Journal of Computers Vol. 33 No. 2, April 2022, pp. 105-114
doi: 10.53106/199115992022043302009

An E-mail Classification Algorithm based on Stacking Integrated Learning

Li-Xia Wan*, Wei-Xing Huang, Qing-Hua Tang

College of artificial intelligence, Jiangxi University of Engineering, Xinyu, China
{1240147131, 1731513987, 1775851458}@qq.com

Received 20 June 2021; Revised 5 October 2021; Accepted 1 November 2021

Abstract. The text filtering of traditional anti spam system mainly focuses on keyword matching and text fingerprint analysis, which is difficult to accurately identify and classify spam. Therefore, an integrated learning algorithm based on stackin g is proposed in this paper. Firstly, the algorithm takes the manually marked text data of various categories as samples, uses TF-IDF algorithm to train the word vector space model, then selects linear SVC, xgboost and logistic regression algorithm to structure the base classifier, uses random forest algorithm to structure the meta classifier, and combines the stacking ensemble learning algorithm to structure the classification model. It achieves the function of dividing e-mail into five categories: illegal, advertisement, news, bill and recruitment. From the simulation results, the AUC values of the stacking integrated learning classification algorithm for each category are 0.92, 0.95, 1.00, 0.93 and 0.97 respectively, and the AP values are 0.86, 0.88, 1.00, 0.88 and 0.94 respectively, which realizes the high performance and high precision of text classification.

Keywords: anti spam system, integrated learning algorithm, TF-IDF algorithm, word vector space model, e-mail classification

1 Introduction

Spam affects the user experience because of its large quantity, large content and fast propagation speed [1]. Spam usually takes up system resources, consumes transmission time and affects the stability of the mail system. And because of the various forms of spam, it also brings a lot of security problems. Therefore, how to accurately identify and classify spam is the main goal of this paper.

There are three anti spam schemes for filtering the text of e-mail: keyword rule filtering, text fingerprint analysis and statistical content filtering [2]. Keyword rule filtering achieves the purpose of identifying spam by manually designing rule thesaurus and text matching. However, a single keyword is difficult to reflect semantic, context and emotional tendency, and is prone to false interception and missing interception. Text fingerprint analysis [3] uses specified technical means, such as text hash algorithm, to calculate the partial characteristic fingerprint of e-mail text and match it with the spam fingerprint database to complete the filtering process. However, the text fingerprint database relies on manual maintenance, which is inefficient and the matching process is strict, so it is difficult to achieve similarity matching. Statistical content filtering [4] generally uses word tags to associate with spam and non spam, and uses Bayesian inference to calculate the possibility that an email is spam. However, this scheme is difficult to realize multi classification of emails and can not differentially intercept spam.

In order to accurately identify and classify spam, an email text classification algorithm based on stacking integrated learning is put forward in this paper. The algorithm mainly includes two parts: constructing word vector space model and training classification model. The word vector space model takes the manually labeled text data as the input sample, segments the text part through Jieba word segmentation library, and uses TF-IDF algorithm to complete the construction of word vector space model. The word vector space model can convert the text into word vector (feature vector). The classification model takes the word vector as the input, uses linear SVC, xgboost and logistic regression algorithm as the base classifier, and takes the random forest algorithm as the meta classifier to complete the training of the whole integrated learning classification model. The algorithm can divide e-mail into five categories: illegal, advertisement, news, bill and recruitment. The simulation results show that text classification algorithm proposed in this paper has the characteristics of high accuracy and stable performance.

2 Related Work and Our Contributions

* Corresponding Author

The purpose of e-mail classification is to identify the semantic features contained in the text, and classify and process the e-mail according to the features. At present, there are mainly two schemes in text semantic classification: filtering dictionary matching and machine learning [5].

