Study of Chinese spam filtering Based on Improved Naive Bayesian Classification Algorithm

Kaiying Zuo*

School of Tianjin University of Technology, Tianjin, China

*Corresponding author e-mail: zuokaiying@tjut.edu.cn

Abstract. Spam is a growing threat to mobile communications. This paper puts forward some mitigation technologies, including white list and blacklist, challenge response and content-based filtering. However, none are perfect and it makes sense to use an algorithm with higher accuracy for classification. Bayesian classification method shows high accuracy in spam processing, so it has attracted extensive attention. In this paper, a Bayesian classification method based on annealing evolution algorithm is introduced into Chinese spam filtering to improve the accuracy of classification. Our simulation results show that the algorithm has better performance in spam filtering.

1. Introduction

According to the spam and phishing in 2019 [1] released by Kaspersky on securelist on April 8, 2020. In 2019, spam accounted for 56.51% of the total mail, 4.03% higher than that in 2018. The largest source of spam is China, accounting for 21.26%. Therefore, how to effectively filter Chinese spam is a very important research topic.

Nowadays, e-mail has become an indispensable and important tool in people's daily life, which is mainly used for a variety of information interaction. The rapid development of e-mail and the development of related business make the number of e-mail users reach an amazing number, followed by the proliferation of spam on the network. Spam mainly comes from anonymous forwarding servers, one-time accounts and zombie hosts. In order to promote their products or promote their websites, some people achieve their goals by sending random e-mails, which makes spam far exceed the normal number of e-mails, occupies a large number of users' mailbox space and network bandwidth, affects users' use, and consumes the legitimate rights and interests of users. It brings heavy economic burden and social pressure to mail service providers [2]. This is not only a technical issue, nor just a policy and legal issue, but a global and comprehensive issue. Therefore, how to quickly and effectively solve the spam problem has great practical significance [3].

As thus, many experts propose to analyze the content of e-mail to identify spam. This is the task of combining spam filtering with text classification and information filtering, and introducing the methods commonly used in text classification and information filtering into spam filtering.

In this paper, naive Bayesian classification based on simulated annealing genetic algorithm is introduced into spam filtering. Compared with naive Bayesian method, the new method improves the accuracy of spam filtering, reduces the misjudgment rate of non-spam, achieves satisfactory results, and improves the quality of email filtering from the algorithm level.
2. Methodology

2.1. Overview of naive Bayesian classification

Naive Bayes relies on two basic assumptions to simplify the calculation: (1) all features in the data set are independent of class label conditions. (2) The prediction process is not affected by other potential variables. According to the first assumption, the following joint probability can be obtained:

\[
P(X_1 = x_1, \ldots, X_n = x_n, C = c) = P(C = c)P(X_1 = x_1, \ldots, X_n = x_n | C = c) = P(C = c) \prod_{i=1}^{n} P(X_i = x_i | C = c)
\]

(1)

Message category in spam determination: \( C \in C = \{0, 1\} \). Set 1 to indicate spam and 0 to indicate non-spam. The task of mail classifier is to calculate the probability that the mail to be classified is spam. If it exceeds a certain threshold, the mail is considered as spam. On the basis of Bayesian theory, for any mail sample \( t_x \):

\[
P(C_j | t_x) = \frac{P(C_j)P(t_x | C_j)}{P(t_x)}
\]

(2)

In formula (2), \( P(C_j | t_x) \) represents the probability that mail \( t_x \) belongs to class \( C_j \), \( P(C_j) \) is the class a priori probability, and \( P(t_x | C_j) \) is the class conditional probability. For the same sample, \( P(t_x) \) remains unchanged.

When \( C = 1 \),

\[
P(C = 1 | t_x) = \frac{P(C = 1)P(t_x | C = 1)}{P(t_x)}
\]

(3)

When \( C = 0 \),

\[
P(C = 0 | t_x) = \frac{P(C = 0)P(t_x | C = 0)}{P(t_x)}
\]

(4)

If \( P(C = 1 | t_x) > P(C = 0 | t_x) \), the email \( t_x \) is considered as spam, otherwise it is a non-spam. It is generally believed that users would rather accept multiple spam messages than miss a non-spam. This cost relationship often introduces a cost factor \( \lambda \). To describe, therefore, for the \( \lambda \) It is feasible to take a larger value, for example \( \lambda = 9999 \), that is, the loss of 9999 spam missed is equivalent to the loss caused by misreporting a non-spam as spam.

\[
\frac{P(C = 1 | t_x)}{P(C = 0 | t_x)} > \lambda
\]

(5)

Since \( P(C = 1 | t_x) = P(C = 0 | t_x) \), equation (5) can be expressed as:

\[
P(C = 1 | t_x) > \frac{\lambda}{1 - \lambda}
\]

(6)

