Application and Research of Spam Classification Based on Cluster Intelligence Algorithm to Optimize SVM

Zhennan Xia¹, Jianzhi Deng²*

¹,²College of Information Science and Engineering, Guilin University of Technology
¹,²Guangxi Key Laboratory of Embedded Technology and Intelligent System
Guilin, China
384524117@qq.com, 175154606@qq.com

Abstract—With the development of science and technology, the increasing popularity of the Internet, email has been widely used. Because of its convenience and low cost, e-mail has been used by more and more criminals to maliciously spread commercial advertisements, computer viruses and bad information. In order to resist spam, this paper proposes a spam classification method based on swarm intelligence algorithm optimized SVM. It classifies English spam and extracts features with special sensitive word coding. The model and method in this paper have certain reference value for spam classification research.

1. INTRODUCTION
With the development of the times, the Internet has been popularized, and network communication has gradually become a common way of people's daily communication. At the same time, a large number of spam appears with the popularity of network communication. Spam is often accompanied by economic, advertising and other activities, is a bad product of normal mail [1]. The emergence of a large number of spam seriously affects the normal network environment, causing unnecessary trouble to Internet users, and the transmission of spam information needs to occupy a large number of network resources, storage of spam also needs to consume huge hard disk space, which causes a huge waste of resources. Therefore, spam filtering technology has become the main problem to be solved in the field of Internet information security.

At present, the research of spam filtering technology is mainly based on the content of email and its encoding results to complete the classification and identification of spam. For the content of e-mail spam classification and identification technology, including text segmentation, document representation, feature selection and classification technology. Generally, e-mail data is a high-dimensional sparse matrix. So dimension reduction is the key to improve the efficiency of spam filtering. Several commonly used feature extraction techniques, such as information acquisition, cross correlation and Chi method, etc. In this paper, the target e-mail is preprocessed and compared with the spam vocabulary established by ourselves. In this way, the target e-mail can be transformed into a matrix and marked as $X$.

At present, scholars at home and abroad have done a lot of research and experiments on the characteristics of spam. Spam detection methods are mainly divided into two main directions, one is based on the sending source detection technology, the other is based on the text content detection technology, among which, the spam identification technology based on the source of the email is...
judged by analyzing the sender's email address and network IP[2]. Both rule-based and content-based statistical methods belong to text content-based recognition technology [3]. Although the source of spam detection can be successfully identified before the mail is sent successfully, spam can not reach the mail path. However, email identification technology based on the source of e-mail can not identify all cases, and it needs a lot of manpower to establish blacklist and whitelist. Therefore, the coverage of this method is not very high, and it may be wrong to intercept non spam. Therefore, we should focus on the content identification method.

In this paper, based on the classical particle swarm optimization algorithm of swarm intelligence algorithm, it is necessary to optimize, and combined with the classical characteristics of SVM, then improve the overall process of SVM training, grouping and classification. The experimental results show that SVM based on particle swarm optimization achieves good results in text classification, which is gradually becoming a research hotspot at home and abroad. Therefore, the application scenario based on spam classification is not only of great significance in theory, but also has practical application value in real scenes.

2. BASIC IDEAS OF SPAM CLASSIFICATION

2.1 Preparatory work
In a classic complete E-mail message, it might contain a URL, an E-mail address, Numbers, symbols, and so on. While many e-mails contain similar types of content (for example, Numbers, other urls, or other E-mail addresses), the specific content (for example, a specific URL or symbol) in almost every E-mail will be different. Because spammers usually randomize urls, urls are basically different from one spam to another, so the probability of seeing a particular or the same URL again in a lot of spam is very small. Based on the above conditions, the classification and identification process of spam will depend more on URL than URL specificity. This helps to improve the performance of the spam classifier. Therefore, it is necessary to pre-process these E-mail messages, and the results should be normalized, including URL, number and so on. For example, in this article, a set of strings "httpAddress" is used to replace the URL that exists in an E-mail message, indicating that the URL exists in the message.