The filtered dictionary matching scheme usually needs to segment the text to be tested, then match the word segmentation words with the filtered dictionary, find the emotional words, degree words, negative words, etc., and calculate the category tendency score of the text to be tested, so as to realize text classification [6]. However, this scheme needs to manually construct a filter dictionary composed of filter words and design a series of correlation matching rules to complete text classification, which is not efficient. In addition, a single filtered phrase matching is difficult to reflect the implied semantics and context of the text, so this scheme is less applied [7].

The machine learning scheme is usually supervised learning, which mainly includes three steps: acquiring text data set, extracting text emotional features and training classification model [8]. The analysis process is shown in Fig. 1. At present, many researchers have done a lot of work in the field of text classification. Jianfeng Deng and others propose a new text classification model, called attention-based BiLSTM fused CNN with gating mechanism (ABLG-CNN), which can capture text context semantics and local phrase features, and perform experimental verification on two long text news datasets [9]. In addition, Xie Jinbao proposed a recursive convolution multitask learning (mtl-rc) model for text multi classification. The model combines the advantages of multi task learning, recurrent neural network (RNN) and convolutional neural network (CNN) to model multi task text. The model can obtain the correlation between multi domain texts and the long-term dependence of texts [10]. However, most of the above traditional machine learning models are weakly supervised models, which have a selective preference for the categories of text classification, and are difficult to adapt to the multi classification scenario of text [11].

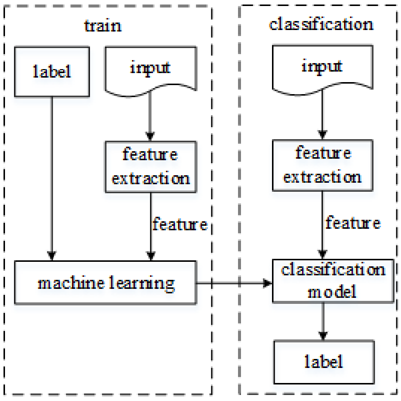

Fig. 1. Algorithm flow of machine learning text classification

The main contributions of this paper are as follows:

(1) The improved TF-IDF algorithm is adopted to construct the word vector space model, which is used to convert the text into word vector (feature vector).

(2) Three weak supervision models of Linear SVC, xgboost and logistics regression are used as the base classifier, random forest algorithm model is used as the meta classifier, and the stacking integrated learning algorithm is combined to train the strong supervision model, so as to complete the construction of the stacking integrated learning model.

3 The Algorithm Flow of Stacking Integrated Learning Classification Algorithm

In the supervised learning algorithm of machine learning, the goal is to train a stable model with good performance in all aspects, but the actual situation is often not so ideal. More often, we can only get multiple preferred models, that is, the weak supervised model with good performance in some aspects. Integrated learning is to combine multiple weak supervised models in order to get a better and more comprehensive strong supervised model. The potential idea is that even if one weak classifier gets the wrong prediction, other weak classifiers can correct the error. In the field of integrated learning, there are mainly three methods: bagging, boosting and stacking. Stacking method refers to training a model to combine other models. As long as the appropriate model combination strategy is adopted, bagging and boosting methods can be represented by stacking, which has better adaptability and stability. Therefore, Stacking integrated learning strategy is selected in this paper.

The principle of stacking integrated learning classification algorithm is shown in Fig. 2. The whole process includes three steps: pretreatment for training sample, training word vector space model and construct classification model. Firstly, processing sample data, which including filtering interference words and invalid characters, text segmentation and removing stop words, and get the labeled text segmentation data. Then, the TF-IDF algorithm is used to construct the word vector model, which is used to calculate the feature vector of each text segmentation. Finally, the feature vector of labeled text is used as the training set, and the stacking integrated learning algorithm is used to train the classification model by fusing linear SVC, xgboost and logistic regression algorithm.

