A Comparative Study of Two Different Spam Detection Methods

Haoyu Wang[1], Bingze Dai[1], and Dequan Yang[2](✉)

[1] School of Information and Electronics, Beijing Institute of Technology,
Beijing 100081, China
[2] Network Information Technology Center, Beijing Institute of Technology,
Beijing 100081, China
yangdequan@bit.edu.cn

Abstract. With the development of the Internet, the problem of spam has become more and more prominent. Attackers can spread viruses through spam or place malicious advertisements, which have seriously interfered with people's life and internet security. Therefore, it is of great significance to study efficient spam detection methods. Currently using machine learning methods for spam detection has become a mainstream direction. In this paper, the machine learning method of Bayesian linear regression and decision forest regression are used to conduct experiments on a data set from UCI Machine Learning Repository. We use the trained models to predict whether a mail is spam or not, and find better prediction scheme by comparing quantitative results. The experimental results show that the method of decision forest regression can get better performance and is suitable for numerical prediction.

Keywords: Bayesian linear regression · Decision forest regression · Spam detection · Machine learning · Numerical prediction

1 Introduction

"Spam" refers to some unsolicited e-mails or text messages that often contain advertisements or trashwares. Spams are sent out through computer network and mobile phone to many different addresses, usually indiscriminately. Twitter spam is usually referred to as the unsolicited tweets that contain malicious links directing victims to external sites with malware downloads, phishing, drug sales, scams, etc. [1].

Spam email is still one of the serious problems that plague Internet communication in the world, and with the continuous development of Online Social Networks (OSNs), such as Facebook, Twitter and Instagram, these social platforms have become a very important part of people's lives, because people are using these platforms to socialize more and more. This environment where a large number of users are active at the same time has become an ideal working environment for spammers. Therefore, it is very necessary to adopt a more effective detection and filtration method for users.

Spam not only seriously wastes network resources, but also takes up users' valuable time. It also poses a threat to Internet security and directly causes huge economic losses. Spam has seriously plagued the normal mail communication of hundreds of

G. Wang et al. (Eds.): DependSys 2019, CCIS 1123, pp. 95–105, 2019.
https://doi.org/10.1007/978-981-15-1304-6_8

millions of Internet users, and has taken up a large amount of limited storage, computing and network resources on the Internet, reducing the efficiency of network use and consuming a large amount of processing time of users. Moreover, spam has gradually become a major way for viruses to spread on the Internet. Faced with the growing problem of spam, more and more technologies are being applied to anti-spam work. Therefore, it is of great significance to study efficient spam detection technology. At present, there are many methods for detecting spam. For example, blacklist is one of the most effective and convenient methods for detecting spam. However, due to its timeliness and lag, people are looking for independent updates or dynamic monitoring method. This can be achieved by the research of Fu et al. [2] Most of the current spam detection methods are implemented by detecting text messages in emails. However, Youn et al. [3] and Li et al. [4] have proposed ways to identify spam by detecting image information. In addition, spam detection using machine learning method is increasingly popular. In this article, we applied the Bayesian linear regression and the method of decision forest regression to predict mails' characteristic value to determine whether the mail is spam, and compared the results of two experiments.

The rest of the paper is organized as follow: In the second part, we will introduce the recent research about the machine learning methods of Bayesian classifiers and decision forests. In the third part, we use the Bayesian method and the decision forest method to experiment on the same mail data set respectively. In the fourth part, the results of two experiments are presented and compared. Finally, we will summarize in the fifth part.

