International Journal of Computer Applications Technology and Research
Volume 9–Issue 08, 246-248, 2020, ISSN:-2319–8656

Application and Research of Naive Bayes Algorithm in Spam Filtering

Jun Li
School of Software Engineering
Chengdu University of Information Technology
Chengdu,China

Abstract: With the advent of the 5G era, the scope of e-mail applications has become more and more extensive, but the various spam messages that follow have also caused more and more serious problems. Among the many existing methods for filtering spam, the probability-based Bayesian classification algorithm is simple and efficient, and the accuracy rate can reach about 90%. This article briefly introduces the Bayesian model, gives an email filtering method based on the naive Bayesian classification model, and briefly analyzes its advantages and disadvantages. Finally, its effectiveness is verified through experiments.
Keywords: 5G; Email; Spam Filtering; Naive Bayesian Classification Model

1 INTRODUCTION

With the popularization of the Internet, e-mail has been loved by many netizens due to its low price and convenient use. However, it has also caused a large number of spam emails to affect normal communication. According to Kaspersky Lab, in 2019, the third In the quarter, the average proportion of spam in global mail traffic was 56.26%. Among them, the top 5 spam source countries: China ranked first (20.43%), followed by the United States (13.37%) and Russia (5.60%). Fourth place is Brazil (5.14%) and fifth place is France (3.35%). It can be seen that the form of spam processing in my country is still not optimistic.

Many countries have formulated anti-spam laws and regulations, and my country has also formulated relevant legal provisions. However, due to interest-driven, currently spam has not been effectively curbed, but has a growing trend. In addition to national prevention and control, many mail servers use technical methods to filter spam [1], such as adding blacklists, adopting sensitive word filtering rules, and using whitelists. At present, the more popular spam filtering methods include decision tree[2], Boosting[3], K nearest neighbor[4], support vector machine[5], Bayesian principle, etc.[6].

This article mainly introduces the use of Naive Bayes algorithm to filter spam, and combines Adaboost to improve the algorithm [7].

2 NAIVE BAYES MODEL AND RELATED PRINCIPLES

2.1 Bayesian Principle

Bayesian principle is a method proposed by British scholar Bayes as early as the 18th century to apply observed phenomena to correct subjective judgments about probability distribution [8]. The theorem states that the probability of something happening in the future can be estimated by calculating how often it has happened. Using Bayesian algorithm to filter spam, first we prepared 5574 samples, used cross-validation, randomly selected 4574 as training samples, generated a vocabulary list (corpus), tested and calculated the average error rate of classification for 1000 test samples .

2.2 Bayesian Classifier

The naive Bayes classifier uses the "attribute conditional independence hypothesis": assuming that all attributes are independent of each other, based on the attribute conditional independence hypothesis. Assume that the words contained in the content of the email are Wi, Spam, and ham. To judge an email, the word contained in the content is Wi, to judge whether the email is spam, that is, to calculate the conditional probability of P(S|Wi). According to Bayesian formula:

$$\Pr(S|W) = \frac{\Pr(W|S) \cdot \Pr(S)}{\Pr(W|S) \cdot \Pr(S) + \Pr(W|H) \cdot \Pr(H)}$$

among them:

• Pr(S|Wi) The conditional probability that a message with the word Wi appearing is spam (that is, the posterior probability);

• Pr(S) The probability of spam in the mail data set during the training phase, or the probability of spam actually investigated (ie, prior probability);

• Pr(Wi|S) the probability of the word Wi in spam emails;

• Pr(H) The probability of normal mail in the mail data set during the training phase, or the probability of normal mail actually investigated;

• Pr(Wi|H) The probability that the word Wi appears in a normal email;

For all words appearing in the email, considering the independence of each word occurrence event, calculate the joint probability Pr(S|W) of Pr(S|Wi), W={W1, W2,...Wn}:

$$p = \frac{p_1 p_2 \cdots p_N}{p_1 p_2 \cdots p_N + (1-p_1)(1-p_2)\cdots(1-p_N)}$$

Among them:-P is Pr(S|W), the conditional probability of spam when the word W={W1, W2.....Wn} appears;-Pi is Pr(S|Wi), the word Wi appears Is a conditional probability of spam.

3 ALGORITHM IMPROVEMENT

Use Bayesian formula to classify emails, calculate Pr(S|W) and Pr(H|W), compare the size of Pr(S|W) and Pr(H|W), and judge whether it is spam or normal email . We find that Pr(S|W) and Pr(H|W) calculate the same denominator, so we only need to compare the numerators. But there are still two problems: 1. When the vocabulary does not exist, that is, ni=0, at this time Pr(S|Wi) = 0, it will cause P=0, which cannot be compared; 2. When Pr(S| When Wi) is small, multiplying operations will cause underflow problems.

