Naive Bayesian Algorithm for Spam Classification Based on Random Forest Method

Lu Wang¹, Guohui Zeng*¹ and Bo Huang¹

¹School of Electrical and Electronic Engineering, Shanghai University of Engineering Science, Shanghai, 201600, China
*Corresponding author’s e-mail: zenggh@sues.edu.cn

Abstract. Aiming at the problem of dimensionality disasters in the spam filtering system, a spam filtering model based on content feature selection random forest algorithm is proposed. The method combines the Naïve Bayesian algorithm and effectively reduces the influence of redundant information on classification performance through feature selection filtering of Gini impurity. The improved algorithm takes advantage of decision tree integration and naïve Bayesian algorithm. The experimental results show that compared with the naïve Bayesian algorithm, the improved algorithm has been improved in the classification performance indicators of spam filtering, which is more feasible and effective in practical applications.

1. Introduction

With the continuous development of network technology, the types and quantity of spam are also increasing, which brings inconvenience to people's daily life. Therefore, spam filtering technology [1-6] has also evolved. For example, based on blacklist filtering technology, rule-based filtering technology, naive Bayesian filtering technology, support vector machine algorithm, KNN algorithm, and the like. Naïve Bayes is popular for its excellent classification and high accuracy. However, Naïve Bayes is vulnerable to the classification of spam classification training when dealing with high feature value dimensions, reducing the classification performance of the classifier.

In the classification process, the naïve Bayesian algorithm assumes that all eigenvalues have the same importance and independence, but this assumption is not in line with the actual situation, and also affects the accuracy of the spam filtering algorithm. In order to solve the above problems, many Scholars have improved the Naïve Bayes algorithm. [7] proposed a weighted Bayesian mail filtering model that gives different weight coefficients to the mail header and mail body. [8] proposed the idea of introducing feature word weighting, and proposed a low-loss Bayesian spam filtering algorithm. Although these methods effectively improve the accuracy of the spam filtering system, they cannot avoid the learning training of high-dimensional eigenvalues [9-11], so the eigenvalue weighting method has become a research hotspot.

The random forest [12-14] is a tree classifier-based classification algorithm proposed by Breiman in 2001. This method has many advantages, such as no pre-processing, no over-fitting, and selection of characteristic genes. It first uses the Bagging method to create a differential training sample set, and uses the classification regression tree as a meta-classifier. When constructing a single tree, a strategy similar to random subspace partitioning is used to randomly select features to perform attribute splitting on internal nodes. This "double random" strategy creates a large difference between the sub-classifiers, making the random forest have superior classification performance and is one of the most successful integrated learning methods.
Considering that the random forest has a natural advantage in processing data classification, this method can select the data to construct the tree classifier, and can perform feature screening through Gini impurity, which can reduce the accuracy of data weight difference to a certain extent. Influences. In this paper, under the research of different scholars, the eigenvalue weights of random forests in classification and feature extraction of sample data are discussed. The naive Bayesian algorithm is used to train the eigenvalues after screening. The experimental results show that the improved algorithm has higher accuracy and better test time when dealing with small sample spam classification, which improves the performance of the classifier to some extent.

2. Methods

In the process of spam training, the eigenvalues assumed by the naive Bayesian algorithm are independent of each other. Although the naive Bayesian algorithm is simpler, it also causes dimensionality disasters and selects important features. Values, building models through these features in subsequent training learning can alleviate dimensional disasters, and removing some unrelated features can also reduce training difficulty. The feature selection methods mainly include principal component analysis, lasso and so on. However, among these methods, the idea of random forest method for feature importance assessment is relatively simple. Random forest (RF) [12] is a commonly used method for analyzing high-dimensional data. When discriminant analysis is performed, the importance weight of variables can be given at the same time. The measures of random forest contribution include: Gini index (gini), out-of-bag data (OOB) error rate, when the data category ratio is not balanced, and the training data is different in the two categories, especially the ratio difference. When you are big, using the Gini index to calculate the importance weight of variables will bring better results.

