A fusion algorithm model based on KNN-SVM to classify and recognize spam

Zengle Ge
Sichuan University Chengdu, China
201814141340@stu.scu.edu.cn

Abstract. Spam usually has the characteristics of large data dimension and large numbers of samples. At present, the common classification learning algorithms have their own advantages and disadvantages. Aiming at the classification tendency of K nearest neighbor algorithm (KNN) in the face of unbalanced sample data and the long time consuming of support vector machine (SVM) training model, a fusion algorithm of KNN and SVM is proposed. Firstly, the content of spam is extracted, and then the text preprocessing is transformed into sparse matrix. The k nearest neighbor samples of the test sample set are screened out by KNN algorithm, and then the SVM algorithm is used to train these screened samples, establish the separation hyperplane, identify the mail, and judge whether it is spam or not. The results show that the KNN-SVM fusion algorithm can effectively reduce the impact of sample imbalance while maintaining the high classification accuracy of SVM algorithm, and its classification efficiency is slightly lower than that of KNN, algorithm but much higher than that of SVM algorithm.

Key words: spam; sparse matrix; K nearest neighbor algorithm; support vector machine; feature sample data.

1. Introduction

With the rapid development of the Internet, our lives are also full of a lot of information and data, the network has gradually become the mainstream way of daily communication and communication [1]. In daily life, people often send email to communicate with each other, but at present, a large numbers of spam is full of our lives, which has brought a lot of trouble to daily life and work, and even brought security risks. Therefore, efficient and accurate detection and classification of spam is the topic of current concern.

For spam detection and classification, the commonly used algorithms are naive NB algorithm, support vector machine algorithm (SVM), neural network algorithm, K nearest neighbor algorithm (KNN) and so on [2] [3] [4] [5]. The classifiers built by different algorithms have different effects. K nearest neighbor algorithm is one of the most commonly used algorithms. Its advantages are that the principle is simple and the classification efficiency is high, but in the face of unbalanced samples, it is easy to produce errors and can not accurately identify the categories of small sample data. Support vector machine (SVM) algorithm is the linear classifier with the largest interval defined in feature space. It can distinguish the categories of samples by constructing separation hyperplane. The accuracy is high, but...
the training time is long [6]. Therefore, according to the advantages and disadvantages of the two algorithms, a fusion algorithm model of KNN-SVM is proposed [7].

In theory, KNN-SVM fusion algorithm can reduce the training time of SVM algorithm and maintain the accuracy of classification. The core of SVM algorithm is to construct separated hyperplane. For a large numbers of training samples, the core is the support vector. KNN-SVM fusion algorithm selects the feature sample data adjacent to the test sample through KNN, and replaces the whole sample set with feature sample, which greatly reduces the numbers of training samples of SVM. Because the feature sample is adjacent to the test sample, the support vector is very likely to be in the feature sample. Therefore, the constructed hyperplane can classify and identify the test samples well, and then replace this kind of samples into the SVM training model to get the classification results [8]. In order to verify the performance of the algorithm, we choose the English mail set with a large numbers of text data as the sample, randomly select a part as the test sample set, and the rest as the training sample set. At the same time, compared with the common classification algorithm, the performance of the algorithm is compared to verify its advantages and disadvantages.

2. Related work

The detection and classification of text files is the focus of the current field, and there have been a lot of research in this area. In reference [9], a text classification method based on KNN+ hierarchical SVM is proposed, in which a multi-class sparse hierarchical SVM classifier is constructed, which is very large in text content and can be classified accurately and effectively, and the time consumption is about the same as the conventional classification algorithm. However, when the text structure does not have a significant level or distinguishes the single sample, its classification effect is not different from that of the general SVM algorithm. In reference [10], a research method combining KNN with SVM is proposed. In order to avoid the deviation of KNN classification of boundary sample data, the subclass of mail set is obtained by K-means clustering, and then the center vector of each subclass is calculated, then the category of each subclass is judged by KNN, and then the probability transformation of the output result of SVM is carried out by using sigmoid function, and the reliability analysis is carried out to make the decision. This method can reduce the sample data and improve the accuracy of classification at the same time. Compared with the single KNN), the training of binary classifiers is obtained by different samples alone, which can not guarantee the cooperation of classification results, and because of the superposition of multiple algorithms, it has a certain impact on the accuracy of classification results.