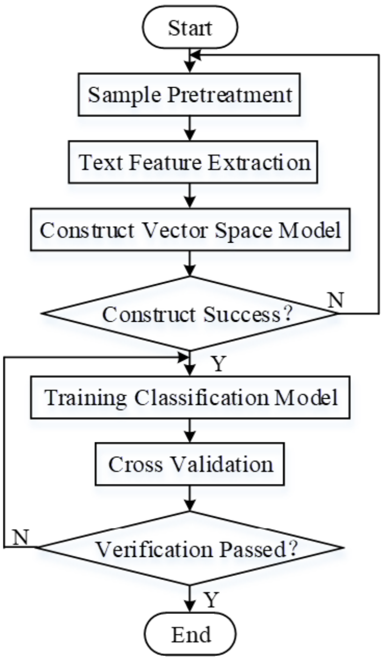

Fig. 2. Text classification algorithm flow

4 Constructing Word Vector Space Model

4.1 Pretreatment for Training Sample

The training samples are mainly from manually labeled text data. Before the feature extraction, the original text data needs to be preprocessed, which is important for feature extraction. A good preprocessing process will significantly improve the quality of feature extraction and the performance of classification algorithm. The sample pretreatment group in this paper should include the following steps.

(1) The words are segmented by the Jieba word segmentation library, then we build a dictionary of stop words, including adverbs, adjectives and conjunctions without actual semantics, and the samples after segmentation are processed by remove stop words.

(2) The word data is marked with part of speech tagging, which is to judge the verb, noun, adjective and other attributes of the segmentation.

4.2 Training Word Vector Space Model

The TF-IDF algorithm is used to calculate the word vector model of a document in this paper. TF-IDF is a statistical method, which is used to evaluate the importance of a word to a document in a corpus [12]. TF (term frequency) represents the word frequency, and the definition of probability representation is as follows.

$$\mathrm{TF}(t_i, d_i) = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

Where $n_{i,j}$ is the number of words t_i occurrences in the document d_i, and the denominator $\sum_k n_{k,j}$ is the sum of all words occurrences in the document. IDF (inverse document frequency) represents the frequency of reverse document, which is used to describe the importance of a word and defined as follows.

$$\mathrm{IDF}(t_i, D) = \log\left(\frac{|D|}{\left|\{d \in D : t_i \in d\}\right|}\right) \tag{2}$$

Where |D| is the total number of documents in the corpus, and $|\{d \in D;\ t_i \in d\}|$ is the number of documents containing the word t_i in the corpus, that is the number of documents of $n_{i,j} \neq 0$.

4.3 Constructing Space Vector Model

Most texts use natural language and contain unstructured information, which is difficult to be processed by computer. Therefore, how to accurately represent the text is the main factor affecting the performance of text classification. In recent years, many researchers have proposed many text representation models, such as Boolean model, vector space model, latent semantic model and probability model, to express the semantics of text with a specific structure. Considering the classification speed, the stacking integrated learning model in this paper uses the space vector model to represent the text. The vector space model is proposed by G Salton of Harvard University. The model transforms a given text into a vector with high dimension, and takes the feature term as the basic unit of text representation. Each dimension of the vector corresponds to a feature term in the text, and each dimension itself represents the weight of the corresponding feature term in the text. The weight represents the importance of the feature to the text, that is, how much the feature can reflect the category of the document.

Vector space model is an algebraic model that represents a document as a vector, and the similarity between documents is calculated by the angle between vectors [13]. Suppose that the number of all words in the corpus is T, the j document is d_j, and the document to be queried is q_i, the vectors of the two are as follows.

$$d_j = \left(w_{1,j}, w_{2,j}, \cdots, w_{T,j}\right) \in \mathrm{R}^T,\ q_i = \left(w_{1,i}, w_{2,i}, \cdots, w_{T,i}\right) \in \mathrm{R}^T \tag{3}$$

Where R^T represents the word vector space model, then the calculation formula of similarity between d_j and q_i is

$$sim(q_i, d_j) = \log\left(\frac{q_i \bullet d_j}{|q_i| \bullet |d_j|}\right) \tag{4}$$

5 Training Stacking Classification Model

Stacking is an integrated learning technology, which constructs a classification model by training n base classifier combinations. The algorithm flow is shown in Fig. 3. The base classifier uses the document vector calculated by the word vector space model as the input to train each classification model, and the meta classifier uses the prediction class labels of each base classifier as the feature data set to train the classification model, and carries out n fold cross validation on the training results to prevent the occurrence of over fitting phenomenon [14].

