Comparison of algorithms that use deep learning to classify spam

Congying Dai

Computer science and technology, East China University of Science and Technology, Shanghai, 200333, China

20001998@mail.ecust.edu.cn

Abstract. Although the application of network security protocols and cryptography provides a certain security guarantee for Internet surfing, it is difficult to cure the persistent security problems. Driven by the promotion of e-mail technology and benefits, bad businesses will also issue promotional emails indiscriminately to a large number of mailboxes, and even drive the underground industry of private mailbox information trading. The existing spam filters use black and whitelists, sensitive word matching and other technologies, but they can not effectively filter all forms of spam, and non-spam is often filtered, which brings more trouble to users. With the rise of artificial intelligence, machine learning algorithms have been applied to spam recognition, such as decision tree algorithm, Boosting algorithm, K nearest neighbour algorithm, SVM support vector machine algorithm, Bayesian principle related algorithms, etc. These methods based on traditional statistics can intelligently classify data sets with large differences and are often used together with expert systems with certain rules to classify spam. However, with the diversification of spam types, the old classification rules are relatively rigid, and new types of mail will be misjudged. In addition, statistics based natural language processing method is based on pre trained fixed dictionaries. For new words and polysemy words, it is impossible to give word vectors with accurate semantics, which brings difficulties to classification. This paper mainly studies the application of five machine learning algorithms in spam detection: improved naive Bayes algorithm, A Lite Bidirectional Encoder Representations from Transformer (ALBERT) dynamic word vector algorithm, Bidirectional Gating Recurrent Unit (BiGRU) algorithm, the Inverted Multi-Index with Weighted Naive Bayes (IMI-WNB) algorithm and clustering analysis algorithm.

Keywords: Spam detection, improved naive Bayes algorithm, ALBERT dynamic word vector algorithm, BiGRU algorithm, IMI-WNB algorithm, clustering analysis algorithm

1. Introduction

This review studied five kinds of algorithms that were applied to spam detection: Bidirectional Gating Recurrent Unit Attention Convolutional Neural Network (BiGRU-Attention-CNN), A Lite Bidirectional Encoder Representations from Transformer (ALBERT) dynamic word vector algorithm, the Inverted Multi-Index with Weighted Naive Bayes (IMI-WNB) algorithm, Naive Bayes and clustering algorithm. The ALBERT dynamic word vector is different from the traditional dictionary mapping mode, and the network structure needs to be re planned when it is used. BiGRU is more suitable for text modeling and obtaining global structure information of text. IMI-WNB algorithm is based on the improvement of IMI mutual information algorithm by introducing word frequency factor and inter class difference factor respectively.

The spam recognition method based on clustering analysis algorithm mainly uses the key words in the mail text as the cluster center in the clustering analysis algorithm, and then recognizes the spam from the huge mail collection. The improved Naive Bayes algorithm is based on the Naive Bayes algorithm, which uses a tree structure to maintain the number of occurrences of the feature words, and at the same time, it extracts the conditional probability of the feature words [1]. This review aimed at studying how different spam classification algorithms work. It also compared the performance of different algorithms that use deep learning to classify spam. This review tried to demonstrate that deep learning is well suited to be applied to spam classification as well. Spam is a message that the recipient refuses to receive or does not agree to receive but still receives. It mainly includes commercial, advertising, training, promotion, quotation and other kinds of emails. The key characteristics of these messages are mandatory and bulk sending. Spam can lead to the disclosure of critical information. It is extremely harmful to enterprises and users. Therefore, it is necessary to study how to detect spam.

Traditional spam detection methods include keywords, blacklist and whitelists, checksum code, etc. These traditional spam detection methods tend to be ineffective. At the same time, they are easily evaded by spam [2]. As a result, deep learning was chosen as the method of dealing with this kind of problem. Deep learning is an effective means of detecting spam. Different algorithms that use deep learning to classify spam have been discussed thoroughly. Meanwhile, application of these algorithms has been developed to a certain degree. However, optimization is still necessary. More algorithms suitable for spam detection need to be developed either. This review is written to investigate and compare the differences between different algorithms. This review was organized as follows: overview of the state of research and problems in the field of spam classification; analyze and introduce different deep learning algorithms (algorithms, advantages and disadvantages, current state of research); compare the differences between different algorithms; demonstrate that deep learning is well suited to be applied to spam classification [3].

