu Ming, trans. China machinery industry press, 2015: 153-168.

[6] Yang shiwen, Yang mingjing, Liu weijing. Research on similarity action recognition method based on unbalanced data [J]. Information communication, 2018(05): 11-14.

[7] Pan zo-qiang, Zhang Lin, zhang lei, et al. Research on disease classification method of clinical imbalance data of traditional Chinese medicine [J]. Journal of intelligent systems, 2017, 12(06): 848-856.

[8] NiuBin, Wu peng, Mali, et al. Based on the generated against the behavior of the network data set extension methods [J]. Computer technology and development, 2019-03-21. http://kns, CNK,. net/KCMS/detail / 61.1450 TP. 20190321.0909.030. HTML

[9] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, et al. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, 2013.