3. Preprocessing of mail samples

For a large numbers of e-mail sets, it is difficult to identify directly which are spam and which are normal. We first analyze the content contained in the message, standardize the content, such as turning all words into lowercase, converting all numbers to "numbers", deleting irregular characters, etc., then creating a vocabulary list to include all the words that have appeared, and establishing a sparse matrix of each message through the vocabulary list (the word appears in this message is 1, if it does not appear, it is 0). Thus, the abstract English words are transformed into a matrix with a certain dimension, and the next algorithm is prepared for the application of the algorithm.

4. KNN-SVM fusion algorithm model

4.1. KNN algorithm

KNN is one of the most commonly used classification algorithms. KNN algorithm is a statistical algorithm based on case learning. The main idea is to select the nearest K nearest neighbor samples in the input sample space, then count the categories of the K samples, and select the cate-gories with the largest numbers of times as the categories of the input samples.
Assuming that there are n, c categories of samples used as training sets, k nearest neighbor samples are selected from n samples, and then their respective categories are counted to judge the categories of test set samples. Distance similarity is measured using Euclidean distance:

\[ d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \]  

Classification discriminant function:

\[ f_i(x) = \max_{k_i} \quad k_i \in c \]

\[ k_i \]

The formula is k nearest neighbor samples for each test sample.

4.2. Support vector machine algorithm

The basic algorithm model of SVM is the linear classifier with the largest interval defined in the feature space, that is, to solve the separated hyperplane which can correctly divide the training data set and the largest geometric interval, and then use the hyperplane as the separation plane, and the two sides of the hyperplane are two different types of samples. For the maximum interval to be sought, that is, the geometric interval between the sample with the smallest distance from the hyperplane to the hyperplane is the largest, the larger the interval is, the smaller the generalization error will be, and the stronger the classification and prediction ability of the new data will be.

\[ (x_i, y_i) \ i = 1,2,...,N; \ x_i y_i \]

Set the training sample set as the sample feature vector, as the sample label, that is, \{-1:1\}, take-1 as spam and 1 as normal mail. The separation hyperplane is

\[ w^T x + b = y \]  

\[ w, b \]

The SVM parameter is in the Formula.

Because there is almost no completely linear divisible data in practice, the "relaxation variable" and the 'penalty parameter' are introduced to obtain:

\[ \min_{w,b,\varepsilon_i} \frac{1}{2}||w||^2 + C \sum_{i=1}^{m} \varepsilon_i \]

s.t. \( y_i(w \cdot x_i + b) \geq 1 - \varepsilon_i \)

\[ \varepsilon_i \geq 0, \ i = 1,2,...,N \]

The Lagrangian function is obtained by Lagrange multiplier method, and then the dual problem is solved.

Select the penalty parameter C and construct the solution:

\[ \min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N} \alpha_i \]

s.t. \( \sum_{i=1}^{N} \alpha_i y_i = 0 \)

\[ 0 \leq \alpha_i \leq C, \ i = 1,2,...,N \]

Get the optimal solution \( \alpha^* = (\alpha_1^*, \alpha_2^*, ..., \alpha_N^*)^T \)

The following is calculated on behalf of the agent:

\[ w^* = \sum_{i=1}^{N} \alpha_i^* y_i x_i \]

\[ b^* = y_i - \sum_{i=1}^{N} \alpha_i^* y_i (x_i \cdot x_j) \]

The separation hyperplane is as follows:

\[ w^* \cdot x + b^* = 0 \]

The classification decision functions are:

\[ f(x) = sign(w^* \cdot x + b^*) \]

The decision functions for nonlinear classification are as follows:

\[ f(x) = sign(\sum_{i=1}^{N} \alpha_i^* y_i K(x, x_i) + b^*) \]

\( K(x, x_i) \) The formula is a kernel function.

