Classification of Spam E-mail based on Naïve Bayes Classification Model

Shaopeng Cheng

University of California, Santa Barbara CA 93106, California, United States

shaopeng@ucsb.edu

Abstract. With the rising number of spam email, the need of more sufficient antispam filter is surging. Phishing attack can lead to extremely large losses of companies and individual, even more than 1 billion dollars in one year. This paper investigates and combines Naïve Bayes Classification and clustering algorithm in the application of identifying spam emails. With sample emails to create a dynamic dictionary containing most frequent words in spam and normal emails, this distribution of spam filter will provide a stricter method to prevent spam emails than those methods used in mail companies, e.g., Google, Yahoo, and Outlook.com. Besides, this paper also compares several algorithms used today in classifying spams and the future techniques of deep learning and machine learning's application in classifying spam emails. According to the analysis, Google's algorithm has the most comprehensive function, but such algorithm has less strict rule than Yahoo's. Outlook.com, as a combination of Microsoft application, it has a unique algorithm for encrypting and filtering spams. Overall, these results shed light on guiding further exploration of both comprehensive and strict rule for classifying spams.

Keywords: Naïve Bayes Classification; Spam Emails; Machine Learning.

1. Introduction

Contemporarily, unwanted emails, which are called spam emails, become a huge problem for the society. Both companies and individuals are suffering from the loss of time and money caused by these spam emails. One report argues that the threat of spam emails keeps increasing, and the amount of spam emails is responsible over 75% of all emails. In order to effectively solve these problems caused by spam emails, leading email companies (e.g., Google, Yahoo, and Outlook.com) have developed advanced email filtering methods based on machine learning and artificial intelligence such as the application of neural network. The combination of machine learning and Naïve Bayes Classification algorithm can achieve 99.9% accuracy. However, there are over 1 billion emails sent one day, even such a system still has numbers of spam emails successfully evading. Bahdad pointed that the improvement of system to prevent spam emails is not enough to stop penetrating [1]. Therefore, such system should also can identify fraud and prevent it.

The most widely used and advanced methods of machine learning used in spam filtering is the Naïve Bayes Classification, which is a filtering method based on content. This method normally analyses words, the occurrence, and distributions of words and phrases in the content of emails and used then use generated rules to filter the incoming email spams [2]. Another common approach of classifying spam is knowledge engineering. However, when people want to use the knowledge engineering, they will create a set of rules to identifying spams, so users will lose amount of time in updating and changing their rules. Therefore, machine learning is gradually replacing the knowledge engineering's place in classifying spam emails. Machine learning algorithms can learn a specific rules in email data and use them to create classification [3]. In addition to these algorithms, stochastic optimization techniques are also involving in filtering spams. They can recognize the optimal solutions for problems which do not have complexed mathematical computation [4]. However, spammers have used more sophisticated methods, such as utilizing arbitrary sender addresses to surmount the barriers created by filtering methods [5]. Therefore, more reliable techniques should be proposed to against new methods used by spammers.

The motivation of this paper is aimed to find a more radical way to classify spam emails in order to achieve an ideal result in filtering spam emails. This paper discusses some algorithm used widely and compare the advantage of them, and then come up a new algorithm based on K-nearest neighbors (KNN) algorithm and Naïve Bayes Classification. With relatively small database, this algorithm still has some vulnerability and lack of training. Therefore, this paper also clarifies some other useful ways for identifying spam emails and the advantage and disadvantage of different algorithms. Finally, this paper exhibits one assumption of spam filtering based on the improvement of machine learning and artificial intelligence.

2. Methodology

Distinguishing spams always needs large database, which contained both Ham and Spam emails. Accuracy is not the only performance for a spam filter. It's also important for a spam filter not to misclassification. When a spam email is wrongly classified to ham email, which is the normal email, people can easily solve this problem by clicking the delete button. Nevertheless, when one ham email is classified to the spam email, it is a big trouble because people will be worried about why they did not receive their email. Therefore, it is inadequate to evaluate the performance of a machine learning algorithm only by its accuracy.

Dada collected and compare common methods used for spam filtering, such as clustering, Naïve Bayes Classifiers [6]. Clustering will be a good method to classify spam emails and ham emails, which is showed by Whissell and Clarke [7]. The output of their research remarkably showed that their technique performed better than those of semi-supervised techniques. Clustering will classify objects or opinions which in the same group are more similar than those in the other group. Two methods of clustering which are used for spam filtering are density-based and K-nearest neighbors (KNN). K-nearest neighbors' distribution is very critical for classifying spam, because this classification model is based on the test instance of training examples. Therefore, when the user receives a new email, the classifier will compare the new email with these K examples and decide whether it's a spam or not.