2.2 E-mail standardization
- Lowercase: converts the entire email to lowercase, avoiding the influence of upper and lower case letters in the title.
- Remove HTML Tags: all HTML tags are removed from the email. Many e-mails are usually in HTML format; this article's approach is to remove all HTML tags so that only the content is retained.
- Normalized URLs: all URLs are replaced with the text "http_address".
- Normalized email address: all email addresses will be replaced with the text "email_address".
- Normalize numbers: all numbers are replaced with the text "number".
- Normalized currency symbol: all currency symbols are replaced with the text "money".
- Stem analysis: words are reduced to stem form. For example, "interest", "interests", "interested" and "interesting" are replaced by "interest". Sometimes, the word stem analyzer actually removes other characters from the end, so "include", "includes", "included" and "including" will be replaced by "include".
- Delete non words: delete non words and punctuation marks. All blanks (tabs, line breaks, spaces) are replaced with a blank character.

2.3 Set up vocabulary coding table
Through the spam corpus, 5624 words were obtained by selecting all the words that appeared not less than 100 times. Given a vocabulary, map each word in the pre-processed E-mail to a list that contains an index of the words in the vocabulary. Look up the word in the vocabulary and determine if the word exists in the vocabulary. If the word exists, its index should be added to the corresponding variable. If
the word does not exist in the vocabulary, skip the word. Then form the word index table corresponding to the E-mail.

2.4 feature extraction
The author transforms each email into a vector in matrix $R^n$ for feature extraction. Use $n = \#$ in the above vocabulary for word mapping comparison. The characteristic seat $x_i \in \{0,1\}$ of the e-mail corresponds to whether the $i$-th word in the above vocabulary appears in the email. If the $i$-th word is in the email, then $x_i = 1$. If the $i$-th word is not in the email, then $x_i = 0$. The following eigenvectors are formed.

$$
\begin{bmatrix}
1 \\
\vdots \\
1 \\
0 \\
\vdots \\
1 \\
0 \\
1 \\
\end{bmatrix} \in R^n
$$

2.5 SVM training
After completing feature extraction, the next step is to load a pre-processed training data set, which will be used to train the SVM classifier. The data set prepared by the author contains 4000 training sets for spam and non-spam, as well as 1000 test sets. Converts the preprocessed raw E-mail into vector $x' \in R^{5624}$.

2.6 Main predictors of spam
After training, we can check the parameters to see how the classifier predicts spam, and find the parameters with large positive value ($y = 1$) in the classifier, and these parameters are the basic factors to judge spam. So, if an email contains high-frequency words such as "incest" "price" "investor" and "cassino" it classifies the approximate rate as spam.

2.7 Swarm intelligence optimization SVM
After the spam classification is encoded by specific rules, the features of the sample feature set are obvious, which are low dimension and high correlation between features. These two characteristics are very helpful for SVM to play its due effect. At the same time, the strategy of swarm intelligence optimization SVM can be used to effectively complete spam classification and identification from the perspectives of training sample reorganization and dynamic adjustment of training weight. In order to strengthen the pertinence of the learning process, it can also speed up the convergence speed of the algorithm and improve the classification accuracy. Swarm intelligence algorithm has generality, and the core process can be reflected in the more classic particle swarm optimization algorithm.

Combined with the existing similar mechanism system feature reorganization rules and the actual composition, and then according to the results of multi batch experiments, after induction and analysis, then targeted to complete the spam encoding feature set and corresponding samples reorganization and selection. On the one hand, in the spam classification system, based on the existing feature set to complete the re selection of samples, the essential role is to increase the influence of the samples with higher discrimination on the classification, that is, to increase the training weight. On the other hand, the training weight of the features with low discrimination is reduced, but it can not disappear.
completely, and there may be some hidden influence among the features according to the different data sets used. At present, there is no general method suitable for all spam classification and identification scenarios, so such hidden conclusions often need to be gradually summarized through a large number of implementations. Based on this feature, and combined with the actual needs of spam classification system, the author uses particle swarm optimization algorithm to upgrade the two aspects mentioned above, that is, from the perspective of training sample reorganization and training weight dynamic adjustment, the optimization results will be reflected in the two aspects of shortening the training time and improving the actual classification accuracy.