2. Different Algorithms

2.1. Improved NB Algorithm

Since NB uses arrays in the email training part to maintain the number of occurrences of the characteristic words, the list generated in the process of traversing the email can be regarded as a matrix, and each element in each line of the matrix represents the number of occurrences of the characteristic words in each email. However, the boundary of the array must be read every time when the array is used to count the characteristic words, and the generated matrix needs to be column added to calculate the conditional probability, which leads to long training time and low system operation efficiency. At the same time, NB algorithm considers that the conditional probability of the appearance of the feature words in all emails is of the same importance for decision classification, and high-frequency words and low-frequency words have the same proportion, so it is unreasonable to simply calculate the conditional probability of the feature words [4].

Therefore, the algorithm introduces the idea of tree structure. Tree is a kind of nonlinear data structure, which can well describe data sets. In the training process, if a characteristic word only appears in the spam, the characteristic word will be stored in the tree structure of the normal mail synchronously and initialized to 1 to avoid the situation that the numerator or denominator is 0 when calculating the posterior probability [5]. The new algorithm is mainly divided into two stages: training stage and testing stage. In the training phase, the mail set is parsed, and then the feature words are extracted, and the tree structure is used for storage training. In the testing phase, the test set is preprocessed, and the tested mail category can be obtained by comparing the posterior probability obtained from the conditional probability processed by the classifier.

2.2. ALBERT Dynamic Word Vector

The ALBERT dynamic word vector is different from the traditional dictionary mapping mode, and the network structure needs to be re planned when it is used. In this spam classification model, the text is first encoded based on word level, and then CLS and SEP tags are added at the beginning and end respectively. After obtaining the formatted representation of the text, input the ALBERT model to learn the dynamic word vector.

The ALBERT-RNN network is divided into four parts. First, the ALBERT network is used to relearn the dictionary mapped fixed word vector, and then the RNN network is used to extract the sequence features in the sentence tags. Then, in the feature fusion part, the classification tags and sequence features are fused. Finally, the fused features are sent to the classification network. In the classification network, the algorithm uses Focal Loss to optimize the cross entropy loss function.In order to ensure that the optimizer can maintain considerable optimization vitality in the late training period and prevent over learning of easily fitting samples, Focal Loss is selected for the algorithm to modify the loss function [6].

2.3. BiGRU-Attention-CNN

RNN is a kind of neural network used to process sequence data. It can extract context relations. However, when training long sequences, RNN has the problems of gradient disappearance and gradient explosion. To solve these problems of RNN, Hochreiter et al. proposed a special RNN Long Short Term Memory (LSTM) network. Later, Cho et al. proposed Gated Recurrent Unit (GRU) network on the basis of LSTM. GRU has similar model effects compared with LSTM, but GRU training requires less resources. BiGRU model includes two GRU models: forward propagation and backward propagation, which have higher classification accuracy than GRU. To sum up, the algorithm adopts BiGRU model [7].

After the feature vector of the email passes through the BiGRU layer, the pre and post information of the email has been fully extracted, but the key information and keywords in the email have not been highlighted. The principle of the attention mechanism is very similar to the logic of people looking at pictures. Instead of seeing all the details of the picture, they focus their attention on the focus area of the picture. The algorithm introduces the attention mechanism to give more weight to the keywords in the email text to highlight the key information. The implementation process of attention mechanism is shown in Figure 1.