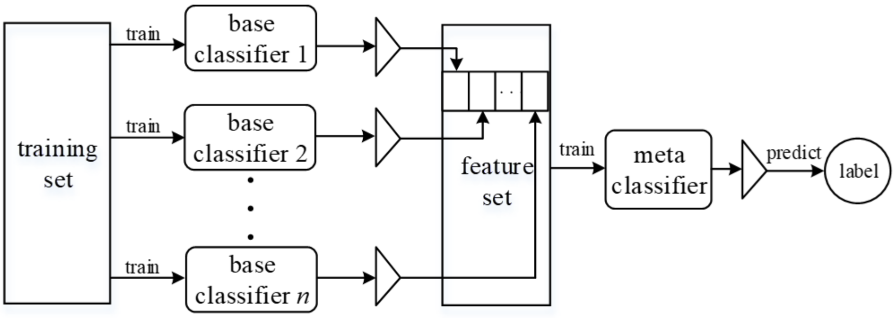

Fig. 3. The algorithm flow of stacking classification

5.1 Selection of Base Classifier

Stacking integrated learning method requires that the implementation principle of each base classifier is different, and is suitable for the same data set. In addition, the classification results is similar [15].

Linearsvc is a supervised machine learning algorithm using hinge loss function as kernel method to achieve linear classification. The trained classifier has the characteristics of sparsity and robustness [16], and the over fitting control mechanism is introduced to ensure the generalization performance of the classifier. It is widely used in pattern recognition fields such as portrait recognition and text classification. Xgboost is a machine learning algorithm based on gradient boosting classifier. It improves the existing gradient lifting algorithm by sparse feature processing and L1/L2 regularization, which has the advantages of reducing over fitting and improving computational efficiency [17]. Logistic regression is a generalized linear regression algorithm based on logistic function. It is often used in data mining, economic prediction, automatic disease diagnosis and other fields [18]. In view of the above reasons, this paper selects three machine learning algorithms, linear SVC, xgboost and logistic regression, as the base classifier.

5.2 Training of Meta Classifier

The meta classifier is a secondary classifier trained with the output combination of each base classifier as the feature set. The randomness of the training set sampling and the estimation variance of the model need to be considered in the selection of the algorithm. Therefore, this paper selects Random Forests algorithm as the meta classifier. Assuming that the number of samples is N and the number of attributes of each sample is M, the training steps of the classifier are as follows:

(1) N samples to be put back are randomly selected to train a decision tree as samples at the root node of the decision tree.

(2) m attributes (m << M) are randomly selected from M attributes, and a specified strategy, such as information gain, is used to select an attribute as the splitting attribute of the subordinate node to obtain the splitting node.

(3) Each child node in the decision tree is split according to step (2). If the attributes of the current node are the same as those used by the parent node, it means that the node has reached the leaf node, and then the splitting is stopped.

(4) Repeat steps (1) ~ (3) to establish a large number of decision trees to complete the training process of random forest.

Assuming that the n-th decision tree after training is f_n, then for the sample x of unknown category, the prediction output of the classifier can be obtained according to the prediction mean value of all decision trees, i.e

$$\hat{f} = \frac{1}{N}\sum_{n=1}^{N} f_n(x) \tag{5}$$

The correlation calculation formula between multiple decision trees is

$$\sigma = \sqrt{\frac{\sum_{n=1}^{N}[f_n(x)-\hat{f}]^2}{N-1}} \tag{6}$$

The calculation results shows that the multi decision tree model of random forest can reduce the correlation between the models and achieve the effect of reduce the noise of each base classifier.