Using the Gaussian kernel function:
\[ K(x, z) = \exp\left(-\frac{||x-z||^2}{2\delta^2}\right) \quad (16) \]

The final classification decision function can be obtained as follows:

\[ f(x) = \text{sign} \left( \sum_{i=1}^{N} \alpha_i^* y_i \exp\left(-\frac{||x-z||^2}{2\delta^2}\right) + b^* \right) \quad (17) \]

**Figure 1.** Schematic diagram of maximum interval hyperplane and support vector

### 4.3. KNN-SVM fusion algorithm

Generally speaking, there are many commonly used algorithms for spam detection and classification. Compared with other algorithms, KNN is simpler and has a short operation time, but it is insensitive to feature sample data and prone to errors when the sample is unbalanced. In contrast, SVM is more accurate, but it takes a lot of time and space to store training samples and kernel matrices. Based on the advantages and disadvantages of the two, a fusion algorithm of KNN and SVM is proposed: first, the K value of the best classification effect in KNN is found, then the k nearest neighbor samples of the selected test samples are screened twice, the duplicate data are filtered out, and a new training set is formed, and then the new training set is used to train the new training set to obtain the classification model of spam.

The flow of the KNN-SVM fusion algorithm is as follows:

**Figure 2.** Working flow chart of KNN-SVM fusion algorithm

### 5. Results and analysis

In order to study the effect of KNN-SVM fusion algorithm, we set up a comparative experiment. The comparative experiment is set as KNN-SVM fusion algorithm and KNN, SVM alone, and set the
accuracy, recall rate, training time $t$ as the evaluation index. The accuracy, recall is defined as follows:

Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

(18)

Recall rate:

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\%$$

(19)

In TP:, normal mail is predicted as normal mail; TN: forecasts spam as spam; FP: forecasts spam as normal mail; and FN: forecasts normal mail as spam.

5.1. Experimental data set and Experimental platform

The experimental data is the spam classification dataset in the UCI data source library. Ham and spam are the labels of the message text, marked in the first column of each message, where ham represents normal mail and spam represents spam. There are 5572 messages in the experimental data set, and there are 7204 dimensions after converting them into sparse matrices. 1393 messages are randomly selected from the messages as test samples, and the rest are training samples.

The experimental platform chooses spyder4.0, in Anaconda3 to choose Python as programming language.

5.2. Experimental results and analysis

(1) After calculation and verification, it is found that when the $K$ value increases, the accuracy rate decreases gradually, which is because the larger the $K$ value is caused by the imbalance of the sample data, the greater the tendency of the large class data in the classification discrimination is, and the accuracy rate will decrease accordingly, but the $K \geq 1-10:00$ can be kept above 89%, among which $K \leq 1-3$ has the best effect and the accuracy can reach 94%.

(2) The KNN-SVM fusion algorithm has been verified:

The KNN-SVM fusion algorithm is used to establish the training model of the mail set, and then the test set samples are classified. The accuracy of the KNN-SVM fusion algorithm model is (Accuracy), the recall rate (Recall) and the numbers of feature samples (numbers) are changed when the $K$ value is 1-10:00.

![Variation of KNN-SVM accuracy, recall rate and numbers of feature samples from 1 to 10:00 with K value](image)

**Figure 3.** Variation of KNN-SVM accuracy, recall rate and numbers of feature samples from 1 to 10:00 with K value
Table 1. When K is taken from 1 to 30, the data of each index are obtained.

| K value | KNN-SVM Accuracy | KNN-SVM Recall | The numbers of feature samples |
|---------|------------------|----------------|--------------------------------|
| 1       | 89.088299        | 100            | 569                            |
| 2       | 94.328787        | 100            | 836                            |
| 3       | 91.17014         | 100            | 990                            |
| 4       | 91.957999        | 100            | 1108                           |
| 5       | 93.826274        | 100            | 1202                           |
| 6       | 92.893037        | 100            | 1288                           |
| 7       | 95.692749        | 100            | 1337                           |
| 8       | 97.702800        | 98.592715      | 1381                           |
| 9       | 97.559225        | 88.162252      | 1423                           |
| 10      | 97.487437        | 100            | 1455                           |
| 10-15   | 96.439345        | 98.377483      | 1550                           |
| 15-20   | 97.415652        | 99.917219      | 1632                           |
| 20-30   | 97.415658        | 99.834438      | 1807                           |

