Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/15/20230844

Spam filter based on naive bayes algorithm

Mengyuan Han

The department of International Software Engineering, Dalian University of Technology, Dalian, China

983822294@mail.dlut.edu.cn

Abstract. The widespread use of Electronic Mail (E-mail) has led to a significant increase in spam, which has severely impeded the growth and well-being of the Internet. To mitigate this issue, the implementation of email filtering techniques has become necessary, requiring the use of specific technological tools. Presently, the K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Naive Bayes (NB) algorithms are commonly used in probability statistical classification methods for email filtering. Among these, the NB algorithm is the most classical, with its rich mathematical theory as the basis, high classification efficiency, and straightforward algorithmic approach. However, the algorithm relies on the conditional independence assumption, making the accuracy susceptible to the correlation between attributes. This study focuses on email filtering techniques based on the NB algorithm, conducting experiments to evaluate the classification accuracy and proposing feasible improvements to weaken the independence assumption. The experimental results demonstrated the effectiveness of the employed method.

Keywords: spam filter, naive bayes algorithm, machine learning.

1. Introduction

With the continuous advancement of the Internet, E-mail has gradually become one of the common means of communication in people's daily life. However, the resulting spam has occupied a large number of network resources occupied and disrupted the normal order of E-mail communication, which has gradually become a significant problem in the work of Internet governance.

Spam can be strictly defined as any unwanted mail that does not adhere to the recipient's acceptance preferences, encompassing advertisements, unsolicited electronic publications, promotional materials, and other forms of communication that the recipient has not requested before, as well as mails that cannot be rejected, those that conceal the sender's identity or address, and mails containing false information about the source, sender, or route. Spam imposes many harms, including occupying a large number of network resources, seriously affecting the normal mail service; Dealing with them is time-consuming, inefficient and frustrating; Be used by hackers, causing the attacked website network paralysis; Spreading harmful information, causing harm to the real society, etc.

According to the report issued by Symantec, an international authoritative network security company [1]. It shows that the global spam accounted for 87.4% of the total number of emails in 2009. Since 2007, spam has risen by an average of 15%. At present, China has become one of the most seriously harmed countries in the world by spam. According to a survey by the ITU, in 2003 alone, the

DOI: 10.54254/2755-2721/15/20230844

cost of dealing with junk mail reached 4.8 billion yuan, and in 2006, the annual loss caused by junk mail to the national economy has exceeded 10 billion yuan.

Common spam prevention measures include IP blacklist and whitelist, Real-time Blacklist List (FBL) and mail filtering. Filtering technology works by following an algorithm or rule to determine whether an email is spam or not. The original mail filtering technology used pattern matching algorithm rules, such as the search of keywords to find spam. Furthermore, regular expression is used to realize fuzzy matching. With the development of technology, people begin to use spam recognition algorithms, such as Bayesian algorithm, which mainly originates from information classification. It is an algorithm for classification by calculating a posterior probability. Because Bayesian filtering operates purely according to statistical rules, and because tags are created solely by users, spammers have no way of guessing how their filters are configured to effectively block all types of spam.

This study will deeply explore the spam filtering technology based on the Park Subayes algorithm, and calculate the accuracy of classification through experiments, as well as explore the possible improvement.

2. Bayesian classification technique

The problem of spam filtering is actually a binary classification problem. At present, the most widely used is Bayesian classification technology, which is also a good spam filtering technology.

2.1. Bayes' theorem

Bayes' theorem is A theorem about the conditional probability of random events A and B [1]. It was proposed by the famous British mathematician Bayes in the 18th century. Its fundamental idea is to find the probability of another event when the probability of one event is known. In mathematics, the probability of event A occurring in the presence of another event B is not necessarily the same as the probability of event B occurring in the presence of another event A.

In situations where the occurrence probability of an event cannot be determined, it is possible to calculate the occurrence probability of an event related to it and, through the application of attribute correlation theory, infer the occurrence probability of the event itself. According to this thought description, Bayes theorem can be widely applied to text classification.

The formula of conditional probability is:

$$P(A|B) = P(A) * \frac{P(B|A)}{P(B)} \tag{1}$$

It can be expressed as

posterior probability = prior probability * adjustment factor

According to the deformation, the formula of Bayes' theorem can be obtained as follows:

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n} P(B_j)P(A|B_j)} \tag{2}$$

2.2. Naive Bayes classifier

Bayesian classifier is a classifier based on Bayesian decision theory [2, 3], which is widely used in text classification at present. Text classification is to judge which category a text belongs to. When making classification, Bayes classifier calculates the probability of the posterior probability of the test text under different categories according to the samples of known text categories through Bayes' theorem, and compares the size of the posterior probability and selects the category of the value with a larger posterior probability as the category of the text. Bayesian reasoning model diagram can be found in Figure 1.