2 Related Work

A number of studies about Bayes Classifier have been reported during last ten years. Nurul Fitriah Rusland, Norfaradilla Wahid et al. (Analysis of Naïve Bayes algorithm for email spam filtering across multiple datasets.) used the naive Bayesian algorithm to test the performance of two data sets. Their test results show that the type of e-mail and the number of data set instances have an impact on the performance of the naive Bayesian algorithm. They found that for naive Bayes classifiers, datasets with fewer e-mails and attributes perform better. Qijia Wei (Understanding of the naive Bayes classifier in spam filtering.) introduced the concept and process of the naive Bayes classifier and gave two examples. He also suggested that although the naive Bayes classifier proved to be a very efficient classification method, the interdependence between its attributes (usually words or phrases in e-mail) was limited. Jieming Yang et al. (A new feature selection algorithm based on binomial hypothesis testing for spam filtering.) proposed a new method called Bi-Test to evaluate whether the probability of being classified as spam satisfies the threshold. They used Naive Bayes (NB) and Support Vector Machines (SVM) classification algorithms to separate the six benchmark spam corpora (pu1, pu2, pu3, pua, lingspam, CSDMC2010). Test was evaluated and compared with four well-known feature selection algorithms (information gain, v2 - statistical, Gini index, Poisson distribution). The experimental results show that when using the naive Bayes classifier, the performance of the double test is significantly better than the v2 - statistic and Poisson distribution, which is equivalent to the information gain and the improved Gini index performance on the F1 measure; when using the SVM classifier,its performance is comparable to other methods. Moreover, Bi-Test performs faster than the other four algorithms. In the study of Lizhou Feng et al. (Quick online spam classification method based on active and incremental learning.) to improve the classification speed of mail, some of them train the classifier according to the incremental learning theory. They will support vector machine (SVM), naive Bayesian classifier (NB) and k-nearest neighbor. The classifier (KNN) is used for the two types of classifiers, Trec2007 and Enron-spam. The experimental results show that compared with the six typical active learning-based incremental learning methods, the proposed method greatly reduces the time-consuming of mail classification while ensuring classification accuracy. Chong-zhi Gao et al. (Privacy-preserving Naive Bayes classifiers secure against the substitution-then-comparison attack.) constructed a privacy-protected NB classifier that is resistant to replacement and then comparison (STC) attacks. In the case of not using the full homomorphic encryption with large computational overhead, a scheme for avoiding information leakage under the STC attack is proposed. Our key technology involves the use of "double-blind" technology and demonstrates how it can be combined with additional homomorphic encryption and unrelated transmission to hide the privacy of both parties.

At the same time, machine learning method of random forest has been widely used in spam detection. He Long (Identification of Product Review Spam by Random Forest.) proposed a random forest-based product spam comment recognition method, which is to repeatedly extract the same number of samples from the large and small categories in the sample or give the same weight to the total samples of the large and small categories to establish the random forest model. Moreover, its experimental results on amazon data set show that the recognition results based on random forest are better than other baseline methods. Al-janabi, M et al. (A systematic analysis of random forest based social media spam classification.) conducted systematic analysis on random forest classification, and assessed the impact of key parameters such as tree number, tree depth and minimum size of leaf nodes on classification performance. Their research results show that controlling the complex random forest classifier is of great significance to the classification of social media spam. Sun Xue et al. (One Email Filtering System Based On Category Feature Selection And Feedback Learning Random Forest Algorithm.) proposed an email filtering model based on category feature selection and feedback learning stochastic forest algorithm. Their experimental results show that this method can alleviate the impact of redundant information and noise data on classification performance effectively, and can realize the self-regulation of email filtering system and timely catch the changing trend of spam.

Together these studies provide important insights into the Bayesian approach and random forest method to spam detection.

3 Experimental Method

This article uses the Spambase Data Set created by Mark Hopkins et al. in the UCI Machine Learning Repository. This data set extracts some characteristics of spam and quantifies these characteristics to build a digital data set. Two regression algorithms for numerical prediction are used in this paper to conduct experiments, which are Bayesian regression algorithm and decision forest algorithm. The two approaches and their advantages are briefly described below.

3.1 Bayesian Linear Regression

When we only have limited data or want to use prior probabilities in the model, Bayesian linear regression can satisfy these needs. The Bayesian linear regression method is special compared with other regression algorithms for that Bayesian linear regression is not to find the optimal value of the target parameter, but to determine the posterior probability distribution of the model parameters. By training the model through input and output parameters in the dataset, the posterior distribution of a parameter in the model can be obtained.