3. CORE TECHNOLOGY ANALYSIS

3.1 Principle of SVM

The basic idea of support vector machine is to construct a hyperplane to delimit the binary data [4]. In order to maximize the interval (the distance between the hyperplane and the nearest sample point), the following optimization problems are obtained:

- Use \{ (x_i, y_i) | (i = 1, 2, \ldots, N) \} as the sample space and use \( x_i \in R^d \) as the sample set for the training and input part. \( y_i \in R \) is the target output, \( i \) is the number of training samples. Hyperplane can be expressed as \( f(x) = \omega^T x + b \). When \( f(x) = 0 \), \( x \) is a point on the hyperplane.
- The basic type of SVM is obtained by maximizing the support vector interval:

\[
H(\omega, b) = \min_{\omega, b} \frac{1}{2} \omega^T \omega \quad \text{s.t.} \, y_i (\omega^T x_i + b) \geq 1 \quad (i = 1, 2, \ldots, n)
\]

Where

- \( \omega \) is the hyperplane direction parameter
- \( b \) is the parameter of dividing displacement
- \( i \) is the total number of samples

- From Lagrange multiplier method to dual problem

\[
L(\omega, b, \alpha) = \max_{\omega, b} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \\
\text{s.t.} \sum_{i=1}^{n} \alpha_i y_i = 0 \\
\alpha_i \geq 0 \quad (i = 1, 2, \ldots, n)
\]

Where

- \( \alpha_i, \alpha_j \) is Lagrange multiplier
- \( x_i, x_j \) is the code in the corresponding vocabulary
- \( y_i, y_j \) is the sample label and the value is \{-1,1\}

3.2 Improved particle swarm optimization algorithm

Particle swarm optimization (PSO) [5] is a swarm based intelligent optimization search method. Firstly, an initial population is generated. At this time, each particle in the feasible space is a solution, and they all move in the feasible space, and the corresponding fitness value is determined by the objective.
function. The flight direction and distance of each particle are determined by their motion speed. Through its unique memory ability, particles follow the optimal particles to search in the feasible space. In each iteration, particles update themselves through two extremum of local optimal solution $P_{id}$ and global optimal solution $P_{gd}$.

Suppose particle swarm optimization searches in an n-dimensional solution space, The population is recorded as

$$X = \{ X_1, X_2, \ldots, X_m \}$$  \hspace{1cm} (3)

The particle number is m and the location of each particle is recorded as

$$X_i = \{ x_{i1}, x_{i2}, \ldots, x_{im} \}$$  \hspace{1cm} (4)

The velocity of each particle is randomly generated, the velocity of each particle is recorded as

$$V_i = \{ v_{i1}, v_{i2}, \ldots, v_{in} \}$$  \hspace{1cm} (5)

The position of each particle is recorded as

$$P_i = \{ p_{i1}, p_{i2}, \ldots, p_{in} \}$$  \hspace{1cm} (6)

After finding the two optimal solutions, each particle updates its velocity according to the following formula.

$$v_{ida}^{i+1} = \omega v_{ida}^i + c_1 r_1 (p_{ida}^i - x_{ida}^i) + c_2 r_2 (p_{gd}^i - x_{ida}^i)$$ \hspace{1cm} (7)

$$x_{ida}^{i+1} = x_{ida}^i + v_{ida}^{i+1}, \quad i = 1, 2, \ldots, m$$ \hspace{1cm} (8)

Where

$\omega$ is inertia weight

$r_1, r_2$ is a random number between 0 and 1

$c_1, c_2$ is the learning factor

$\omega$ is a concept introduced by Shi et al. [6-7] in 1998. Its size determines the inheritance of the current velocity of particles. Proper selection can make particles have the ability of balanced search and development. There are several ways to select weights, such as fixed weights, time-varying weights, fuzzy weights and random weights [8]. The time-varying weight can be dynamically adjusted with the number of iterations. Moreover, in order to make the particles have strong search ability in the early stage and better development ability in the later stage of flight, the time-varying weight is selected. The weight range is $[\omega_{min}, \omega_{max}]$, and the maximum iteration number is $\text{iter}_{max}$. The formula is as follows:

$$\omega_i = \omega_{\max} - \frac{\omega_{\max} - \omega_{\min}}{\text{iter}_{\max}} \times i$$ \hspace{1cm} (9)

In order to prevent the particle from getting too fast, let $V \in [\omega_{min}, \omega_{max}]$

$$\begin{cases} v_{\max} = (X_{\max} - X_{\min}) / 2 \\ v_{\min} = (X_{\min} - X_{\max}) / 2 \end{cases}$$ \hspace{1cm} (10)