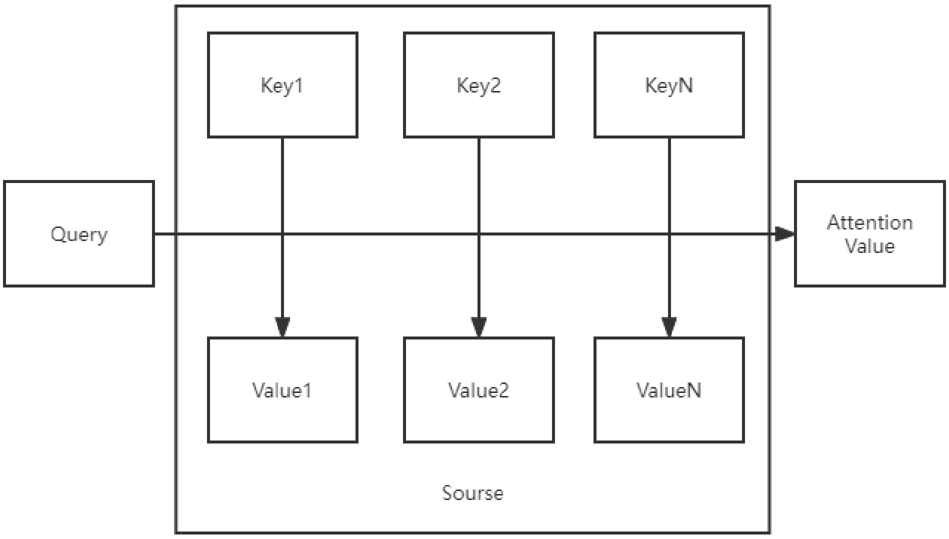

Figure 1. the implementation process of attention mechanism.
(Photo credit: Original)

The attention mechanism calculates the similarity between Query and Key to get the weight value, and then normalizes the weight value to get the weight. Finally, the weight and the weight value are weighted and summed to get the Attention Value.

The algorithm uses CNN model to further process the mail text feature vectors containing weight information, and obtains the final spam classification results.

2.4. *IMI-WNB Algorithm*

The traditional mutual information algorithm only deals with the frequency of words, but does not analyze the number of occurrences of words. In industrial networks, the filtering efficiency of junk text with unbalanced word frequency is very low. For example, the word frequency characteristics of the same language are very different. In traditional classification methods, words with larger frequency characteristics have higher correlation with classification. However, in this context, the traditional mutual information calculation method has a certain degree of correlation, which is obviously inconsistent with the reality. Based on this, this method proposes a method based on the lexical frequency factor and the difference between types, which improves the IMI mutual information. When the lexical frequency of a feature is higher than the word frequency, its corresponding lexical frequency coefficient is higher, and the proportion of this feature in spam screening is larger.

In addition, if the uneven distribution of characteristics of multiple classifications affects classification determination, it usually occurs in a specific classification, and rarely occurs in other classifications. In this case, it can be generally considered that this feature has a greater impact on the filtering of industrial Internet spam. In fact, such characteristics are called characteristics with large standard deviation in statistics

The feature can reflect the dispersion degree of the mail text, which is conducive to the filtering of spam [8].

2.5. *Clustering analysis algorithm*

Email is a characteristic and semi-structured text form. A standard email sample includes the header, body, and some attachments. The format of each email is not uniform. Therefore, the preprocessing of mail before mail recognition mainly involves word segmentation and removal of stop words. After the mail is preprocessed, some word segmentation messages are continuous messages [9].

Some of them cannot be used directly and need to be characterized. To this end, the feature attribute selected in the mail feature space contains the structural feature attribute of the mail. The segmentation is taken as the basic feature to describe the mail content, and all features contained in the mail text are obtained. The feature set of multiple mail paragraphs is merged to obtain a feature set containing multiple mail features. Each email is regarded as a feature vector and projected into the vector space represented by the feature set, where the dimensions of the vector space correspond to the features in the feature set one by one. If the mail contains a feature once, the weight value of the mail feature vector is assigned to the dimension corresponding to the feature, starting with 0 and then gradually increasing by 1. The value of the vector dimension corresponding to a participle in the final mail is approximately equal to the frequency of the word in the mail text. By feature extraction, the dimension of features is reduced. Email samples are quantized into feature vectors with more simplified features. The feature vectors are used as the clustering center in the clustering analysis algorithm, and the purpose of identifying spam is achieved through iterative calculation. Take the extracted mail features as the initial cluster center, calculate the distance between each mail sample and the initial cluster center, find the nearest cluster center, assign it to the nearest cluster, and recalculate the average value of each cluster. The Euclidean distance and assignment process are repeatedly calculated until the maximum number of iterations is reached or the difference between the newly calculated F value and the F value obtained in the previous iteration is less than a given threshold. After the calculation, the output result is the identified spam. So far, the design of spam identification method based on clustering analysis algorithm has been completed [10].