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} \tag{1}$$

In this function, *P(x|y)* is the posterior probability distribution of a model parameter calculated from a pair of given input and output. It is equal to the likelihood of the output *P(y|x)* multiplied by the prior probability *P(x)* of the parameter x for a given input and divided by the normalization constant. This is a simple form of expression of Bayes' theorem which is the basis for supporting Bayesian inference.

$$Posterior = \frac{Likelihood * Prior}{Normalization} \tag{2}$$

Linear regression model is a linear combination of the basis function of a set of input variable x, and the mathematical expression is

$$y(x, w) = w_0 + \sum_{j=1}^{M} w_j \phi_j(x) \tag{3}$$

M is the number of basis functions, we assume that $\phi_0(x) = 1$, then

$$y(x, w) = \sum_{j=0}^{M} w_j \phi_j(x) = w^T \phi(x) \tag{4}$$

where $w = \{w_0, \ldots, w_M\}$, $\phi = \{\phi_0, \ldots, \phi_M\}$, then the probability density function of the linear model is

$$P(T|x, w, \beta) = \prod_{i=1}^{N} N(t_i|y(x, w), \beta^{-1} I) \tag{5}$$

T is the target data vector, $T = \{t_1, \ldots, t_N\}$.

Assuming that the prior probability $P(w)$ obeys the Gaussian distribution

$$P(w) = N(w|0, \alpha^{-1}I) \tag{6}$$

So the posterior probability can be expressed as

$$P(w|X,T) = \frac{P(T|w,X) \cdot P(w)}{P(T|X)} \tag{7}$$

After the posterior probability distribution is obtained through training the model, we can acquire the value of the estimated parameter with the maximum posterior probability density, which is $\hat{w}$. So based on this estimated parameter, the output estimate with new data input can be estimated.

Compared with other typical regression algorithms such as Ordinary Least Squares (OLS) and Maximum Likelihood Estimation (MLE), Bayesian linear regression has three main advantages:

1. Prior distribution: If there are data or reasonable guesses about a domain or model parameters, then they can be included in the process of using Bayesian linear regression, rather than when using OLS, all required information about the parameters needs to be obtained from the data. If there is no prediction in advance, non-information priori can applied on the parameter, such as a normal distribution. Using this estimation may produce larger errors when the data is small, but as the data points increase, estimates will increasingly trend towards the values predicted by OLS.
2. Posterior distribution: the result of Bayesian linear regression is a distribution of model parameters based on training data and prior probability. This allows the quantification of the uncertainty of the model: if there are fewer data points, the posterior distribution will be more dispersed. As the amount of data points increase, the effect of the a priori will reduce. When there are have infinite data, the output parameters converge to the values obtained using the OLS method.
3. Prevent over-fitting: Since the maximum likelihood estimation would make the model too complex to produce over-fitting, simply using maximum likelihood estimation is not always an effective method. While the Bayesian linear regression can solve the problem of over-fitting in the maximum likelihood estimation.

This formula that uses model parameters as a probability distribution reflects the essence of Bayesian theory: starting with the initial estimate and the prior distribution, the model makes fewer mistakes as more data is collected, and gets closer to the truth. Bayesian reasoning can also be understood as a natural extension of our intuition. For example, we have an initial hypothesis at the beginning, and with the collection of data that supports or denies ideas, our model of the world's perceptions will change.