5.3 Cross Validation

Since the stacking integrated learning method proposed in this paper adopts a scheme of two-level classifier, using the same sample to train base classifier and meta classifier may lead to over fitting. In order to train a reliable and stable classification model, the k-fold cross validation method is usually used to evaluate the accuracy of the model. This paper adopts a 5-fold cross validation scheme. Firstly, the sample set is randomly divided into five groups, four groups of samples are selected to train the base classifiers, and the remaining group of samples is input into the trained base classifiers to get the prediction label as the feature vector for training the meta classifier. Finally, five of full sample sets are used to train each base classifier. At this time, the newly trained base classifier and meta classifier constitute the final stacking integrated learning model. Algorithm 1 is the specific implementation process of cross validation.

Algorithm 1. Cross validation algorithm

```
Input:training data set $D = \{x_i, y_i\}_{i=1}^{m}$ ($x_i \in R^T, y_i \in Y$)
Output:stackingEnsemble learning classifier $H$
a) Adopt cross validation approach in preparing a training set
Randomly split $D$ into 5 equal-size subsets: $D = \{D_1, D_2, ..., D_5\}$
    for $k \leftarrow$ to 5 do
        // Train base classifiers
        for $t \leftarrow 1$ to $T$ do
            Train a classifier $h_{kt}$ from $D_k \in D$
        end for
        // Construct a training set for meta classifier
        for $x_i \in D_k$ do
            Get one record $\{x'_i, y_i\}$, where $x'_i = \{ h_{k1}(x_i), h_{k2}(x_i), \cdots, h_{k5}(x_i),\}$
        end for
    end for
b) Train a meta classifier
    Learn a new classifier $h'$ from the collection of $\{x'_i, y_i\}$
c) Re- train base classifiers
    for $t \leftarrow 1$ to $T$ do
        Train a classifier $h_t$ , based on $D$
    end for
return $H(x) = h'(h_1(x), h_1(x), \cdots, h_1(x))$
```

6 The Effect of Text Classification

Precision and recall are commonly used indicators to evaluate machine learning models [19]. Assuming that the number of positive samples predicted by the model to be positive is TP, the number of positive samples predicted by the model to be negative is FP, the number of positive samples predicted by the model to be negative is FN, and the number of negative samples predicted by the model to be negative is TN, then the calculation formulas of accuracy rate and recall rate are shown as follows.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}, \ \text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{7}$$

In addition, in order to evaluate the sensitivity and misjudgment of the machine learning model, the true positive rate TPR (true positive rate) and false positive rate FPR (false positive rate) indicators are usually introduced [20], which are defined as follows:

$$\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}, \ \text{FPR} = \frac{\text{FP}}{\text{FP+TN}} \tag{8}$$

In order to intuitively evaluate the machine learning model, ROC curve and PR curve are usually used to show the classification effect [21]. In this paper, 1000 labeled word segmentation samples are selected as the training data. The ROC curve obtained from the test simulation is shown in Fig. 4, in which FPR is the abscissa, TPR is the ordinate, and AUC (area under curve) is the area under the ROC curve, indicating the probability that the predicted positive example is ahead of the negative example. According to the simulation diagram, the AUC values corresponding to each category are more than 0.9, indicating that the model performance is good. The PR curve corresponding to the classification model is shown in Fig. 5. The curve takes recall as the abscissa and precision as the ordinate. The average accuracy value (AP) corresponding to each category is more than 0.85, indicating that the multi classification accuracy of the model is high.