2.1 Weighted random forest characteristics importance assessment

The method of selecting feature filtering in this paper is the Gini index (also known as Gini impurity, Gini index value). That is, the children are randomly selected from a data set to measure the proportion of other data they are incorrectly assigned to other groups. The principle is as follows:

If there are n features \( x_1, x_2, \ldots, x_n \), then calculate the gini index of each feature \( x_j \), and the formula is as follows:

\[
I_G(f) = \sum_{k=1}^{m} \hat{f}_i(1 - \hat{f}_i) = \sum_{k=1}^{m} (\hat{f}_i - \hat{f}_i^2) = \sum_{k=1}^{m} \hat{f}_i - \sum_{k=1}^{m} \hat{f}_i^2 = 1 - \sum_{k=1}^{m} \hat{f}_i^2
\]

where \( k \) indicates that there are \( k \) categories, and \( \hat{f}_i \) indicates the proportion of nodes \( i \).

If the eigenvalues of the same category have different eigenvalues in the same category, and the kinetics of the eigenvalues are the same, the eigenvalues are considered to have the same influence on the category of the test judgment message, in order to reduce the dimension, improve the test time, and pass the random forest feature. The value selection filters out the feature values and stores the feature values that affect the classification results in a dictionary and trains the new training set using the naive Bayesian algorithm.

2.2 Naive Bayesian algorithm

The Bayesian classifier divides messages into normal mail and spam. Calculate the probability of spam and normal mail according to the naive Bayes formula, and then compare the magnitudes of these two probabilities. Calculated as follows.

For a sample \( D \) to be classified, its sample attribute and categorical variable are, as can be known from Bayes' theorem, the formula for finding the posterior probability is

\[
P(C | X) = \frac{P(C)P(X | C)}{P(X)}
\]

Naive Bayes assumes that each characteristic variable is independent of each other. Given a
category C and a sample attribute X, the conditional independence hypothesis can be expressed as

$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i)$$  \hspace{1cm} (3)

It can be seen from equations (2) and (3) that under the premise of conditional independence assumption, the calculation formula of posterior probability is

$$P(C \mid X) = \frac{P(C) \prod_{k=1}^{n} P(x_k \mid C)}{P(X)}$$  \hspace{1cm} (4)

In the process of calculating the probability, if there are too many eigenvectors, and the probability of some eigenvectors is extremely small, these extremely small numbers are multiplied, so that the probability of multiplication is close to 0, which leads to the above equation (4). Underflow, so the classification result is not accurate. In this paper, the natural logarithm method is adopted, and the case where the original probability is multiplied is converted into the probability addition by log fitting, thereby solving the underflow problem. Since P(X) is fixed, it is only necessary to compare the size of the molecule. Equation (4) can be converted to

$$C_h = P(C_h \mid X) = \ln P(C_h) + \sum_{k=1}^{n} \ln P(x_k \mid C_h)$$  \hspace{1cm} (5)

$$C_s = P(C_s \mid X) = \ln P(C_s) + \sum_{k=1}^{n} \ln P(x_k \mid C_s)$$  \hspace{1cm} (6)

C_h and C_s respectively indicate that the classification result belongs to the posterior probability of normal mail and spam, and the size of the above test can be used to obtain the category of the test mail.

2.3 Combination algorithm

Based on the combination of random forest and naive Bayesian algorithm, this paper proposes a combined classifier, and shows that the improved classifier has higher accuracy and shorter test time. The specific algorithm flow is shown in Figure 1

![Figure 1](image_url)

Figure 1. Filtering flow chart after improving the algorithm.

The algorithm steps are described as follows:
1) The random forest calculates the importance of each feature through Gini's impureness and stores it in a dictionary.
2) Determine the feature value to be culled. This paper chooses to eliminate the feature value with the importance of 0 to get a new feature set.
3) Find the conditional probability and prior probability of each feature word in the new feature set.
4) Find the posterior probability using the naive Bayesian algorithm, compare the size, and get the category of the test mail.

3. Experiment and analysis of results

3.1 Evaluation criteria and experimental environment

The evaluation criteria used in this paper are shown in Table 1.
### Table 1. Evaluation standard table

| result                  | The system determines that it is ham mail | The system determines that it is spam |
|-------------------------|------------------------------------------|--------------------------------------|
| Actually ham mail       | TN                                       | FP                                   |
| Actually spam           | FN                                       | TP                                   |

Accuracy rate: In simple terms, the accuracy rate is the proportion of the correct result predicted by the classifier. The formula is as follows:

\[
Accuracy = \frac{TP + TN}{TP + FP + TN + FN}
\]

Precision rate: The percentage of the correct classification made by the classifier in a category and the total classification of the classifier made on that class. Calculated as follows

\[
Precision = \frac{TP}{TP + FP}
\]

Recall rate: The percentage of the correct classification made by the classifier in a message category and the actual number of classifications in that class. Calculated as follows

\[
Recall = \frac{TP}{TP + FN}
\]

### 3.2 Experimental comparison

The experimental data is derived from the spam database in the UCI machine learning database. In order to better compare the improved algorithm and the naive Bayesian algorithm in the small sample spam filtering, the UCI machine learning database is randomly selected, 300 messages in the training, 100 emails were tested.