DOI: 10.54254/2755-2721/15/20230844

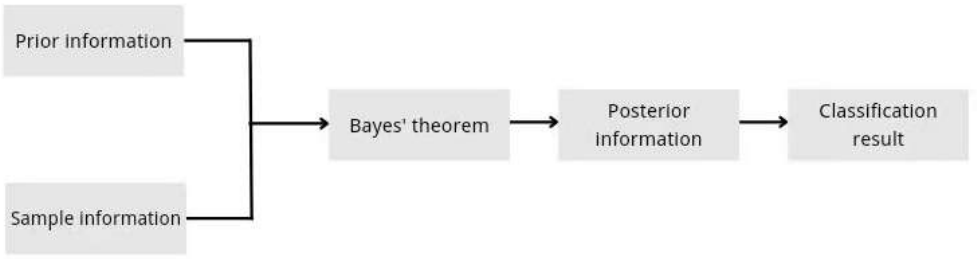

Figure 1. The procedure of the Bayes model.

In the field of classification, naive Bayes model is commonly used [4-5]. Naive Bayes algorithm is a certain improvement on the basis of Bayes algorithm, it is based on the conditional independence assumption between eigenvalues, the algorithm is very easy to understand, and has a high accuracy.

The difference between naive Bayes and Bayes is that it assumes that the attributes are independent of each other, which can be expressed by the following formula:

$$P(X|Y = c_k) = \prod_{j=1}^{n} P(x_j|y = c_k) \quad (3)$$

When determining the category, the following optimization formula is usually used:

$$y = \text{argmax}\, P(Y = c_k) \prod_{k=1}^{n} P(X_k|Y = c_k) \quad (4)$$

3. The experiment

3.1. Training procedure

First, the collected data set is labeled with ham or spam, where ham represents normal mail and spam represents spam. Then the data in each mail is preprocessed, the basic process is: identify and cut the words in the mail -- > the words out of the cut lowercase processing -- > the length of less than 3 words and stop words out -- > count the number of remaining words. The flow chart can be found in Figure 2.

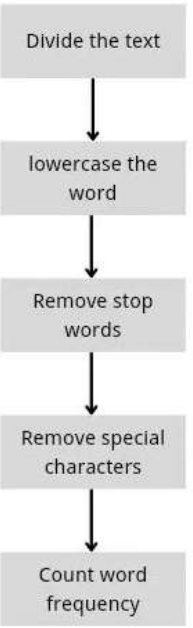

Figure 2. The procedure of the training process.

DOI: 10.54254/2755-2721/15/20230844

The classification is determined by naive Bayes algorithm. Calculate and compare the probability that the mail is classified as ham and spam. The higher probability is the category to which the mail belongs.

3.2. Test model

First prepare the data set, which contains 150 emails. Then divide the data set into 3 groups, with each group including 25 ham and 25 spam. From the 50 emails in each group, randomly select 40 for training and 10 as test set. Each group is tested for ten times. The results can be found in Figure 3 and Table 1.