3.2 Decision Forest Regression

The random decision forest regression method can generate a new decision model. Obviously, the random decision forest is to establish a forest model in a random way. The forest model consists of many decision trees, and there is no correlation between each decision tree. A decision tree consists of nodes and directed edges. Generally, a decision tree contains a root node, several internal nodes, and several leaf nodes. The node contains the attributes of the objective function it depends on, and the value of the objective function reaches the leaf nodes through the branch. The decision process of the decision tree needs to start from the root node of the decision tree, and the data to be tested is compared with the feature nodes in the decision tree, and select the next comparison branch according to the comparison result until the leaf node is the final decision result. Repeat the above process to get a forest with t decision trees. The decision tree algorithm in stochastic decision forest regression is a process of recursively constructing a decision tree. The minimum error criterion is used to select features and generate a binary tree [13]. After obtaining the forest model, once there is a new input, each decision tree in the forest will discriminate the sample and give a predicted value. Finally, the value of the sample is taken as the average of the predicted values for all decision trees. Figure 1 shows us the random decision forest frame. In the process of establishing a decision tree, the sample needs to be sampled. The sampling method with a put back is applied here. Assuming that there are N input samples, the sampled samples are also N. We also assume that the number of input features is M. When splitting on each node of each decision tree, m input features are randomly selected from M input features, and then choose the best one from the m input features for splitting. m does not change during the construction of the decision tree. As a result, each tree's sample size is not all samples during training which leads to the advantage that it's not easy to come to over-fitting. Each tree in a regression decision forest outputs a Gaussian distribution as a prediction. An aggregation is performed over the ensemble of trees to find a Gaussian distribution closest to the combined distribution for all trees in the model.

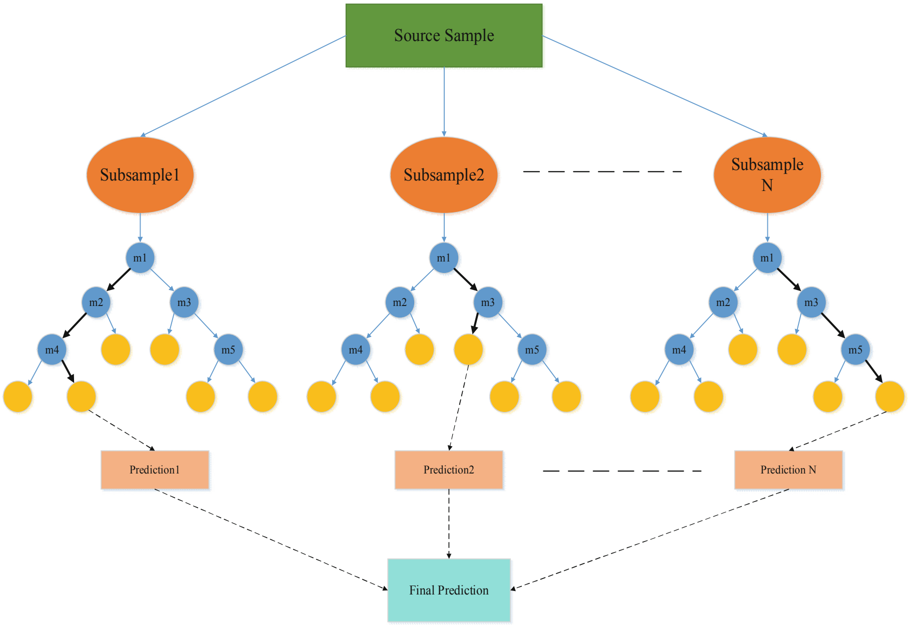

Fig. 1. Random decision forest framework

Random decision forests have several advantages:

1. Training can be highly parallelized, can run efficiently on large data sets, can produce high-accuracy classifiers;
2. Can handle a large number of input variables;
3. While classifying samples, can output the importance of each feature to the predicted target;
4. When some features are missing, the accuracy can still be maintained, and the tolerance for feature loss is high;
5. The training process of random forest is very fast.

3.3 Experimental Process

The experiments in our paper were carried out in Microsoft's Azure Machine Learning Studio, using existing machine learning models: Bayesian linear regression model and stochastic decision forest model. First, we upload the spam dataset downloaded from the UCI database to the studio platform and use it as the first module of the process. Then we divide the data set and use 75% (3,450 emails) of the data set for the training of the model. The features used for training were all the attributes given in the data set, and the predicted attribute is whether the email is spam or not. The remaining 25% (1150 emails) of data were used to test the model after training, and the predicted values were finally obtained and compared to known results. The specific flow charts are as follows (Figs. 2 and 3).