#### Table 2. Naive Bayesian filter mailing list

| result                  | The system determines that it is ham mail | The system determines that it is spam |
|-------------------------|------------------------------------------|--------------------------------------|
| Actually ham mail       | 49                                       | 4                                    |
| Actually spam           | 11                                       | 35                                   |

#### Table 3. Improved Naive Bayesian filter mailing list

| result                  | The system determines that it is ham mail | The system determines that it is spam |
|-------------------------|------------------------------------------|--------------------------------------|
| Actually ham mail       | 49                                       | 4                                    |
| Actually spam           | 10                                       | 36                                   |

#### Table 4. Comparison of classification performance

|                          | Precision rate /% | Recall rate /% | Accuracy rate /% |
|--------------------------|-------------------|----------------|------------------|
| Naïve Bayesian algorithm | 89.74             | 76.09          | 84.85            |
| Improved Naïve Bayesian algorithm | 90.00             | 78.26          | 85.86            |

Table 4 shows that the improved algorithm is superior to the original algorithm in system performance, the precision is 90%, 0.26% higher than the original algorithm, and the recall rate is 78.26%, which is higher than the original algorithm 2.17%, the accuracy rate reached 85.86%, which was 1.01% higher than the original algorithm, and the classification effect was better, which proved the effectiveness of the improved algorithm.

### 4. Conclusion

In view of the importance and independence of feature words in the process of spam training, the use of random forest feature selection method to select the optimal feature value for model establishment can achieve higher accuracy and better effect, it is useful for spam filtering technology.
Acknowledgments
This paper is supported by the National Natural Science Foundation under Grant No.61603242; Jiangxi provincial collaborative innovation center of economic crime investigation and prevention and control technology open project under Grant No. JXIZXTCX-030.

References
[1] Wang B., Pan W.F. (2005) Overview of Content-based Spam Filtering Technology. J. Chinese Journal of Information Science., 19(5): 3-12.
[2] Chen L., Liang Y.W. (2018) Tan Chengyu. Spam Detection Based on Adaptive Classifier. J. Computer Engineering., 44(5): 194-200.
[3] Liu Y.F., Zhang Y.B., Yuan J.H. (2018) Research on spam filtering of NB algorithm in cloud environment. J. Microelectronics & Computer., 411(8):60-63.
[4] Fan Y.T., Lai H.C. (2008) A spam filtering method based on SVM algorithm. J. Computer Engineering and Applications., 44(28): 95-97.
[5] Zhang J.L., Zhang F. (2007) Application of Improved KNN Algorithm in Spam Filtering. J. Modern Library and Information Technology., (4): 75-78.
[6] Xu M.L., Huang J.W. (2018) Application of Naive Bayesian Algorithm in Spam Filtering. J. Network Security Technology and Application., 211(7):49-50.
[7] Wang H., Huang Z.W., Liu S.F. (2015) A new weighted rough and naive Bayesian algorithm and its application,. J. 290 (12): 154-158.
[8] Wang H.L., Zheng G. (2013) Application of Improved Bayesian Algorithm in Spam Filtering. J. Information and Communications., (9):91-92.
[9] Bu H.L., Xia J., Zheng S.Z. (2014) Spam Detection Based on Feature Extraction. J. Journal of Chaohu University., (3): 28-31.
[10]Lei J.G., Sun X.B. (2013) Simulation Research of a Smart Spam Filtering Model. J. Computer Simulation., 30(5): 370-373.
[11]Li H., Li Z. (2015) An Improved Random Forest Algorithm Based on Integrated Non-Returning Sampling. J. Computer Engineering and Science., 37(7): 1233-1238.
[12]Breiman L. (2001) Random forests. J. Machine Learning, 45(1): 5-32.
[13]Guo S.Q., Gao C., Yao J., et al. (2005) Intrusion Detection Model Based on Improved Random Forest Algorithm. J. Journal of Software., 16(8): 1490-1498.
[14]Wang Y.S., Xia S.T. (2018) Overview of Random Forest Algorithms for Integrated Learning. J. Information and Communication Technology., 49-55.