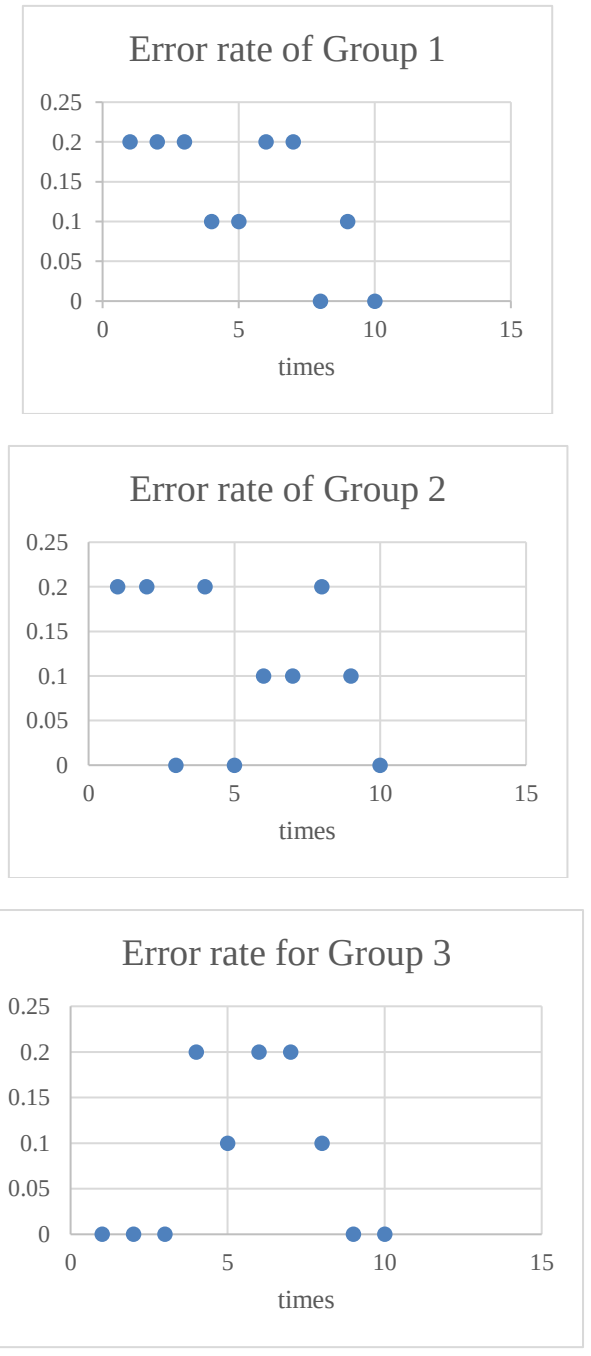

Figure 3. The error rate of three groups.

DOI: 10.54254/2755-2721/15/20230844

Table 1. The performance of the model.

Group Number	Group 1	Group 2	Group 3
Average error rate	0.13	0.11	0.18

The average error rate in each group is between 0 and 20%(0.13,0.11,0.18 respectively), which means Naive Bayes classifier has quite high accuracy.

4. Algorithm analysis

4.1. Advantage and disadvantage

Because of its rich mathematical theory, naive Bayes algorithm has high classification efficiency and classification performance, especially based on the assumption of feature independence, which simplifies the form of Bayes classifier. In addition, naive Bayes algorithm can solve multiple classification problems, the algorithm is simple, suitable for incremental training.

However, the corresponding disadvantage is that it depends on the assumption of conditional independence among features, which makes the influence of each attribute on the global must be independent of each other. But obviously this assumption is not in line with reality and is difficult to satisfy. Therefore, when there are too many attributes or the attributes are correlated, the efficiency of naive Bayes classifier will be reduced. In contrast, when the number of attributes is small or the correlation is not strong, the performance of classification will reach the best.

4.2. Possible improvement: Naive bayes classifier based on random forest

Random forest is a group of tree structure classifier, the basic unit is decision tree. Since random forest is the idea of integrated learning, it makes up for the shortcomings of overfitting and low precision caused by a single tree [6]. As a combination classifier that integrates a variety of weak classifiers to form a strong one, the advantages of random forest are high accuracy, not easy to overfit, simple implementation, etc.

Bagging algorithm belongs to a classical ensemble learning algorithm, which is generally used to solve regression and classification problems. The weight of Bagging algorithm is the same, and data is extracted in the way of sampling with retractions. The method is to randomly extract n samples from the sample set for each training session. If s training sessions are performed, s training sets will be generated. The prediction result will be decided by the vote of s classifiers obtained by training.

Random forest algorithm uses Bagging method to carry out the selected samples with retractions and generate random samples of training sets for each tree [7-10]. Since each new training set is constructed on the original training set, there are differences among training sample sets. In the construction process of decision tree, the optimal segmentation on the selected random features is used to segment nodes, and the grown tree does not need to be pruned, so that the training error rate of random forest is low, and the anti-noise ability is strong.

5. Conclusion

In the present age of swift advancement of Internet technology, email has become an integral component of individuals' daily lives. The pressing need for anti-spam measures to purify the network environment necessitates the application of email filtering technology, which has been identified as one of the most efficacious approaches. This study primarily focuses on the widely used Naive Bayes filtering algorithm and experimentally verifies its high accuracy in email classification. Additionally, it proposes that the independence assumption can be weakened by incorporating the Random Forest algorithm. Despite an ever-increasing interest in spam mails, the constantly evolving email filtering technology can effectively mitigate the inundation of unwanted mails.

DOI: 10.54254/2755-2721/15/20230844

References

[1] Malakoff D A 1999 Brief Guide to Bayes Theorem Science 286(5444) 1461-1461
[2] Zhang K Chen X Song Y et al. 2019 improved method for TAN method Computing technology and Automation (in Chinese) (01):55-61
[3] Lin S Tian F 2000 Research on Bayesian classifier for data mining Computer Science (in Chinese) 27(10) 73-76
[4] Lu L Wang J Wang C 2017 Research on data mining Classification Algorithm for cloud computing Microcomputers and applications (in Chinese) 36(06):7-9
[5] Ma G 2018 The improvement and application of Naive Bayes Algorithm (in Chinese) Hefei: Anhui University
[6] Wang L 2020 Research on spam filtering based on Bayesian (in Chinese) Shanghai: Shanghai University of Engineering Science
[7] Liu Y Xing Y 2019 Research and application of text classification based on improved random forest algorithm (in Chinese) Computer system application 28(05):222-227
[8] Breiman L 2001 Random Forests Machine Learning 45(01):5-32
[9] Wang Y Xia S 2018 A survey of random forest algorithms for ensemble learning Information and Communication Technology (in Chinese) 12(01):49-55
[10] Biau G Scornet E 2016 A random forest guided tour Test 25: 197-227