It can be seen that when K is taken from 1 to 10:00, the accuracy and accuracy of KNN-SVM fusion algorithm are on the rise, and the accuracy of K is above 97%. The recall rate does not change significantly with K value. The numbers of feature samples increases significantly with the increase of K value when K value is small, but when K is above 10, the average value of K value 10-15, 15-20, 20-30 is taken. The growth rate of feature samples slowed down obviously, and the indexes changed little. It can also be proved that the sample characteristics which are close to the test sample space are more obvious, that is, the probability of support vector is also greater, and with the increase of K value, the increment of the numbers of feature samples is very small, and the influence on the accuracy can be ignored.

In order to compare the performance difference between KNN-SVM, SVM and KNN, the accuracy and recall rate of KNN (K ≤ 3) and SVM alone were tested, and the results are as follows:

Table 2. Accuracy and recall rate of KNN and SVM

| Algorithm | Accuracy | Recall    |
|-----------|----------|-----------|
| KNN       | 91.888011| 38.918919 |
| SVM       | 98.205312| 86.486486 |

It can be seen that when K value is appropriate, the accuracy of KNN-SVM fusion algorithm is little different from that of SVM, the recall rate is higher than that of SVM, and the accuracy of KNN-SVM is much higher than that of KNN. This shows that the ability of single KNN and SVM to classify and recognize normal mail is insufficient, and the information of mail can not be described accurately and fully, and the recall rate of the classifiers established is low. In addition, in the aspect of algorithm classification efficiency, KNN-SVM is between KNN and SVM, and the classification model can meet the practical application.

6. Conclusion

The recognition of spam is a focus in the field of machine learning at present. The selection of different algorithm models as classifiers directly affects the final screening effect. For the disadvantage of KNN and SVM, which are the two most commonly used classification algorithms, a KNN-SVM fusion algorithm is proposed. The results show that the KNN-SVM fusion algorithm can maintain high classification accuracy and recall rate, and maintain high classification efficiency, so it has good practical application value.
In the aspect of improving the algorithm, we will study the introduction of better support vector machine parameters to further improve the efficiency of the algorithm, so that it can obtain better spam recognition and classification ability.

Reference to the text of the study

[1] WANG Bin, PAN Wen-feng, A Survey of Content-based Anti-spam Email Filtering [J]. Journal of Chinese Information Processing, 2005, 19(5): 4-5.

[2] Lewis D D, Ringuette M. Comparison of two learning algorithms for text categorization [C]// Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR’94), 1994: 81-93.

[3] Masand B, Lino G, Waltz D. Classifying news stories using memory-based reasoning [C]// 15th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’92), 1992: 59-64.

[4] Ng H T, Goh W B, Low K L. Feature selection, perceptron learning, and a usability case study for text categorization [C]// 20th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’97), 1997: 67-73.

[5] Vapnic V. The Nature of Statistical Learning Theory. Springer [M]. New York, 1995.

[6] MA Jian-bin, LI Ying, TENG Gui-fa. The comparison studies on the algorithm of KNN and SVM for Chinese text classification. Journal of Agricultural University of Hebei, 2008, 31(3): 120-123.

[7] LIN Yin. Spam mail filtering model based on K nearest neighbor algorithm and support vector machine, 2016, 39(23): 90-97.

[8] LIU Bin, HUANG Tie-jun, CHENG Jun. A New Statistical-based Method in Automatic Text Classification. Journal of Chinese Information Processing, 2002, 16(6): 18-24.

[9] Wang Jinhua, Yu Hui, Chan Wen. INTEGRATING KNN AND HIERARCHICAL SVM FOR AUTOMATIC TEXT CLASSIFICATION [J]. Computer Applications and Software, 2016, 33(2): 38-41.

[10] Wang Qiang, Wang XiaoLong, Guan Yi. A research on text categorization based on the fusion of K-NN and SVM [J]. Chinese High Technology Letters, 2005, 5(15): 19-24.