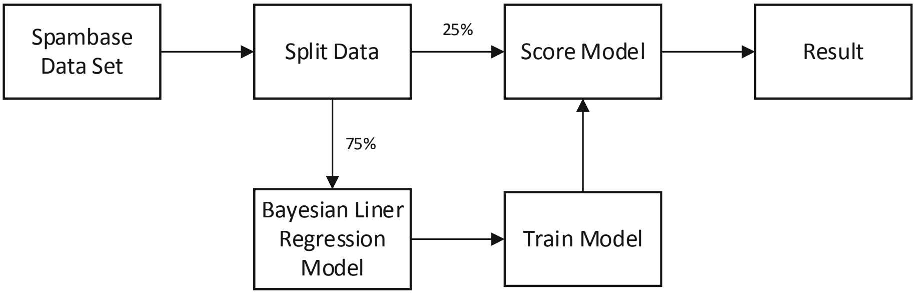

Fig. 2. Bayesian liner regression method

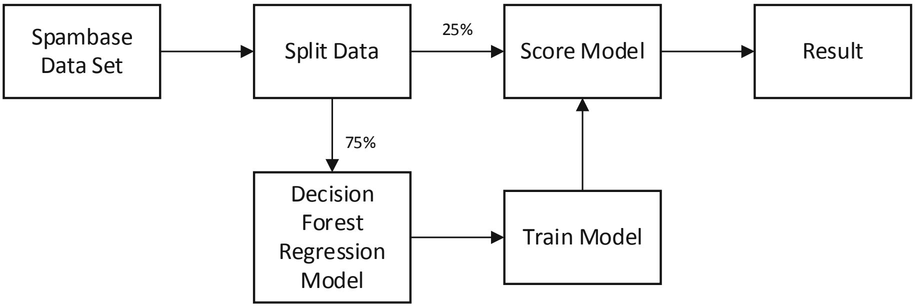

Fig. 3. Decision forest regression method

In the random decision forest model, the decision tree node parameters are set as Table 1.

Table 1. Decision tree node parameters.

Number of parameter	Value
Number of decision trees	8
Maximum depth of the decision trees	32
Number of random splits per node	128
Minimum number of samples per leaf node	1

When everything is ready, the experiment can begin.

4 Experimental Result

In UCI's Spambase Data Set, spams are quantified by number 1, and non-spams are quantified by number 0, so the trained model predicts the characteristic value of an email based on the input values of attributes, and finally classifies the characteristic value of the email as 1 or 0. The email is considered spam within a certain range of values close to 1, and we consider the email to be non-spam within a certain range close to 0. In order to verify the performance of the model predictions, we know in advance that 25% (1150 emails) of the datasets participating in the test have 435 spams and the rest are non-spams. After constructing the model according to the flow chart in Sect. 3.3, the two experiments using different methods are carried out, and the predicted value distribution results of the mails that participated in the test are shown in Figs. 4 and 5.