If \( \lambda = 9999 \), then when \( P(C = 1 | t_x) > 0.9999 \), the mail \( t_x \) is considered as spam. This is the naive Bayesian classification method based on minimum risk.
2.2. Overview of Annealing Evolution Algorithm

2.2.1. Basic thought of Genetic Algorithm. The earliest proponent of genetic algorithm theory was American scientist holder, who deepened it according to Mendelian genetics and Darwin's theory of evolution and grasped the basic characteristics of information transmission; For example, Darwin believed that species would survive in the process of evolution and improve their adaptability to the environment; Although the nature of the species has not changed and still integrates the basic characteristics of the previous generation, there will be differences under the evolution of the species, adapt to the environment and realize the survival of the fittest and evolution. Mendelian genetic theory research shows that the genetic instruction code in the cell is contained in the chromosome, because the position of each gene is fixed and related to special properties, Each gene is also adaptive to the environment of future generations. Genes that adapt to survival can be selected by gene hybridization and gene mutation, and unsuitable individuals can be eliminated.

2.2.2. Basic thought of Simulated Annealing algorithm. In 1982, Kirkpatrick and others introduced the idea of annealing into the field of combinatorial optimization and proposed an effective approximate algorithm for solving large-scale combinatorial optimization problems, especially NP complete combinatorial optimization problems - Simulated Annealing algorithm. Compared with the previous approximation algorithms, it has the advantages of simple description, flexible use, wide application, high operation efficiency and less limited by initial conditions, and has high practical value [4].

Simulated annealing algorithm is a stochastic optimization technology used to solve continuous, ordered discrete and multimodal optimization problems. It simulates the cooling process of classical particle system in thermodynamics to solve the extreme value of programming problem. It regards the objective function of solving the problem as the energy function of the object, the solution of the problem as each state of the annealing process, and the optimal solution is the optimal solution after annealing. The lowest energy state. It imitates the solid annealing process and divides the whole process of the algorithm into three phases: heating process, that is, the process of setting the initial high temperature; isothermal process, that is, the search process at a certain temperature; cooling process, that is, the cooling process from the initial temperature to the final temperature. Through simulation calculation, the optimal solution or approximate optimal solution of the problem is finally obtained [5].

Simulated annealing algorithm is a general optimization algorithm. At present, it has been widely used in various fields and has become an optimization method with good development prospects. The specific algorithm steps are as follows:

Step 1: given the initial temperature \( t = t_0 \), randomly generate the initial state \( S = S_0 \), let \( k = 0 \);
Step 2: generate a new state \( S_1 \) by random disturbance, and calculate \( \Delta S = S_1 - S_0 \);
Step 3: if \( \min \left[ 1, \exp \left( -\frac{\Delta S}{t_k} \right) \right] \geq \text{random}[0,1] \), accept state \( S_1 \) as the current state;
Step 4: desuperheating \( t_k + 1 = \text{update}(t_k) \), make \( k = k + 1 \);
Step 5: if the end is not reached, return to step 2; otherwise, the algorithm ends and the result is output.

2.2.3. Annealing Evolution Algorithm. Simulated annealing algorithm (SA) and genetic algorithm (GA) are algorithms derived from some laws of nature. They are two branches of calculation according to natural laws. It is meaningful to study how to organically combine them and improve their efficiency, which is also in line with the essence of opposition and unity in nature. It combines SA and GA algorithms with complementary advantages, gives full play to SA local search ability and GA global search ability, and overcomes the problems of poor SA global search ability and low efficiency, poor GA local search ability and premature phenomenon.

Genetic algorithm can make the following improvements to simulated annealing algorithm: by dynamically adjusting the KNN subset, the capacity \( m \) of the KNN subset: \( M = \partial \left( f_{\text{max}} - f_{\text{avg}} \right) \).
$f_{\text{max}}$ is the maximum fitness value, $f_{\text{avg}}$ is the average fitness value, and $\theta$ is the coefficient. Then, in the preliminary phase of the algorithm, there is a big difference between the maximum fitness value and the average fitness value. Hence that the capacity of KNN subset is large. In the middle phase of the algorithm, the gap between the maximum fitness value and the average fitness value began to narrow. In the later phase of the algorithm, there is hardly any deviation between the capacity of the maximum fitness value and the average fitness value, and the capacity of KNN subset is getting smaller. In order to improve the operation efficiency of the algorithm, we should make it search in a more accurate range, so it is necessary to add dynamic adjustment parameters to the algorithm.

2.3. Bayesian classification method based on simulated annealing genetic algorithm

2.3.1. Fitness Function. Most genetic algorithms only use the objective function as the fitness function in evolutionary search. The evaluation of fitness function is the basis of selection operation, and the design of fitness function directly affects the performance of genetic algorithm. In specific applications, the design of fitness function should be combined with the requirements of solving the problem itself. For different problems, the selection of fitness function has many forms. On the basis of the treatment of objective function, fitness function can be divided into the following two categories: objective function mapping into fitness function and fitness function calibration.