3. Comparison

Five different algorithms have their own advantages and disadvantages. The following figure is a comparison of them.

Table 1. the comparison of five algorithms.

Algorithm	Advantages	Disadvantages
ALBERT Aynamic Word Vector	1.High performance 2.The accuracy and recall of dynamic word vector model have been significantly improved.	The precision index is slightly low.
BiGRU-Attention-CNN	Excellent performance in training speed and accuracy	When selecting the mail data set for model training and detection, only the body part of the mail is selected, and other parts of the mail are not considered. However, these non body parts are also very helpful for spam recognition. Therefore, the algorithm needs to be improved.
IMI-WNB	1.Get more robust spam filtering results than traditional algorithms, 2.Effectively reduce the false detection rate and missed detection rate when filtering spam	It has a higher recall rate than traditional algorithms, so the leak rate of spam is significantly lower than that of traditional algorithms.
Improved NB	1.The training time of the improved NB algorithm is obviously less than that of the NB algorithm. 2.The increase of the improved algorithm is relatively gentle, and the effect is better	The recall rate has decreased.
Cluster Analysis	1.The text processing in the mail is more accurate and reliable 2.The identification of spam is more effective.	The balance between clustering accuracy and efficiency needs further study.

4. Conclusion

According to the word segmentation characteristics of spam, clustering analysis algorithm can identify other spam in the mail collection. in this process, it achieves the purpose of identifying spam by using the real-time discrimination ability of clustering analysis algorithm.

The improved Naive Bayes algorithm has improved the efficiency of email training and the performance of email classification to a certain extent, and has a good effect, which can be used for reference in spam filtering technology.

The accuracy of ALBERT-LSTM is high, and the accuracy and recall of the dynamic word vector model are significantly improved. After Focal Loss, the model is fitted to non spam, effectively reducing the problem of over fitting the original model to positive samples.

Compared with the traditional text classification model and other spam detection methods, the BiGRU Attention CNN model method has greatly improved, and the training speed and accuracy are also excellent.

The IMI-WNB algorithm uses the speech frequency coefficient and the type difference coefficient to improve the calculation of mutual information, and uses them as the attribute weight of Bayesian classification, thereby constructing a new spam filtering algorithm based on IMI-WNB. Through the test on open source data, it is found that this method can better improve the filtering effect of spam, and can effectively reduce false detection and missed inspection.

Spam is sent in a variety of ways and forms, and the detection technology is constantly updated. These algorithms are very effective for spam detection.

References

[1] Junjun F and Li L 2021 *Computer knowledge and technology* 154-155
[2] Junjun F and Li L 2021 *Computer knowledge and technology* 36-37
[3] Pan D 2019 *Yanshan University* 10
[4] Lu W, Weizhi L, Chengde Z and Yongjiu L 2020 *Sensors and Microsystems* 46-48
[5] Ge P 2020 *Computer knowledge and technology* 244-245
[6] Zhining Z, Bingjun W, YIming D and Xin T 2020 *Information Network Security* 107-111
[7] Yuxuan Z and Huaixiang H 2021 *Computer and Modernization* 122-126
[8] Xiaopeng J 2021 *Electronic Components and Information Technology* 165-167
[9] Jing W 2020 *Qufu Normal University* 106
[10] Xuan G 2020 *Computer and Modernization* 17-22