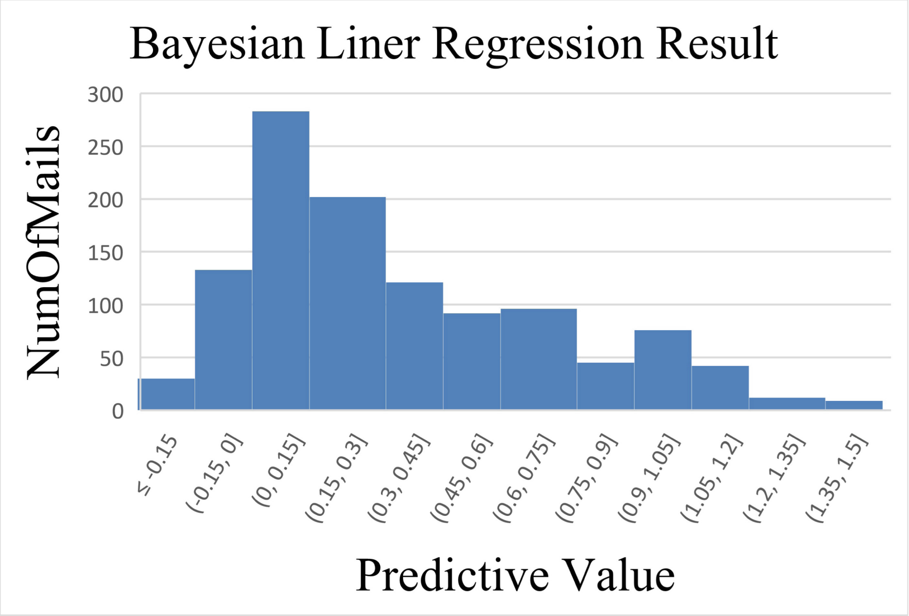

Fig. 4. Predicted value of Bayesian linear regression.

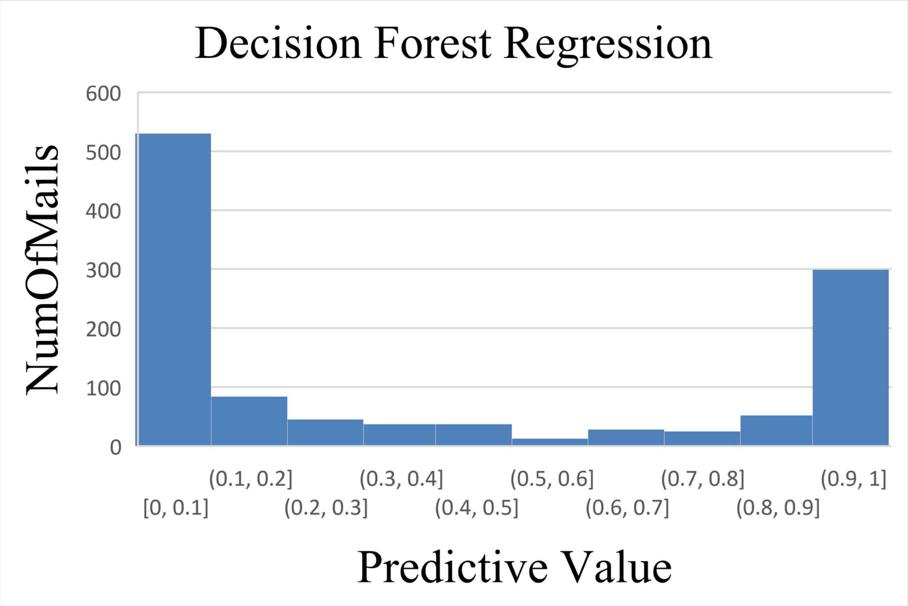

Fig. 5. Predicted value of decision forest regression.

It can be seen from Fig. 4 that there is a peak between the predicted value range (–0.15, 0.15] (i.e. the predicted value is near 0) and another peak between the predicted value range is (0.9, 1.05] (i.e. the predicted value is near 1). This result indicates that about half of the emails in the data set participating in the test can be predicted as spam or non-spam, and the distinction between tow peaks is obvious. However, since Bayesian linear regression predicts the posterior probability distribution of a parameter, so it can be seen that forecasts are widely distributed. Moreover, there are still some emails that were divided into the fuzzy area of the middle of predicted range. This means that these mails are not obviously distinguished whether they belong to the non-spam or spam. The presence of these mails also reflected one flaw of the Bayesian linear regression method in the practical application: it requires a certain amount of samples to train the model to obtain good results.

From Fig. 5 we can see that all the predicted values are distributed in the interval [0, 1], and most of the mail can be classified as spam or non-spam. Compared to Bayesian linear regression method, the number of characteristic values predicted by the decision forest model in the middle fuzzy region is less. And the closer to the middle area (i.e. the predictive value of 0.5), the less number of divided mails are. The result indicates that the random decision forest has the advantages of high accuracy. The advantage is that only using more trees or setting a higher tree depth will make the model more adequately trained and ultimately make the prediction performance even better.