In this paper, the fitness function is set as:

$$F = R + \lambda D$$  \hspace{1cm} (7)

Where, $R$ is the classification accuracy of NBC on the verification set, and $D$ is the difference of NBC on the verification set, $\lambda$ is the coefficient that determines the influence of the degree of difference.

2.3.2. Crossover and mutation operations. Crossover operation: generate $N = 30$ random numbers $r_{\text{nd}}(i = 1,2,\ldots,N)$ of $[0,1]$, judge each chromosome, if $r_{\text{nd}} \leq p_c$, cross chromosome $i$, randomly generate chromosome $j$ and random number $\alpha$, and the crossed chromosome is $\alpha X_i + (1-\alpha)X_j$, otherwise it will not be crossed.

Mutation operation: generate $N = 30$ random numbers $r_{\text{nd}}(i = 1,2,\ldots,N)$ of $[0,1]$, judge each chromosome. If $r_{\text{nd}} \leq p_m$, mutate chromosome $i$, randomly generate a direction $D$, and the mutated chromosome is $X_i + mD$ ($M$ is the coefficient). If the mutated chromosome is not feasible, reduce the value of $M$ until it is feasible. Otherwise, no variation will be carried out.

This paper adopts two-point crossover, that is, exchanging the middle part of the two loci of the male parent to produce the corresponding offspring. In order to avoid prematurity, the crossover probability and mutation probability are dynamically adjusted by adaptive method. Thus, if the fitness changes, the crossover probability $p_c$ and mutation probability $p_m$ will change automatically. When the fitness of each individual in the population incline to be consistent or reaches local optimization, the $p_c$ and $p_m$ raise; When the population fitness is dispersed, $p_c$ and $p_m$ will decrease. Simultaneously, for individuals whose fitness value is exceed than the average fitness value of the population, the next generation will be protected; The individuals below the average fitness value correspond to higher $p_c$ and $p_m$, hence the solution is sifted out. The calculation expressions of the selected crossover probability $p_c$ and mutation probability $p_m$ are as follows:

$$p_c = \begin{cases} \left(1 - \frac{f' - f_{\text{avg}}}{f_{\text{max}} - f_{\text{avg}}} \right) p_{c1}, & f' \geq f_{\text{avg}} \\ \frac{p_{c2}}{p_{c1}}, & f' < f_{\text{avg}} \end{cases} \hspace{1cm} (8)$$
\[
p_m \begin{cases} 
    \left( p_{m1} - \frac{p_{m2}}{f_{\text{max}} - f_{\text{avg}}} \right) \left( f' - f_{\text{avg}} \right), & f' \geq f_{\text{avg}} \\
    \frac{p_{m2}}{f_{\text{max}} - f_{\text{avg}}}, & f' < f_{\text{avg}}
\end{cases} 
\]

(9)

Where, \( p_{c1} = 0.9, \) \( p_{c2} = 0.6, \) \( p_{m1} = 0.1, \) \( p_{m2} = 0.001, \) \( f' \) is the larger fitness value of the two individuals to be crossed, and \( f \) is the fitness value of the individual to be mutated.

2.3.3. Individual simulated annealing operation. The fitness is used as the energy in the simulated annealing genetic algorithm, and the KNN subset in the algorithm is selected by dynamic adjustment: \( M = \hat{c} \ast (f_{\text{max}} - f_{\text{avg}}) \). When the fitness value increases, accept the solution as the next current solution; otherwise, accept the solution with a certain probability \( p = \exp(-\Delta \text{fitness}/T) \), i.e:

\[
p' \begin{cases} 
    1, & f_{i+1} \geq f_i \\
    \exp\left( \frac{f_{i+1} - f_i}{T} \right), & f_{i+1} < f_i
\end{cases}
\]

(10)

Among them, \( f_{i+1} \) is the fitness value of the child, \( f_i \) is the fitness value of the parent, and the temperature \( T \) is borne by the control parameter that decreases its value with the process of the algorithm.

2.3.4. Select operation. The operation of selecting the superior individuals in the group and eliminating the inferior individuals is called selection. On the basis of the dimensions of chromosome fitness value, select the chromosome with stronger adaptability to generate a new population. Therefore, the greater the fitness value, the greater the probability of being selected. This paper adopts roulette selection method.

The roulette selection process is as follows:

1. Calculate the selected probability \( p(x_i) \) of each individual
2. Calculate the cumulative probability \( q(x_i) \) of each part
3. Randomly generate an array \( m \), and the value range of elements in the array is between 0 and 1, and sort them from small to large. If the cumulative probability \( q(x_i) \) is greater than the element \( m[i] \) in the array, the individual \( x(i) \) is selected. If it is less than \( m[i] \), the next individual \( x(i+1) \) is compared until an individual is selected.
4. If \( n \) individuals need to be transferred, repeat step (3) \( n \) times.