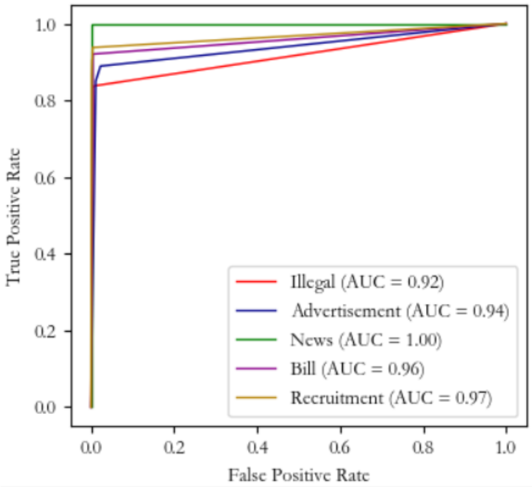

Fig. 4. ROC curve of stacking integrated learning model

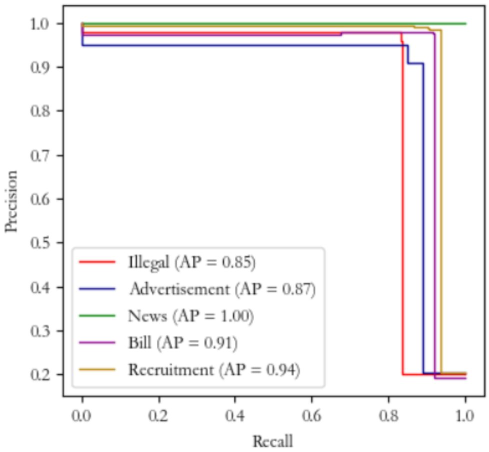

Fig. 5. PR curve of stacking integrated learning model

In order to compare the performance differences between single base model and ensemble learning model, 1000 labeled samples is selected as test data, and inputs them into linear SVC, xgboost, logistic regression and stacking ensemble learning algorithm models respectively. Then taking precision and recall as evaluation indexes, the detailed test data are shown in Table 1 and Table 2, and the comparison of each classification model are shown in Fig. 6 and Fig. 7.

Table 1. The precision of each classification model

Model	Precision				
	Illegal	Advertisement	News	Bill	Recruitment
LinearSVC	0.95	0.92	0.94	0.88	0.91
xgboost	0.94	0.89	0.96	0.90	0.90
logistic regression	0.94	0.90	0.92	0.89	0.94
stacking	0.98	0.96	0.99	0.95	0.98

Table 2. The recall of each classification model

Model	Recall				
	Illegal	Advertisement	News	Bill	Recruitment
LinearSVC	0.81	0.80	0.89	0.91	0.88
xgboost	0.79	0.82	0.87	0.88	0.87
logistic regression	0.84	0.79	0.87	0.92	0.91
stacking	0.89	0.87	0.94	0.95	0.93