Another useful and reliable method is Naïve Bayes Classifiers, which is a supervised learning technique for machine learning. This classification provides a learning algorithm based on the previous message and data. Two obvious advantage of this classification is that it only needs small group of databases of ham and spam emails for the computer to train and every result of classifying emails is independent from each other, which is very important for filtering spam. Naïve Bayes has become a simple and computationally efficient algorithm for classification of text, spam email, spam reviews and other online application. According to Li and Zhong's research [8], the token T which denote the rate of spam is computed as illustrated:

$$S[T] = Cspam(T)/Cspam(T) + Cham(T) \quad (1)$$

where Cspam(T) and Cham(T) are the number of spam or ham messages containing token T, respectively. Then, the total rate of spam should be calculated as the product of individual emails:

$$H[M] = \prod_{i=1}^{N}(1 - S[T_i]) \quad (2)$$

3. Results & Discussion

Machine learning involves two phases: the training phase and the testing phase. As mentioned in section 2, both clustering and Naïve Bayes Classifiers are useful and reliable methods for classifying spams in such training phase. In order to achieve the goal to filter spams, Python gives a set of packages to help. First, as mentioned in clustering method, it is important to create a K-set of emails, which can provide the K-example to classify spams. Therefore, one way for creating such K-example is to create a dictionary for most frequent words in spams. Besides, Naïve Bayes Classification allows

fewer cases and examples for emails. Thus, this K-example is created by 1400 emails, containing both 700 normal emails and 700 spam emails. This study creates two separate dictionaries, which contain 4000 words, for each spam and normal emails. Such dictionary will give the spam filter a set of K-example to help distinguish spam and ham emails.

Table 1. The most frequent words in normal and spam emails.

Most Frequent Words in Normal Emails									
re	to	the	for	in	a	with	on	is	of
Most Frequent Words in Spam Emails									
spam	hibody	for	you	your	80	to	off	on	the

Then, one creates an array containing these words as K-example for K-nearest neighbors' distribution. In this case, it is obvious for users and testers to observe these most frequent words as summarized in Table. 1. It is clear that spam emails will contain more personal pronoun and numbers in it. Therefore, spam filter will begin to classify spams based on these results and the algorithm provided by Naïve Bayes Classifier.

The tester chose 400 emails containing 200 spams and 200 normal emails. Tester tests these two sets of emails separately, and the results are given in Table. 2. This method of spam filter successfully classifies most of these emails, above 90%. The data center of Google will use hundreds of rules to distinguish one email is whether a spam or not. Each feature also has their own weighted importance to be calculated. Besides, the reports said Google also has the method of neural networks and logistic regression for their spam filter. However, the algorithm of Google is more conservative and since such algorithm is based on the IP address, some emails from institutions and other schools will be classified as spam, which will give users lots of inconvenience. Unlike Google's algorithm, Yahoo spam mail filter is based on the domains to detect users who sent spam emails. The classification methods used by Yahoo is stricter and such filter method will prevent any emails that is not approved and authorized. Anonymous senders' emails will always be sent to spam emails instead of inbox. Outlook.com 's spam filtering is different from both Google's and Yahoo's. Since Outook.com is a combined of application provided by Microsoft, and it has cloud service for users to receive and check their emails in the Google Drive or Dropbox. Outlook.com has a unique encrypted method for users to protect their email. According to report, the Outlook.com server will receive over eight billion emails, and the algorithm of their email filtering will prevent over 60% emails.

Table 2. The results of the models.

MultinomialNB	Normal	Spam
Normal	181	12
Spam	11	188

Comparing to different spam filtering distributions used in Google, Yahoo, or Outlook.com, such method will give more accurate results. The reports argued that about 20 percent of authorized emails cannot be receive as expected. Within this distribution of spam filter, users can also create an array containing words that they do not expect to receive from emails. For example, if the user does not want the promotion email of pizza shop, they can add pizza to this dictionary and the spam filter will automatically intercept these promotion emails. Additionally, other examples of such filter distribution include white lists, recipient verification tools, etc.

The email is made up of two messages, which are header parts and the body parts. Normally, when an email is sent and went through an in-between server, the header always has some modification. Therefore, some spam filter will be cheated when these emails changed their headers. However, the spam filter distribution mentioned in section 3, will involve in both header parts and body parts. This distribution will help those people who want harsher and stricter classification of promotion and spam emails. Besides, this distribution will be useful in other aspect, such as recommender system, text analysis, and other online applications.