2.3.5. Termination condition. When the fitness of the optimal individual and the fitness of the population are not improved, the algorithm terminates. And when the fitness of the optimal individual reaches the given threshold, or the number of iterations reaches the preset algebra, the algorithm terminates.

2.3.6. Algorithm implementation. Stratified random sampling, initially divides the units in each group are divided into different types according to regular standards; Then, the number of sample units from each type is determined according to the proportion of the number of units to the total of units; Ultimately, according to the random principle, samples are taken from each type.

Step1: The database is divided into training set, verification set and test set by stratified random sampling method;
Step2: Randomly generate \( s \) random attribute subsets;
Step3: Genetic algorithm is used to optimize the NBC corresponding to \( s \) random attribute subsets as the initial population;
Step 4: Adjust $\lambda$, Repeat step 3.

Genetic algorithm uses selection operation to realize the survival of the fittest of individuals in the population: individuals with high fitness are more likely to inherit to the next generation than individuals with poor adaptability. The selection operation is to select some individuals from the parent population and inherit them to the next generation population.

3. Implementation and Experimental Analysis

Trec06 is a public spam corpus provided by the international text retrieval conference. It is divided into English data set (trec06p) and Chinese data set (trec06c). The mail contained in it comes from real mail and retains the original format and content of the mail. The training set, validation set and test set data used in this experiment are from trec06 data set. 3000 emails, 1500 training set data, 750 validation set data and 750 test set data (including 500 spam and 250 non-spam) are randomly selected for spam filtering. Simultaneously, in order to make the results more universal, the experiment also uses the 2006 TREC public spam corpus Chinese version trec06c as the experimental sample. Trec06c uses the first 10000 emails as the data source, including 6631 spam and 3369 non-spams.

In non-spam and spam classifications confusion matrix is used. In this case four possible cases are possible:

Table 1. Spam system decision distribution

| Reality | Spam | Non-spam |
|---------|------|----------|
| Spam    | True Positives | False Positives |
| Non-spam| True Negatives | False Negatives |

Precison = \( \frac{TP}{TP + FP} \)  \hspace{1cm} (11)
Recall = \( \frac{TP}{TP + FN} \)  \hspace{1cm} (12)

In our case, accuracy is to measure the accuracy of spam identification, and recall rate is a measure of broadness of correctness.

Table 2. Comparison of test results of two algorithms

| Contrast conditions | Naive Bayesian Classification | Improved Algorithm |
|---------------------|-----------------------------|--------------------|
| Precision           | 91.73%                      | 96.93%             |
| Recall              | 90.13%                      | 92.27%             |

It can be seen from table 2 that the improved Bayesian classification method proposed in this paper is superior to the naive Bayesian algorithm in both precision and recall in spam filtering. The data comparison of the experimental results can fully show that the method in this paper does improve the overall performance of spam filtering.

4. Conclusion

In this paper, naive Bayesian classification based on simulated annealing genetic algorithm is introduced into spam filtering. Compared with naive Bayesian method, the new method improves the accuracy of spam filtering, reduces the misjudgment rate of non-spam, achieves satisfactory results, and improves the quality of email filtering from the algorithm level. This paper has made some research achievements in the classification of spam, but there are still many deficiencies, which need to be further studied in the future. The main performance is that because the naive Bayesian algorithm introduces the hypothesis.
of "independence between features", and the meaning of Chinese words is related, there are some limitations in using naive Bayesian algorithm to study the classification of Chinese e-mail.

Acknowledgments
There are many people who want to thank today. Of course, the most grateful person is my teacher. I sincerely thank you for your training. Without your training, I won't have my achievements today. Then I want to thank my classmates. Thank you for your help. Finally, thank my family, relatives and friends for your concern and long-term support.

References
[1] Information on https://securelist.com/spam-report-2019/.
[2] Liu Yuefeng, Zhang Yabin, Yuan Jianghao. Research on spam filtering of Nb algorithm in cloud environment [J]. Microelectronics and computer, 2018, 35 (8): 60-63
[3] Zeng Yusheng. Detection and analysis of spam in ship communication network [J]. Ship science and technology, 2019, 41 (10): 160-162.
[4] Kang Lishan. Non numerical parallel algorithms Volume 1 - simulated annealing algorithm [M]. Beijing: Science Press. 1995 American Institute of Physics, Melville, NY, 2002, pp. 651–654.
[5] Wang Dingwei, Wang Junwei, Wang Hongfeng, et al. Intelligent optimization method [M]. Beijing: Higher Education Press. 2007.