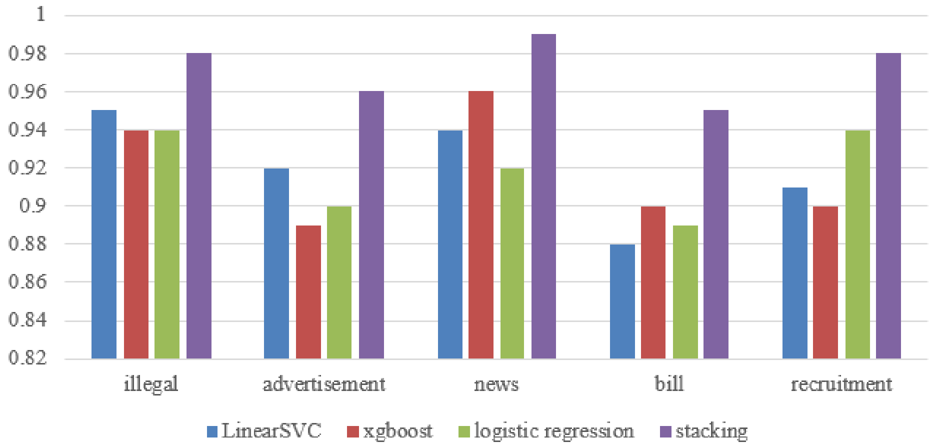

Fig. 6. Accuracy comparison of each classification model

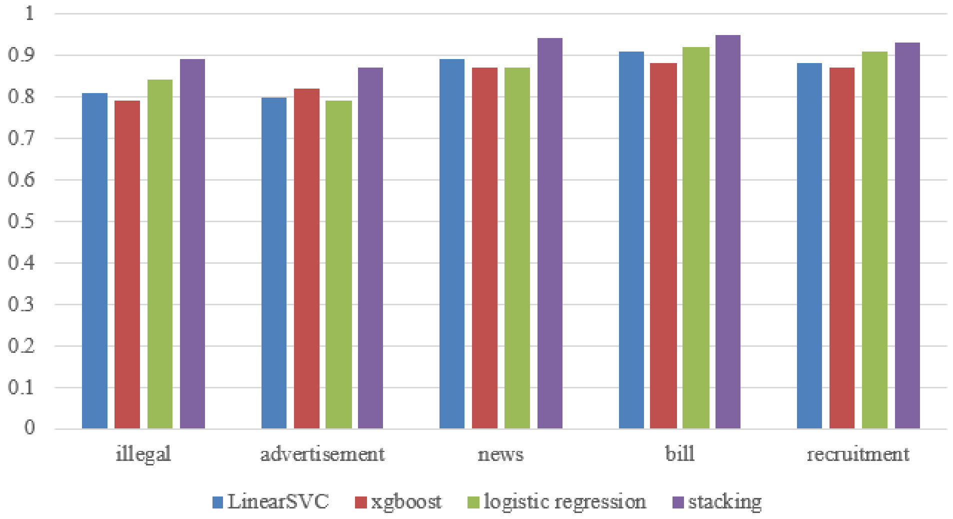

Fig. 7. Recall comparison of each classification model

According to the test results, compared with a single machine learning model, the accuracy and recall rate of stacking integrated learning model in each category of data set are higher than that of a single machine learning model. It shows that the algorithm has better multi classification performance than a single machine learning model.

7 Summary and Prospect

An e-mail classification scheme based on stacking integrated learning is designed in this paper, which can divide e-mail into five categories of illegal, advertisement, news, bill and recruitment according to the text part of e-mail. The scheme is based on the word vector space model constructed by TF-IDF algorithm, and a two-level classification architecture composed of base classifier and meta classifier is designed. In addition, a 5-fold cross validation scheme is introduced to eliminate the over fitting phenomenon between the sub models. From the simulation results, we can see that the AUC value of the classification algorithm for each category is above 0.9 and the average accuracy value is above 0.85, which realizes the high accuracy and high performance of text multi classification. At present, the e-mail classification algorithm based on stacking integrated learning has been successfully connected with the anti spam system of China Telecom 189 mailbox, and applied for relevant invention patents, which intercepting 1.48 million illegal and advertising e-mails per day and improving the experience of mailbox users. The proposed algorithm has a positive practical significance to purify the network environment and reduce the storage pressure of the e-mail system.

Due to the limitation of the number and types of samples, the algorithm proposed in this paper can only divide the mail into five categories, which can not meet the needs of intelligent tags and differentiated interception of the mailbox system. In the later stage, the training sample set will be further enriched, the training parameters will be continuously optimized, and the number of categories of stacking integrated learning model will be increased.

Acknowledgement

This work is supported in part by the Science and technology research project of Jiangxi Provincial Department of Education under Grant 181145.

References

[1] D. Gaurav, S.-M. Tiwari, A. Goyal, Machine intelligence-based algorithms for spam filtering on document labeling, Soft Computing 24(13)(2020) 9625-9638.