4. Limitations & Prospects

It is clear that the accuracy of this spam classification has not achieved 100%. Some normal emails are classified as the spams, and some spam emails are distinguished as normal emails. As mentioned in section 2, it is a tremendous mistake to put normal emails into spam emails because users will not receive the emails as they expected. Naïve Bayes Classification is useful but also has some limitation. The discovery shows the performance of Naïve Bayes Classification cannot achieve 100% correctness in filtering spams. Therefore, with the development of artificial intelligence, some advanced algorithm combined Naïve Bayes Classification algorithm will show up. One aspect which artificial intelligence is neural networks. It can communicate units and their neighbor units to calculate the output. Neural network is the algorithm that worked for most machine learning problem. With the combination of neural networks and Naïve Bayes Classifier, the accuracy will almost achieve 100%, which is mentioned in Paul Graham's work "A Plan for Spam" [9]. These well-interconnected neurons will work together to solve certain problem, such as distinguish spam emails. The combination of Naïve Bayes Classification and neural networks is worked in most email filter, such as Google and Outlook.com.

Besides, there are also other ways for classifying spams including firefly algorithm, rough set classifiers, support vector machine classifiers, decision tree, and other methods. Firefly algorithm and decision tree can also provide reliable results of classifying emails. Yang got inspiration from sparkly behavior of fireflies and develop this method of classifying based on metaheuristic algorithm [10]. When fireflies are about to copulate, they are exposure to the danger because they will share their information by sparkling. There are over 2000 kinds of fireflies in the world, and each one has a specific sparkling way. Depending on the way of sparkling, the companion will communicate by imitating the same form of sparkling. This algorithm is similar with K-nearest neighbors (KNN) algorithm. However, the firefly algorithm is an optimization algorithm because this method will follow a better or brighter, as mentioned in the copulation of fireflies, solution.

The decision tree algorithm is different form Naïve Bayes Classification. Each node of decision tree has the connection with other nodes. The principle of decision tree is recursion. The decision tree model will provide a value of goal variable centered on other values, which become nodes in the decision tree. The test email will go through some process in the tree and finally achieve one leaf node. Nevertheless, such algorithm will spend more time than Naïve Bayes Classification. Each tested email will go through all leaf nodes in the tree, which will cost lots of time. Hence, if the tested value can achieve the final leaf node in a quicker and more accurate way, decision tree will also have its own significance in classify spam emails.

5. Conclusion

In summary, spam emails have become a virus of global security and economy. Although mail companies have come up with different useful and advanced methods in classifying spams, users are still influenced by tremendous number of spams in the worlds. In this paper, with comparing to nowadays spam filtering of mail companies, such as Google, Yahoo, and Outlook.com, it provided an algorithm based on Naïve Bayes Classification and clustering algorithm. This algorithm can allow users to modify their own dictionaries for both spam emails and normal emails and a stricter algorithm to identify spam emails with over 90% accuracy. However, the tested emails in this paper are relatively small, and the accuracy cannot achieve to an ideal goal. There are still aspects to improve. In the future, the spams senders will shift their attention from syntactic attacks, such as attack based on the technical vulnerability, to semantic attacks. Thus, with the development of artificial intelligence, more reliable neural network will become a big help of Naïve Bayes Classification, which is the most used algorithm in distinguishing spams in the world. In addition, some new algorithms can also improve the efficiency of spam filtering (e.g., firefly algorithm), which is a good extension to clustering algorithm. Overall, these results offer a guideline for filtering the spam in a

way with the combination of Naïve Bayes Classification and clustering algorithm to achieve both comprehensive and strict function.

References

[1] Behdad, M., Barone, L., Bennamoun, M., et al.: Nature-inspired techniques in the context of fraud detection. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(6), 1273-1290 (2012).

[2] Christina, V., S. Karpagavalli, G. Suganya.: Email spam filtering using supervised machine learning techniques. International Journal on Computer Science and Engineering (IJCSE) 2.09, 3126-3129 (2010).

[3] Guzella, T. S., Walmir M. C.: A review of machine learning approaches to spam filtering." Expert Systems with Applications 36.7, 10206-10222 (2009).

[4] Qasem, S. N., Siti, M. S., Azlan, M. Z. Multi-objective hybrid evolutionary algorithms for radial basis function neural network design. Knowledge-Based Systems 27, 475-497 (2012).

[5] Stuart, I., Cha S. H., Charles T.: A neural network classifier for junk e-mail. International Workshop on Document Analysis Systems. Springer, Berlin, Heidelberg, 2004.

[6] Dada, E. G., Bassi, J. S., Chiroma, H., et al.: Machine learning for email spam filtering: review, approaches and open research problems. Heliyon 5.6, e01802 (2019).

[7] Whissell, J. S., Charles L. A. C.: Clustering for semi-supervised spam filtering. Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference. 2011.

[8] Li, K., Zhong, Z.: Fast statistical spam filter by approximate classifications. Proceedings of the joint international conference on measurement and modeling of computer systems. 2006.

[9] Graham, Paul.: A plan for spam. Retrieved from: http://www. paulgraham. com/spam. html (2002).

[10] Yang, X. S.: Firefly algorithms for multimodal optimization. International symposium on stochastic algorithms. Springer, Berlin, Heidelberg, 2009.