By contrast, we can see that the decision forest regression algorithm has better performance than the Bayesian linear regression algorithm, and the predicted value is more close to the actual situation. The Mean Square Error (MSE) between the predicted value and the actual value in the random forest algorithm is 0.053, while the MSE between the predicted value and the actual value in the Bayesian linear regression algorithm is 0.111. Therefore, compared with the Bayesian linear regression algorithm, the accuracy of the random forest regression algorithm is improved by more than half, so it is a machine learning method with more accurate prediction, and can more accurately describe and predict experimental data.

5 Conclusion

In this paper, we conduct two experiments on Azure Machine Learning Studio platform using the Machine Learning methods of Bayesian linear regression and random decision forest regression to detect a given set of spam data. Through the experimental results, we can see the difference of the prediction results caused by the difference of the two methods. The Bayesian linear regression method is based on the posterior probability distribution of the characteristic parameters, so its predicted values are relatively scattered. However, the random decision forest method uses the least square error criterion to generate the binary tree for feature selection. Compared to the Bayesian linear regression method, it has higher accuracy, and the random decision forest regression also has the advantages of simple modeling and fast training speed, so it is very suitable as a benchmark model of machine learning. With the continuous development of artificial intelligence technology and the continuous advancement of machine learning technology, researchers will surely reach a new level of spam classification to meet the needs of users for a good email communication environment. The likely direction for spam detection is to produce a better classification standard, such as extracting more complex and accurate attributes that can determine spam attributes as feature signatures. In addition, researchers can develop more excellent low complexity gain algorithms based on random decision forests, such as neural networks. Therefore, the development of a more accurate and faster spam detection method is one of the future development directions in the field of machine learning.

Acknowledgments. This research was supported by CERNET Innovation Project (NGII20180407).

References

1. Benevenuto, F., Magno, G., Rodrigues, T., Almeida, V.: Detecting spammers on twitter. In: Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS), p. 12, July 2010
2. Fu, Q., Feng, B., Guo, D., Li, Q.: Combating the evolving spammers in online social networks. Comput. Secur. **72**, 60–73 (2018)
3. Youn, S., Cho, H.C.: Improved spam filter via handling of text embedded image e-mail. J. Electr. Eng. Technol. **10**(1), 401–407 (2015)
4. Li, S., et al.: WAF-based chinese character recognition for spam image filtering. Chin. J. Electron. **27**(5), 1050–1055 (2018)

5. Rusland, N.F., Wahid, N., Kasim, S., Hafit, H.: Analysis of Naïve Bayes algorithm for email spam filtering across multiple datasets. In: IOP Conference Series: Materials Science and Engineering. IOP Publishing, August 2017
6. Wei, Q. Understanding of the naive Bayes classifier in spam filtering. In: AIP Conference Proceedings. AIP Publishing (2018)
7. Yang, J., Liu, Y., Liu, Z., Zhu, X., Zhang, X.: A new feature selection algorithm based on binomial hypothesis testing for spam filtering. Knowl.-Based Syst. **24**(6), 904–914 (2011)
8. Feng, L., Wang, Y., Zuo, W.: Quick online spam classification method based on active and incremental learning. J. Intell. Fuzzy Syst. **30**(1), 17–27 (2016)
9. Gao, C.Z., Cheng, Q., He, P., Susilo, W., Li, J.: Privacy-preserving Naive Bayes classifiers secure against the substitution-then-comparison attack. Inf. Sci. **444**, 72–88 (2018)
10. He, L.: Identification of product review spam by random forest. J. Chin. Inf. Process. **29**(3), 150–154 (2015)
11. Al-Janabi, M., Andras, P.: A systematic analysis of random forest based social media spam classification. In: Yan, Z., Molva, R., Mazurczyk, W., Kantola, R. (eds.) Network and System Security. NSS 2017. LNCS, vol. 10394, pp. 427–438. Springer, Cham. https://doi.org/10.1007/978-3-319-64701-2_31
12. Sun, X., Han, L., Li, K.: One email filtering system based on category feature selection and feedback learning random forest algorithm. Comput. Appl. Softw. **32**(4), 67–71 (2015)
13. Huawei Cloud. https://support.huaweicloud.com/algnoderef-mls/mls_02_0054.html. Accessed 05 July 2019