[2] C. Cui, D. Lü, S.-F. Jiang, Effect of Cost Parameters Adjustment on the Accuracy of Bayesian Anti-Spam Filtering System, Transaction of Beijing Institute of Technology 39(2)(2019) 142-146.

[3] S. Pang, J. Yao, T. Liu, A text similarity measurement based on semantic fingerprint of characteristic phrases, Chinese Journal of Electronics 29(2)(2020) 233-241.

[4] T. Yuan, J. Cheng, X. Zhang, Enriching one-class collaborative filtering with content information from social media, Multimedia Systems 22(1)(2016) 51-62.

[5] H. Amazal, M. Kissi, M. Kissi, A New Big Data Feature Selection Approach for Text Classification, Scientific Programming, 2021.

[6] A. Moreo, A. Esuli, F. Sebastiani, Word-class embeddings for multiclass text classification, Data Mining and Knowledge Discovery 35(3)(2021) 911-963.

[7] Z.-H. Chen, J.-T. Ren, Multi-label text classification with latent word-wise label information, Applied Intelligence 51(2)(2021) 966-979.

[8] X.-Y. Luo, Efficient English text classification using selected Machine Learning Techniques, Alexandria Engineering Journal 60(3)(2021) 3401-3409.

[9] J.-F. Deng, L.-L. Cheng, Z.-W. Wang, Attention-based BiLSTM fused CNN with gating mechanism model for Chinese long text classification, Computer Speech and Language 68(2021).

[10]J.-B. Xie, J.-H. Li, S.-Q. Kang, A Multi-domain Text Classification Method Based on Recurrent Convolution Multi-task Learning, Journal of Electronics and Information Technology 43(8)(2021) 2395-2403.

[11]W. Huang, M.-Y. Liu, W.-Q. Shang, LSTM with compensation method for text classification, International Journal of Wireless and Mobile Computing 20(2)(2021) 159-167.

[12]F. Wang, S. Yang, Z. Zhao, Automated signature generation algorithm for polymorphic worms based on improved TF-IDF, Journal of Huazhong University of Science and Technology (Natural Science Edition) 48(2)(2020) 79-84.

[13]Y. Li, Y. Zhao, A Virtual Space Vector Model Predictive Control for a Seven-Level Hybrid Multilevel Converter, IEEE Transactions on Power Electronics 36(3)(2021) 3396-3407.

[14]Z.-Y. Algamal, Shrinkage parameter selection via modified cross-validation approach for ridge regression model, Communications in Statistics: Simulation and Computation 49(2020) 1922-1930.

[15]H. Dai, W. Wu, J. Li, Incorporating feature selection in the improved stacking algorithm for online learning analysis and prediction, Engineering Letters 28(4)(2020) 1011-1022.

[16]Y. Zou, P. Dong, K. Liu, Design of nonlinear coordinate damping controller for HVDC and SVC based on synergetic control, IEEJ Transactions on Electrical and Electronic Engineering 15(1)(2020) 61-69.

[17]X. Gu, Y. Han, J. Yu, A novel lane-changing decision model for autonomous vehicles based on deep autoencoder network and XGBoost, IEEE Access 8(2020) 9846-9863.

[18]B. Su, Q. Yang, J. Yang, Encryption algorithm for network communication information based on binary logistic regression, Journal of Intelligent and Fuzzy Systems 39(2)(2020) 1627-1637.

[19]J.-D. Wang, G.-H. Shi, F.-Q Meng, Research on sentiment classification method base on multi-objective evaluation subject, Journal of Computers (Taiwan) 32(2)(2021) 137-148.

[20]S. Kong, W. Shen, Y. Zheng, False positive rate control for positive unlabeled learning, Neurocomputing 367(2019) 13-19.

[21]S. Song, Y. Zhou, Nonparametric estimation of the ROC curve for length-biased and right-censored data, Communications in Statistics - Theory and Methods 49(19)(2020) 4648-4668.