The flow chart of SVM based on particle swarm optimization is shown in Figure 1.
3.3 Summary
To sum up, this section is based on the classic PSO algorithm in swarm intelligence algorithm, based on which, appropriate upgrade is carried out. Due to the statistical characteristics of the results of the corresponding characters and string encoding methods in the corresponding e-mail, based on the introduction of the improved dynamic adjustment mechanism of random characteristics, the training pertinence in the improved PSO optimized SVM training process is strengthened, that is to say, the training weight of the samples with low training discrimination and good effect is appropriately increased, and the training differentiation is obvious. The training weight of the training samples is appropriately reduced, but can not be 0, that is to say, the distance between clusters increases, while the inter class spacing decreases, but it can not disappear completely.

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Experimental data set and reference toolbox
In this paper, the experimental data set comes from the network public data set, a total of 5000 samples are used, of which 4000 are randomly selected as the training set and 1000 as the test set. The samples have two labels: spam and non spam. In addition, with the help of libsvm3.23 toolbox developed by Chih Jen Lin, the relevant codes and experiments are completed, and the kernel functions are all RBF.

4.2 Illustration and analysis of experimental results
As shown in Figure 2, after particle swarm optimization (PSO), the convergence speed of the algorithm is faster, and it is better to find two parameters $c$ and $g$ in SVM. Through repeated experiments, the results show that when $c = 16.326$, $g = 1.2724$, the classification effect is the best.
The result of spam classification is shown in Figure 3.

As shown in Table 1, after PSO optimized SVM, the running time of the system is reduced by nearly half, and the classification accuracy is improved by 12.841%.

| TABLE I. | TABLE TYPE STYLES |
|----------|--------------------|
|          | Running time | Accuracy |
| Conventional SVM          | 158.771 | 82.424% |
| PSO optimized SVM          | 84.207  | 95.265% |

5. CONCLUSION

Aiming at the problem of garbage classification, this paper proposes a classification algorithm based on swarm intelligence optimization SVM. Finally, based on the following two aspects, the final spam classification results are improved: on the one hand, thanks to the characteristics and foundation setting of PSO or other swarm intelligence algorithms, based on the random mechanism mechanism mechanism of appropriate weight and dynamic adjustment, the overall training process can be improved to a certain extent, and then strong. On the other hand, compared with the original PSO algorithm, the improved PSO algorithm in this paper will be more prominent for the insight of the characteristics of the spam samples that are difficult to identify. Although to a certain extent, it is subject to the specificity of the coding results and data sets, it also has a certain promotion value.

By introducing particle swarm optimization, the convergence speed of the algorithm is faster, and it is better to find the two parameters and the effect of SVM. Compared with the conventional SVM, the running time of the system is reduced by nearly half, and the classification accuracy is improved by 12.841%.
REFERENCES

[1] Hu Rui. Research and improvement of Chinese spam filtering method based on Bayesian classification [D]. Tsinghua University, 2006.

[2] Zhao Jingkai, Zhang Jia, bu Hong, Huang Hao. Spam filtering based on letter source [J]. computer engineering and application, 2004 (09): 139-142.

[3] Zhao Xiaodan, 2014, a comparative study of spam technology

[4] TREvorHASTIE, TIBShIRANI R, FRIEDMAN J M. The elements of statistical learning[M]. 12th Springer series in statistic, 2017.

[5] G. T. Kaya and H. Kaya, "Optimization of SVM parameters using High Dimensional Model Representation and its application to hyperspectral images," 2014 22nd Signal Processing and Communications Applications Conference (SIU), Trabzon, 2014, pp. 642-645, doi: 10.1109/SIU.2014.6830311.

[6] Shi Y H, Eberhart R C. A Modified Particle Swarm Optimizer[C]// Evolutionary Computation Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Conference on. IEEE, 2002.

[7] Shi Y H, Eberhart R C. A Modified Particle Swarm Optimizer[C]// Evolutionary Computation Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Conference on. IEEE, 2002.

[8] Li Baosheng, Qin Chuandong. Research on price prediction of electric vehicle based on SVM multi classification based on particle swarm optimization [J / OL]. Computer science: 1-4 [2020-07-02].