3.1 Solution

To solve these two problems, we have adopted the following solutions: 1. When calculating P(Wi|S) and P(Wi|H), initialize the number of occurrences of all words to 1, and initialize the denominator to 2 (or Adjust the denominator value according to the sample/actual survey results); 2. When calculating P(Wi|S) and P(Wi|H), take the logarithm of the probability.So the final comparison is, P(W1|S)P(W2|S)...P(Wn|S)P(S) and P(W1|H)P(W2|H)....P The size of (Wn|H)P(H).

Test effect: 5574 samples, using cross-validation, randomly selected 4574 as training samples to generate a vocabulary list (corpus), for 1000 test samples, the average error rate of classification is about 2.5%.

3.2 Improve The Algorithm Combined with Adaboost

When we calculate the joint posterior probability of ps and ph, we can introduce an adjustment factor DS, whose function is to adjust the "spamicity" of a word in the vocabulary, where DS is iteratively obtained by the Adaboost algorithm to obtain the best value. The process is as follows:

Step 1 Set the number of adaboost cycles count;

Step 2 Cross validation randomly select 1000 samples;

Step 3 DS is initialized to an all-one vector equal in size to the vocabulary list;

Step 4 Iterate the loop count times:

Step 4.1 Set the minimum classification error rate inf

Step 4.2 For each sample:

Step 4.2.1 Classify the sample under the current DS

Step 4.2.2 If the classification is wrong:

Step 4.2.2.1 Calculate the degree of error, that is, compare the alpha difference between ps and ph

Step 4.2.2.2 If the sample was originally spam, it was classified as ham by mistake:

Step 4.2.2.2.1 DS[Vocabulary contained in sample] = np.abs(DS[Vocabulary contained in sample]-np.exp(alpha) / DS[Vocabulary contained in sample])

Step 4.2.2.3 If the sample was originally ham, it was classified as spam by mistake:

Step 4.2.2.3.1 DS[Vocabulary contained in sample] = DS[Vocabulary contained in sample] + np.exp(alpha) / DS[Vocabulary contained in sample]

Step 4.3 Calculate the error rate

Step 5 Save the minimum error rate and the vocabulary list at this time, P(Wi|S) and P(Wi|H), DS and other information, that is, save the information of the best trained model

Test effect: 5574 samples, get the best model information for Adaboost algorithm training (including vocabulary list, P(Wi|S) and P(Wi|H), DS, etc.), for

1000 test samples, the average error rate of classification Approximately: 0.5%.

4 CONCLUSION

The harm of spam is self-evident, so our treatment of spam is also urgent. This article introduces the application of the Naive Bayes algorithm in spam processing, introduces the content and working principle of the Naive Bayes algorithm model in detail, and finds that underflow and probability of 0 occur during the construction of the algorithm model In response to the discovered problems, we found out the corresponding solutions, and finally implemented the coding combined with the improved Adaboost algorithm to greatly reduce the average error rate of spam classification.

5 REFERENCES

[1] Yang Shan, He Yue, Yan Jinjiang. Discussion on anti-spam technology based on Bayesian [J]. Network Security Technology and Application, 2007 (08): 54-56.

[2] Zhang Fuzhi, Wu Chaohui, Yao Fang, et al. Research and improvement of spam filtering technology based on Bayesian algorithm [J]. Journal of Yanshan University, 2009, 33(1): 47-52.

[3] Wang Qingsong, Wei Ruyu. Phrase-based Bayesian Chinese spam filtering method [J]. Computer Science, 2016, 43 (04): 256-259+269.

[4] Qu Meiting, Yu Jingxiao, Bu Wei, et al. Interactive architectural design model based on semantically guided Bayesian algorithm[J]. Bulletin of Science and Technology, 2016, 32(5): 133-136.

[5] Zheng Wei, Shen Wen, Zhang Yingpeng, etc. Research on Spam Filter Based on Improved Naive Bayes Algorithm [J]. Journal of Northwestern Polytechnical University, 2010, 28 (4): 622-627.

[6] ZHAN Chuan，LU Xianliang，ZHOU Xu，et al. Spam filtering method based on Bayesian formula［J］. Computer Science，2005，2（32）：73-75.

[7] Cao Ying, Miao Qiguang, Liu Jiachen, et al. Research progress and prospect of AdaBoost algorithm［J］. Acta Automatica Sinica, 2013, 39(6): 745-758.

[8] Ma Xiaolong. Research on an improved Bayesian algorithm in spam filtering[J]. Application Research of Computers, 2012, 29(3): 1091-1094.